SIAM J. OPTIM.
Vol. 19, No. 1, pp. 1–20

ERROR BOUNDS IN METRIC SPACES AND APPLICATION TO THE PERTURBATION STABILITY OF METRIC REGULARITY*

HUYNH VAN NGAI† AND MICHEL THÉRA‡

Abstract. This paper was motivated by the need to establish some new characterizations of the metric regularity of set-valued mappings. Through these new characterizations it was possible to investigate the global/local perturbation stability of the metric regularity and to extend a result by Ioffe [*Set-Valued Anal.*, 9 (2001), pp. 101–109] on the perturbation stability of the global metric regularity when the image space is not necessarily complete. It was also possible to give a characterization of the local metric regularity and to derive a local version of the perturbation stability of the metric regularity. In this work we also describe an application of this perturbation stability and give a simple proof of a result on the error bound of 2-regular mappings established by Izmailov and Solodov [*Math. Program.*, 89 (2001), pp. 413–435] and generalized by He and Sun [*Math. Oper. Res.*, 30 (2005), pp. 701–717].

Key words. error bound, perturbation stability, metric regularity, generalized equations

AMS subject classifications. 49J52, 49J53, 90C30

DOI. 10.1137/060675721

1. Introduction.

Let X and Y be metric spaces endowed with metrics both denoted by $d(\cdot,\cdot)$. The open ball with center x and radius $r > 0$ is denoted by $B(x, r)$. We recall that a set-valued (multivalued) mapping $F : X \rightrightarrows Y$ is a mapping which assigns to every $x \in X$ a subset (possibly empty) $F(x)$ of Y. As usual, we use the notation $\text{gph } F := \{(x, y) \in X \times Y : y \in F(x)\}$ for the graph of F, $\text{Dom } F := \{x \in X : F(x) \neq \emptyset\}$ for the domain of F, and $F^{-1} : Y \rightrightarrows X$ for the inverse of F. This inverse (which always exists) is defined by $F^{-1}(y) := \{x \in X : y \in F(x)\}, y \in Y$, and satisfies

$$(x, y) \in \text{gph } F \iff (y, x) \in \text{gph } F^{-1}.$$

It is well known that a large amount of problems, for instance, inequalities and equalities systems, variational inequalities, or systems of optimality conditions, are covered by the solvability of a generalized equation (in the terminology of Robinson).

For a given $y \in Y$, determine $x \in X$ such that $y \in F(x)$.

In general F is of the form $f + T$, where $f : X \to Y$ and $T : X \rightrightarrows Y$. An important subcase is furnished by variational inequalities, that is, the problem of finding a solution to the equation $y \in f(x) + N_C(x)$, where $T = N_C$ is the normal-cone operator. For each $x \in \mathbb{R}^n$, the set $N_C(x)$ is the normal cone (in the sense of convex analysis) to a closed convex set C of $\mathbb{R}^n$ at x.

A central issue in variational analysis is to investigate the behavior of the set of solutions of a generalized equation associated to F, that is, the behavior of the

*Received by the editors November 22, 2006; accepted for publication (in revised form) June 25, 2007; published electronically February 6, 2008.

http://www.siam.org/journals/siopt/19-1/67572.html

†Department of Mathematics, University of Quynhon, 170 An Duong Vuong, Qui Nhon, Vietnam (uguiakhiem@yahoo.com). This author's research was supported by XLIM (Department of Mathematics and Informatics), UMR 6172, Université de Limoges, and by PICS CNRS Formath Vietnam.

‡Laboratoire XLIM, UMR-CNRS 6172, Université de Limoges, 87060 Limoges cedex, France (Michel.thera@unilim.fr). This author's research was partially supported by "Fondation EADS" and by Agence Nationale de la Recherche under grant ANR NT05 − 1/43040.

set $F^{-1}(y)$ when y and/or F itself are perturbed. A key to this is the concept of metric regularity. Recall that a mapping F is said to be *metrically regular* on a region $V \subseteq X \times Y$ with modulus τ if there exists a real $\tau > 0$ such that

$$d(x, F^{-1}(y)) \leq \tau d(y, F(x)) \quad \text{for all } (x, y) \in V, \tag{1.1}$$

where $d(x, C)$ denotes, as usual, the distance from x to a set C and is defined by $d(x, C) = \inf_{z \in C} d(x, z)$, with the convention that $d(x, S) = +\infty$ whenever S is empty. In the case, for example, of a set-valued mapping F with closed and convex graph, the Robinson–Ursescu theorem says that F is metrically regular at (x_0, y_0) if and only if y_0 is an interior point to the range of F, i.e., to Dom F^{-1}.

If relation (1.1) holds for all (x, y) close to a given $(\bar{x}, \bar{y}) \in X \times Y$, then we say that F is *(locally) metrically regular* around $(\bar{x}, \bar{y})$. According to the long history of metric regularity there is abundant literature on conditions ensuring this property. This concept goes back to Lyusternik [29] and Graves [18] in connection to the extension to nonlinear operators of the celebrated Banach open mapping theorem. For a detailed account the reader is referred to the works [2, 6, 7, 8, 9, 11, 14, 20, 21, 23, 27, 28, 30, 31, 32, 33, 34, 38, 41] and the references given therein for many theoretical results on the metric regularity as well as its various applications.

In the present paper, we are concerned with the stability of the metric regularity with respect to perturbations of F. Historically, it follows from Banach that when F is a bounded linear operator between two Banach spaces X and Y, if F is surjective, then all operators G sufficiently close to F have the metric regularity property (or equivalently, the covering property). This classical result has been extended to the case of continuously Fréchet differentiable mappings by Lyusternik [29] and Graves [18] for the local metric regularity and by Dmitruk, Miljiutin, and Osmolovskii [13] for the global case. In [15] Dontchev, Lewis, and Rockafellar, and [16], Dontchev and Lewis have studied the case of set-valued mappings under perturbations of a single-valued mapping; that is, perturbation mappings of the form $F(x) + g(x)$, where F is a set-valued mapping and g is a single-valued mapping.

Recently, in [22], Ioffe studied the general case when perturbation mappings are not necessarily expressed in the form $F(x)+g(x)$. He constructed a measure of "closedness" between two set-valued mappings allowing him to significantly extend the classical result (on the global covering property) of Dmitruk, Miljiutin, and Osmolovskii to the general case of set-valued mappings.

Inspired by the work of Ioffe [22], our main objective in this paper is to use the theory of error bounds to study stability of global and/or local metric regularity of set-valued mappings under a perturbation of F. The approach based on error bounds to investigate the metric regularity has been recently used by Azé, Corvellec, and Lucchetti [3] and by Ngai and Théra [35] to study implicit multifunctions in smooth spaces. Especially in the survey paper by Azé [2], it was shown that this approach is powerful to provide a unified theory of the metric regularity. The organization of this paper is as follows. In section 2, we prove new characterizations of the global/ local error bound in complete metric spaces. Using this result, we derive in section 3 a new criterion assuring the metric regularity. Based on this criterion, we extend the result by Ioffe [22] on the perturbation stability of global metric regularity when the image space is not necessarily complete. We establish in section 4 a characterization of the local metric regularity. Based on this characterization, we derive the local version of the perturbation stability of metric regularity. As an application, we use this perturbation stability result to give a simple proof of a result on error bounds of

2-regular mappings established by Izmailov and Solodov [24] and extended by He and Sun [19].

2. Error bound in complete metric spaces. Let X be a metric space. Let $f : X \to \mathbb{R} \cup \{+\infty\}$ be a given function. As usual, $\mathrm{dom} f := \{x \in X : f(x) < +\infty\}$ denotes the domain of f. We set

$$S := \{x \in X : f(x) \leq 0\}. \tag{2.1}$$

We use the symbol $[f(x)]_+$ to denote $\max(f(x), 0)$. We shall say that the system (2.1) admits *an error bound* if there exists a real $c > 0$ such that

$$d(x, S) \leq c\big[f(x)\big]_+ \quad \text{for all } x \in X. \tag{2.2}$$

For $x_0 \in S$, we shall say that the system (2.1) has an error bound at x_0 when there exist real $c > 0$ and $\varepsilon > 0$ such that relation (2.2) is satisfied for all x around x_0, i.e., in an open ball $B(x_0, \varepsilon)$ with center x_0 and radius ε.

Several conditions using subdifferential operators or directional derivatives and ensuring the error bound in Banach spaces have been established, for example, in [10, 26, 37, 35, 40]. Recently, Azé [1] and Azé and Corvellec [5] have used the so-called strong slope introduced by De Giorgi, Marino, and Tosques in [12] to prove criteria for error bounds in complete metric spaces.

The following result, whose proof is strictly based on the Ekeland variational principle [17], gives an estimation for the distance $d(\bar{x}, S)$ from a given point $\bar{x}$ outside of S to the set S in complete metric spaces. Such an estimation using the Fréchet subdifferential in Asplund spaces has been established in [36].

THEOREM 2.1. *Let X be a complete metric space, and let $f : X \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous function and $\bar{x} \notin S$. Then, setting*

$$m(\bar{x}) := \inf \left\{ \sup_{y \in X, y \neq x} \frac{f(x) - [f(y)]_+}{d(x, y)} \;:\; \begin{array}{l} d(x, \bar{x}) < d(\bar{x}, S) \\ f(x) \leq f(\bar{x}) \end{array} \right\}, \tag{2.3}$$

one has

$$m(\bar{x}) d(\bar{x}, S) \leq f(\bar{x}). \tag{2.4}$$

Here and in what follows the convention $0.(+\infty) = 0$ is used.

Proof. As the conclusion holds trivially if $f(\bar{x}) = +\infty$, let us suppose that $\bar{x} \in \mathrm{dom} f$. If $S = \emptyset$, then, obviously, applying the Ekeland variational principle, one obtains $m(\bar{x}) = 0$. Hence, the conclusion follows from the convention $0.(+\infty) = 0$. Now, assume that $S \neq \emptyset$ and consider the function $g : X \to \mathbb{R} \cup \{+\infty\}$ defined by $g(x) = [f(x)]_+$. Let $\varepsilon \in (0, 1)$ be given. Obviously, since $\inf_{x \in X} g(x) = 0$, we can write

$$g(\bar{x}) = \inf_{x \in X} g(x) + \frac{g(\bar{x})}{(1 - \varepsilon) d(\bar{x}, S)} (1 - \varepsilon) d(\bar{x}, S).$$

By virtue of the Ekeland variational principle [17], we can select $z \in X$ satisfying $d(z, \bar{x}) \leq (1 - \varepsilon) d(\bar{x}, S)$ and $f(z) = g(z) \leq g(\bar{x}) = f(\bar{x})$ such that the function

$$g(\cdot) + \frac{g(\bar{x})}{(1 - \varepsilon) d(\bar{x}, S)} d(\cdot, z)$$

attains a minimum at z. That is,

$$[f(x)]_+ + \frac{f(\bar{x})}{(1-\varepsilon)d(\bar{x},S)}d(x,z) \geq f(z) \quad \text{for all } x \in X. \tag{2.5}$$

Since $d(z,\bar{x}) < d(\bar{x},S)$, thanks to the definition of $m(\bar{x})$, just select a point $y \in X, y \neq z$ such that

$$\frac{f(z)-[f(y)]_+}{d(z,y)} \geq m(\bar{x}) - \varepsilon.$$

Therefore, from (2.5), we obtain

$$\frac{f(\bar{x})}{(1-\varepsilon)d(\bar{x},S)} \geq \frac{f(z)-[f(y)]_+}{d(z,y)} \geq m(\bar{x}) - \varepsilon.$$

Taking the limit as ε goes to zero we obtain (2.4), establishing the proof. ∎

The following corollary of Theorem 2.1 gives characterizations of the error bound.

COROLLARY 2.2. *Let X be a complete metric space, and let $f : X \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous function. Let $\tau \in (0,+\infty)$; $r \in (0,+\infty]$ be given. Consider the following statements.*

(i) *$d(x,S) \leq \tau[f(x)]_+$ for all $x \in X$ with $f(x) < r$.*

(ii) *For each $x \in X \setminus S$ with $f(x) < r$ and for any $\varepsilon > 0$, there exists $z \in X$ such that*

$$0 < d(x,z) < (\tau+\varepsilon)(f(x) - [f(z)]_+). \tag{2.6}$$

(iii) *For each $x \in X \setminus S$ with $f(x) < r$ and for any $\varepsilon > 0$, there exists $z \in X$ with $f(z) \geq 0$ such that (2.6) holds.*

(iv) *For each $x \in X \setminus S$ with $f(x) < r$ and for any $\varepsilon > 0$, there exists $z \in X$ with $f(z) > 0$ such that (2.6) holds.*

Then, one has (i) ⇔ (ii) ⇐ (iii) ⇐ (iv). *In addition, if X is a Banach space and f is a continuous function, then all of the statements are equivalent.*

Proof. The implications (iv) ⇒ (iii) ⇒ (ii) are obvious. (ii) ⇒ (i) follows directly from Theorem 2.1.

Let $x \in X \setminus S$ with $f(x) < r$, and let $\varepsilon > 0$ be given. For (i) ⇒ (ii), take $z \in S$ such that $d(x,z) < (1+\varepsilon/\tau)d(x,S)$. Then, obviously, one has $d(x,z) \leq (\tau+\varepsilon)(f(x) - [f(z)]_+)$.

Now let X be a Banach space and f be a continuous function. For (ii) ⇒ (iii), let $z \in X$ such that (2.6) is satisfied. When $f(z) \leq 0$; since $f(x) > 0$ and the function f is continuous, we can find $y \in [x,z] := \{tx + (1-t)z : \ t \in [0,1]\}$ such that $f(y) = 0$. Hence,

$$0 < d(x,y) \leq d(x,z) < (\tau+\varepsilon)f(x) = (\tau+\varepsilon)(f(x) - f(y)).$$

Finally, for (iii) ⇒ (iv), let $z \in X$ with $f(z) \geq 0$ satisfying (2.6). If $f(z) > 0$, then the conclusion obviously holds. Suppose that $f(z) = 0$. Let $A \subseteq \mathbb{R}$ be defined by $A = \{t \in [0,1] : f(tx+(1-t)z) \leq 0\}$. Since A is nonempty, closed, and bounded in $\mathbb{R}$, we may define $\max A := t_0$ with $t_0 \in [0,1)$. For each $t \in (t_0,1)$ if $y_t := tx + (1-t)z$, then $f(y_t) > 0$. Pick a real $\delta > 0$ such that $\frac{1-t_0}{1-\delta(\tau+\varepsilon)} < 1$. Noticing that $f(y_{t_0}) = 0$ and using the continuity of f, we can find $t_1 \in (t_0,1)$ such that $f(y_t) < \delta d(x,z)$ for all $t \in (t_0,t_1)$. Then for all $t \in (t_0,t_1)$, one has

$$\begin{aligned} d(x,y_t) &= (1-t)d(x,z) < (1-t)(\tau+\varepsilon)f(x) \\ &< (1-t)(\tau+\varepsilon)(f(x) - f(y_t)) + (1-t)(\tau+\varepsilon)\delta d(x,z). \end{aligned}$$

Thus,

$$d(x, y_t) < \frac{(1-t)(\tau+\varepsilon)}{1-\delta(\tau+\varepsilon)}(f(x) - f(y_t)) < (\tau+\varepsilon)(f(x) - f(y_t)),$$

which completes the proof. □

Note that (iv) ⇒ (i) has been established by Wu and Ye in [40]. Similarly, we also obtain characterizations for the local error bound.

COROLLARY 2.3. *Let X be a complete metric space, and let $f : X \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous function. Let $\bar{x} \in S, \tau \in (0, +\infty)$ and $\eta \in (0, +\infty)$ be given. Consider the following statements.*

(i) *$d(x, S) \le \tau[f(x)]_+$ for all $x \in B(\bar{x}, \eta/2)$.*

(ii) *For each $x \in B(\bar{x}, \eta) \setminus S$ and for any $\varepsilon > 0$, there exists $z \in X$ such that*

$$0 < d(x, z) < (\tau+\varepsilon)(f(x) - [f(z)]_+). \tag{2.7}$$

(iii) *For each $x \in B(\bar{x}, \eta) \setminus S$ and for any $\varepsilon > 0$, there exists $z \in X$ with $f(z) \ge 0$ such that (2.7) holds.*

(iv) *For each $x \in B(\bar{x}, \eta) \setminus S$ and for any $\varepsilon > 0$, there exists $z \in X$ with $f(z) > 0$ such that (2.7) holds.*

Then, one has (iv) ⇒ (iii) ⇒ (ii) ⇒ (i). *Conversely, if* (i) *holds, then* (ii) *holds with $\eta/2$ instead of η. In addition, if X is a Banach space and f is a continuous function, then the three statements* (ii), (iii), *and* (iv) *are equivalent.*

Recall from [12, 5, 4] that the strong slope $|\nabla f|(x)$ of a lower semicontinuous function f at $x \in \mathrm{dom} f$ is the quantity defined by $|\nabla f|(x) = 0$ if x is a local minimum of f, otherwise,

$$|\nabla f|(x) = \limsup_{y \to x} \frac{f(x) - f(y)}{d(x, y)}.$$

For $x \notin \mathrm{dom} f$, we set $|\nabla f|(x) = +\infty$. Obviously, for all $\bar{x} \notin S$, one has

$$m(\bar{x}) \ge \inf\{|\nabla f|(x) : d(x, \bar{x}) < d(\bar{x}, S),\ f(x) \le f(\bar{x})\}.$$

Therefore, Theorem 2.1 implies directly the following result, which is also a corollary of Theorem 2.1 in Azé and Corvellec [5].

COROLLARY 2.4 (see [5, Theorem 2.1]). *Let X be a complete metric space and let $f : X \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous function. Let $\tau \in (0, +\infty)$ and $r \in (0, +\infty]$ be given. If there exists $m > 0$ such that $|\nabla f|(x) \ge m$ for all $x \in X \setminus S$ with $f(x) < r$, then*

$$m d(x, S) \le [f(x)]_+ \quad \text{for all} \quad x \in X; f(x) < r.$$

Let a real $\gamma > 0$ be given. By noting that for all $x \in X$ with $f(x) > 0$, $|\nabla f^\gamma|(x) = \gamma f^{\gamma-1}(x)|\nabla f|(x)$, one obtains the following error bound with exponent γ.

COROLLARY 2.5. *Let X be a complete metric space, and let $f : X \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous function. If there exists $m > 0$ such that*

$$\gamma f^{\gamma-1}(x)|\nabla f|(x) \ge m \quad \text{for all } x \in X \setminus S \text{ with } f(x) < r,$$

then

$$m d(x, S) \le [f(x)]_+^\gamma \quad \text{for all } x \in X; f(x) < r.$$

3. Perturbation stability of global metric regularity. Let X, Y be metric spaces and $F : X \rightrightarrows Y$ be a set-valued mapping. First, let us recall the notion of metric regularity (see, for example, [22, 21]).

DEFINITION 3.1. *Let $V \subseteq X \times Y$ be a given subset of $X \times Y$. The mapping F is said to be metrically regular on V with modulus $\tau > 0$ if*

$$d(x, F^{-1}(y)) \leq \tau d(y, F(x)) \quad \text{for all } (x, y) \in V. \tag{3.1}$$

It is worth pointing out that (see [22]) F is metrically regular with modulus τ if and only if F *covers* on V with a constant not smaller than $a = \tau^{-1}$. By the notion of covering we mean that for any $t > 0$ and any triplet (x, y, v) such that

$$(x, y) \in V, y \neq v \in F(x), d(y, v) < t,$$

there exists $u \in X$ such that $y \in F(u)$ and $ad(u, x) \leq t$.

The following theorem gives a characterization of the metric regularity on a subset V of the form

$$V = V(F, R) = \{(x, y) \in X \times Y : d(y, F(x)) < R\},$$

where $R \in (0, +\infty]$.

In what follows, we make use of the lower semicontinuous envelope $(x, y) \mapsto \varphi(x, y)$ of the function $(x, y) \mapsto d(y, F(x))$, i.e., for each $(x, y) \in X \times Y$,

$$\varphi(x, y) := \liminf_{(u,v)\to(x,y)} d(v, F(u)) = \liminf_{u\to x} d(y, F(u)).$$

THEOREM 3.2. *Let X be a complete metric space, and let Y be a metric space, (which is not necessarily complete). Let $F : X \rightrightarrows Y$ be a set-valued mapping with a closed graph and $V := V(F, R)$ for some $R \in (0, +\infty]$. Then the following statements are equivalent*:

(i) *For all $(x, y) \in V$, one has*

$$d(x, F^{-1}(y)) \leq \tau d(y, F(x)).$$

(ii) *Let $(x, y) \in X \times Y$ with $y \notin F(x)$; $\varphi(x, y) < R$ and let $\varepsilon > 0$. Then for every sequence $\{x_n\}_{n\in\mathbb{N}} \subseteq X$ converging to x, there exists a sequence $\{u_n\}_{n\in\mathbb{N}} \subseteq X$ with $\lim_{n\to\infty} d(u_n, x) > 0$ such that*

$$\limsup_{n\to\infty} \frac{d(y, F(x_n)) - d(y, F(u_n))}{d(x_n, u_n)} > \frac{1}{\tau + \varepsilon}. \tag{3.2}$$

(iii) *Let $(x, y) \in X \times Y$ with $y \notin F(x)$; $\varphi(x, y) < R$ and let $\varepsilon > 0$. Then, for every sequence $\{x_n\}_{n\in\mathbb{N}} \subseteq X$ converging to x with*

$$\lim_{n\to\infty} d(y, F(x_n)) = \liminf_{u\to x} d(y, F(u)),$$

there exists a sequence $\{u_n\}_{n\in\mathbb{N}} \subseteq X$ with $\lim_{n\to\infty} d(u_n, x) > 0$ such that (3.2) *holds.*

Proof. (ii) $\Rightarrow$ (iii) is obvious. For (i) $\Rightarrow$ (ii), take some $\varepsilon > 0$ and $(x, y) \in X \times Y$ with $y \notin F(x)$; $\varphi(x, y) < R$. Let $\{x_n\}_{n\in\mathbb{N}} \subseteq X$ be a sequence converging to x. We

consider the following two cases.

Case 1. $\limsup_{n\to\infty} d(y, F(x_n)) < R$. Since gph F is closed when n is sufficiently large, say $n \geq n_0$, hence, $y \notin F(x_n)$ and, moreover, $d(y, F(x_n)) < R$. Hence $d(x_n, F^{-1}(y)) \leq \tau d(y, F(x_n))$. For each integer n, pick $u_n \in F^{-1}(y)$ such that $d(x_n, u_n) < (1 + \varepsilon/\tau) d(x_n, F^{-1}(y))$. By relabeling if necessary, we can assume that $\lim d(x, u_n)$ exists. Then using the closedness of gph F, we have $\lim d(x, u_n) > 0$ and for all $n \geq n_0$,

$$d(x_n, u_n) < (1 + \varepsilon/\tau) d(x_n, F^{-1}(y)) \leq (\tau + \varepsilon)\big[d(y, F(x_n)) - d(y, F(u_n))\big].$$

This shows that (3.2) holds.

Case 2. $\limsup_{n\to\infty} d(y, F(x_n)) \geq R$. Let $\{z_n\}_{n\in\mathbb{N}} \subseteq X$ be a sequence converging to x such that

$$\lim_{n\to\infty} d(y, F(z_n)) = \varphi(x, y) (< R).$$

By using the argument of Case 1 applied to $\{z_n\}_{n\in\mathbb{N}}$ instead of $\{x_n\}_{n\in\mathbb{N}}$, we can find a sequence $\{u_n\}_{n\in\mathbb{N}}$ with $\lim d(x, u_n) > 0$ and some $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$, one has

$$d(z_n, u_n) \leq (\tau + \varepsilon)\Big[d(y, F(z_n)) - d(y, F(u_n))\Big].$$

Hence

$$\begin{aligned}\limsup_{n\to\infty} \frac{d(y, F(x_n)) - d(y, F(u_n))}{d(x_n, u_n)} &\geq \limsup_{n\to\infty} \frac{d(y, F(z_n)) - d(y, F(u_n))}{d(x_n, u_n)} \\ &> \frac{1}{\tau + \varepsilon} \limsup_{n\to\infty} \frac{d(z_n, u_n)}{d(x_n, u_n)} = \frac{1}{\tau + \varepsilon}.\end{aligned}$$

Thus, relation (3.2) holds.

Let us prove (iii) $\Rightarrow$ (i). Let $(x, y) \mapsto \varphi(x, y)$ denote the lower semicontinuous envelope of the function $(x, y) \mapsto d(y, F(x))$. Then, $0 \leq \varphi(x, y) \leq d(y, F(x))$ for all $(x, y) \in X \times Y$. Observe that for each $y \in Y$, $F^{-1}(y) = \{x \in X : \ \varphi(x, y) = 0\}$. Indeed, let $(x, y) \in X \times Y$. Obviously, if $x \in F^{-1}(y)$, then $\varphi(x, y) = 0$. Conversely, suppose $\varphi(x, y) = 0$. There exists a sequence $\{x_n\}_{n\in\mathbb{N}}$ with limit x such that $d(y, F(x_n))$ converges to 0. Then, we can find a sequence $\{z_n\}_{n\in\mathbb{N}} \subseteq Y$ such that $z_n \in F(x_n)$ and $d(y, z_n) \to 0$. Since the graph of F is closed, then $(x, y) \in$ gph F, i.e., $x \in F^{-1}(y)$. From this and by virtue of Corollary 2.2, it suffices to show that for each $(x, y) \in X \times Y$, with $y \notin F(x)$; $\varphi(x, y) < R$ and for any $\varepsilon \in (0, 1)$ there exists $u \in X$ with $u \neq x$ such that

$$d(x, u) \leq (\tau + \varepsilon)(\varphi(x, y) - \varphi(u, y)).$$

To see this, let $\{x_n\}_{n\in\mathbb{N}} \subseteq X$ be a sequence with limit x and $d(y, F(x_n)) \to \varphi(x, y)$ as $n \to \infty$. Then $(x_n, y) \in V$ and $x_n \notin F^{-1}(y)$ when n is sufficiently large, say $n \geq n_0$. Let $\{u_n\}_{n\in\mathbb{N}}$ be a sequence of elements in X satisfying (3.2) with respect to the sequence $\{x_n\}_{n\in\mathbb{N}}$. Pick $\delta \in (0, \lim d(u_n, x))$. Then there exists an index $n_1 \geq n_0$ such that for all $n \geq n_1$, one has $d(x_n, u_n) \geq \delta$; $d(x_n, x) < \epsilon\delta$;

$$d(y, F(x_n)) < \varphi(x, y) + \frac{\varepsilon}{\tau + \varepsilon} d(x_n, u_n)$$

and

$$d(x_n, u_n) < (\tau + \varepsilon)(d(y, F(x_n)) - d(y, F(u_n))).$$

Hence,

$$d(x_n, u_n) < (1-\varepsilon)^{-1}(\tau+\varepsilon)(\varphi(x,y) - \varphi(u_n, y)).$$

It follows that for $n \geq n_1$,

$$d(x, u_n) \leq (1+\varepsilon)d(x_n, u_n) < (1-\varepsilon)^{-1}(\tau+\varepsilon)(1+\varepsilon)(\varphi(x,y) - \varphi(u_n, y)).$$

As $\varepsilon \in (0,1)$ is arbitrary, the proof is complete. □

Based on Theorem 3.2, we can now prove an extension of a result by Ioffe [22, Theorem 2] on perturbation stability of metric regularity. Precisely, we extend Ioffe's result to the case where the image space Y is not necessarily complete. It is worth noting that the following proof is more simple than the one given in [22] that is based on a criterion of metric regularity [22, Theorem 6] which depends heavily on the completeness of Y.

Let X be a metric space and Y be a normed linear space. Let $F, \Phi : X \rightrightarrows Y$ be two set-valued mappings. For $(x,r) \in X \times (0, +\infty)$, let us denote the following quantity introduced by Ioffe [22] by

$$\sigma_{F,\Phi}(x,r) := \sup_{\eta \in \Phi(x)} \inf_{v \in F(x)} \sup_{d(u,x)<r, w \in F(u)} \inf_{\xi \in \Phi(u)} \|\eta - v + w - \xi\|. \tag{3.3}$$

Note that when $\Phi(x) := F(x) + G(x)$, where $G : X \rightrightarrows Y$ is a set-valued mapping, then (see [22])

$$\sigma_{F,\Phi}(x,r) \leq \sup_{d(x,u)<r} e(G(x), G(u)) := \sup_{d(x,u)<r} \sup_{y \in G(x)} d(y, G(u)) \quad \text{for all } (x,u) \in X \times X. \tag{3.4}$$

THEOREM 3.3. *Let X be a complete metric space and Y be a normed linear space. Let $F, \Phi : X \rightrightarrows Y$ be set-valued mappings with closed graphs. Suppose that F is metrically regular on $V(F,R)$ for some $R \in (0, +\infty]$ with modulus $\tau > 0$. If there exists $\lambda \in (0, \tau^{-1})$ such that*

$$\sigma_{F,\Phi}(x,r) \leq \lambda r \quad \text{for all } x \in X,\ r \in (0, \tau R), \tag{3.5}$$

then Φ is metrically regular on $V(\Phi, R)$ with modulus $(\tau^{-1} - \lambda)^{-1}$.

Proof. It suffices to show that statement (iii) in Theorem 3.2 is verified. Indeed, let $(x,y) \in X \times Y$ with $y \notin \Phi(x)$; $\liminf_{u \to x} d(y, \Phi(u)) < R$. Let $\{x_n\}_{n \in \mathbb{N}} \subseteq X$ be a sequence converging to x with $\lim_{n\to\infty} d(y, \Phi(x_n)) = \liminf_{u \to x} d(y, \Phi(u))$. Then $(x_n, y) \in V(\Phi, R)$ and $y \notin \Phi(x_n)$ for large n, say $n \geq n_0$. For each $n \geq n_0$, let $\delta_n \in (0, \varepsilon)$ sufficiently small such that $(1+\delta_n)d(y, \Phi(x_n)) < R$ and $\delta_n \to 0$ as $n \to \infty$. Take $\eta_n \in \Phi(x_n)$ such that

$$\|y - \eta_n\| < (1+\delta_n) d(y, \Phi(x_n)).$$

Set $r_n = (1+\delta_n)\tau d(y, \Phi(x_n)) (< \tau R)$. By (3.5), we can find $v_n \in F(x_n)$ such that

$$\sup_{d(u,x_n)<r_n, w \in F(u)} \inf_{\xi \in \Phi(u)} \|\eta_n - v_n + w - \xi\| < (1+\delta_n)\lambda r_n. \tag{3.6}$$

Setting $z_n := y - \eta_n + v_n$, then, obviously, $(x_n, z_n) \in V(F,R)$. Moreover, since assumption F is metrically regular on $V(F,R)$ with modulus τ, one has

$$d(x_n, F^{-1}(z_n)) \leq \tau d(z_n, F(x_n)) \leq \tau \|z_n - v_n\| = \tau \|y - \eta_n\| < (1+\delta_n)\tau d(y, \Phi(x_n)) := r_n.$$

Therefore, there exists $u_n \in F^{-1}(z_n)$ such that $d(x_n, u_n) < r_n$. Hence, relation (3.6) implies that

$$\inf_{\xi \in \Phi(u_n)} \|\eta_n - v_n + z_n - \xi\| < (1+\delta_n)\lambda r_n,$$

that is, $d(y, \Phi(u_n)) < (1+\delta_n)\lambda r_n$. It follows that

$$\limsup_{n\to\infty} d(y, \Phi(u_n)) \le \lambda\tau \liminf_{u\to x} d(y, \Phi(u)) < \liminf_{u\to x} d(y, \Phi(u)),$$

and, consequently, $\liminf_{n\to\infty} d(u_n, x_n) > 0$. Moreover,

$$\begin{aligned} d(y, \Phi(x_n)) - d(y, \Phi(u_n)) &> r_n[(1+\delta_n)^{-1}\tau^{-1} - (1+\delta_n)\lambda] \\ &> d(x_n, u_n)[(1+\delta_n)^{-1}\tau^{-1} - (1+\delta_n)\lambda]. \end{aligned}$$

By letting $n \to \infty$, we obtain

$$\limsup_{n\to\infty} \frac{d(y, \Phi(x_n)) - d(y, \Phi(u_n))}{d(x_n, u_n)} \ge \tau^{-1} - \lambda.$$

This completes the proof. □

Recall that a set-valued mapping $G : X \rightrightarrows Y$ is said to be Lipschitz on X with a constant $\lambda > 0$ if

$$e(G(x), G(u)) \le \lambda d(x, u) \quad \text{for all } (x, u) \in X \times X, \tag{3.7}$$

where, $e(G(x), G(u)) = \sup_{y\in G(x)} d(y, G(u))$. If for some $\bar{x} \in X$ (3.7) holds for all $(x, u) \in B(\bar{x}, \delta) \times B(\bar{x}, \delta)$ for some $\delta > 0$, then we say that G is Lipschitzian around $\bar{x}$.

By relation (3.4), Theorem 3.3 yields the following corollary (see [22, Corollary 3]).

COROLLARY 3.4. *Let X be a complete metric space, and let Y be a normed linear space. Let $F, G : X \rightrightarrows Y$ be set-valued mappings such that both F and $\Phi := F + G$ have closed graphs. Suppose that F is metrically regular on $V(F, R)$ with modulus $\tau > 0$ and G is Lipschitz on X with constant $\lambda \in (0, \tau^{-1})$. Then, Φ is metrically regular on $V(\Phi, R)$ with modulus $(\tau^{-1} - \lambda)^{-1}$.*

4. Perturbation stability of local metric regularity. In this section, we consider the local case of perturbation stability of metric regularity. Let $(\bar{x}, \bar{y}) \in \operatorname{gph} F$. The mapping F is said to be metrically regular at $\bar{x}$ with respect to $\bar{y}$ with modulus $\tau \in (0, +\infty)$ if there exist neighborhoods U of $\bar{x}$ and V of $\bar{y}$ such that

$$d(x, F^{-1}(y)) \le \tau d(y, F(x)) \quad \text{for all } (x, y) \in U \times V. \tag{4.1}$$

First, we prove a characterization of the local metric regularity, which is a local version of Theorem 3.2. Similar to section 2, let $\varphi(\cdot, \cdot)$ denote the lower semicontionuous envelope function of the mapping $d(\cdot, F(\cdot))$.

THEOREM 4.1. *Let X be a complete metric space, and let Y be a metric space. Let $F : X \rightrightarrows Y$ be a set-valued mapping with a closed graph, and let $(\bar{x}, \bar{y}) \in \operatorname{gph} F$ and $\tau \in (0, +\infty)$ be given. Then the following statements are equivalent:*

(i) *There exists a neighborhood $U \times V \subseteq X \times Y$ of $(\bar{x}, \bar{y})$ and such that*

$$d(x, F^{-1}(y)) \le \tau d(y, F(x)) \quad \textit{for all } (x, y) \in U \times V;$$

(ii) *There exists a neighborhood* $U \times V \subseteq X \times Y$ *of* $(\bar{x}, \bar{y})$ *such that for any* $(x, y) \in U \times V$ *with* $y \notin F(x)$, *any* $\varepsilon > 0$, *and any sequence* $\{x_n\}_{n\in\mathbb{N}} \subseteq X$ *converging to* x *with*

$$\limsup_{n\to\infty} d(y, F(x_n)) \leq d(y, F(x)),$$

there exists a sequence $\{u_n\}_{n\in\mathbb{N}} \subseteq X$ *with* $\lim_{n\to\infty} d(u_n, x) > 0$ *such that*

$$\limsup_{n\to\infty} \frac{d(y, F(x_n)) - d(y, F(u_n))}{d(x_n, u_n)} > \frac{1}{\tau + \varepsilon}. \tag{4.2}$$

(iii) *There exist a neighborhood* $U \times V \subseteq X \times Y$ *of* $(\bar{x}, \bar{y})$ *and a real* $\gamma \in (0, +\infty)$ *such that for any* $(x, y) \in U \times V$ *with* $y \notin F(x)$ *and* $\varphi(x, y) < \gamma$ *and any* $\varepsilon > 0$, *then for any sequence* $\{x_n\}_{n\in\mathbb{N}} \subseteq X$ *converging to* x *with*

$$\lim_{n\to\infty} d(y, F(x_n)) = \liminf_{u\to x} d(y, F(u)),$$

we can find a sequence $\{u_n\}_{n\in\mathbb{N}} \subseteq X$ *with* $\lim_{n\to\infty} d(u_n, x) > 0$ *such that* (4.2) *holds.*

Proof. (ii) $\Rightarrow$ (iii) is obvious, while (i) $\Rightarrow$ (ii) is similar to the proof of the respective implication in Theorem 3.2. It remains to prove (iii) $\Rightarrow$ (i). Let us remind the reader that

$$\varphi(x, y) = \liminf_{u\to x} d(y, F(u)), (x, y) \in X \times Y.$$

Then, for all $y \in Y$, $F^{-1}(y) = \{x \in X : \varphi(x, y) = 0\}$. Let $U \times V := B(\bar{x}, \alpha) \times B(\bar{y}, \beta)$ and τ, γ as in (iii). Let $\varepsilon \in (0, \tau/2)$ be given and set $\delta = \min\{\alpha, \frac{\alpha}{6(\tau+2\varepsilon)}, \beta, \gamma/4\}$. If y is fixed in $B(\bar{y}, \delta)$, then

$$\varphi(\bar{x}, y) \leq d(y, F(\bar{x})) \leq d(y, \bar{y}) < \delta.$$

Hence

$$\varphi(\bar{x}, y) \leq \inf_{x\in X} \varphi(x, y) + \delta.$$

By virtue of the Ekeland variational principle [17] (applied to the function $x \mapsto \varphi(x, y)$), we can select $z \in X$ satisfying $d(\bar{x}, z) \leq \delta(\tau + 2\varepsilon)$ and $\varphi(z, y) \leq \varphi(\bar{x}, y) (< \delta)$ such that

$$\varphi(z, y) \leq \varphi(x, y) + \frac{1}{\tau + 2\epsilon} d(x, z) \quad \text{for all } x \in X.$$

It follows that for any sequence $\{z_n\}_{n\in\mathbb{N}} \subseteq X$ converging to z with

$$\lim_{n\to\infty} d(y, F(z_n)) = \liminf_{u\to z} d(y, F(u)),$$

for all $\{u_n\}_{n\in\mathbb{N}} \subseteq X$ with $\lim_{n\to\infty} d(u_n, z) > 0$, one always has

$$\limsup_{n\to\infty} \frac{d(y, F(z_n)) - d(y, F(u_n))}{d(z_n, u_n)} \leq \limsup_{n\to\infty} \frac{\varphi(z, y) - \varphi(u_n, y)}{d(z, u_n)} \leq \frac{1}{\tau + 2\varepsilon} < \frac{1}{\tau + \varepsilon}.$$

Therefore, by assumption we must have $z \in F^{-1}(y)$. Consequently, $B(\bar{x}, 2\delta\tau) \cap F^{-1}(y) \neq \emptyset$.

Now take $(x,y) \in B(\bar{x}, 2\delta\tau) \times B(\bar{y}, \delta)$. We distinguish two cases.

Case 1. $d(y, F(x)) \geq \gamma$. Since $B(\bar{x}, 2\delta\tau) \cap F^{-1}(y) \neq \emptyset$, then

$$d(x, F^{-1}(y)) \leq d(x, \bar{x}) + d(\bar{x}, F^{-1}(y)) < 2\delta\tau + 2\delta\tau = 4\delta\tau \leq \tau d(y, F(x)). \tag{4.3}$$

Case 2. $d(y, F(x)) < \gamma$. Let $z \in X$ with $d(x,z) < d(x, F^{-1}(y))$; $\varphi(z,y) \leq \varphi(x,y)$. One has

$$d(z, \bar{x}) \leq d(z,x) + d(x, \bar{x}) \leq d(\bar{x}, F^{-1}(y)) + 2d(x, \bar{x}) < 6\delta\tau.$$

Thus, $z \in B(x, \alpha)$. Let $\{z_n\}_{n\in\mathbb{N}} \subseteq X$ be an arbitrary sequence converging to z with

$$\lim_{n\to\infty} d(y, F(z_n)) = \varphi(z,y) (< \gamma).$$

According to (iii), we can find a sequence $\{u_n\}_{n\in\mathbb{N}} \subseteq X$ with $\lim_{n\to\infty} d(u_n, z) > 0$ such that

$$\limsup_{n\to\infty} \frac{d(y, F(z_n)) - d(y, F(u_n))}{d(z_n, u_n)} > \frac{1}{\tau + \varepsilon}.$$

It follows that

$$\limsup_{n\to\infty} \frac{\varphi(z,y) - \varphi(u_n, y)}{d(z, u_n)} > \frac{1}{\tau + \varepsilon}. \tag{4.4}$$

Consequently,

$$m(x) := \inf \left\{ \sup_{u\in X, u\neq z} \frac{\varphi(z,y) - \varphi(u,y)}{d(z,u)} : \begin{array}{l} d(z,x) < d(x, F^{-1}(y)) \\ \varphi(z,y) \leq \varphi(x,y) \end{array} \right\} > \frac{1}{\tau + \varepsilon}.$$

By virtue of Theorem 2.1 and as ε is arbitrarily small, we obtain

$$d(x, F^{-1}(y)) \leq \tau d(y, F(x)) \quad \text{for all } (x,y) \in B(\bar{x}, 2\delta\tau) \times B(\bar{y}, \delta),$$

where $\delta = \min\{\alpha, \alpha\tau^{-1}/6, \beta, \gamma/4\}$. The proof is complete. □

REMARK 4.2. (i) *In* [16], *Dontchev and Lewis have established a characterization of the local metric regularity under the assumptions that the function* $(x,y) \mapsto d(x, F(y))$ *is locally lower semicontinuous (and both* X, Y *are complete). Note that, when this condition is satisfied, Theorem* 4.1 *yields directly Lemma* 1.7 *in* [16].

(ii) *When both* X, Y *are complete metric spaces, a characterization using the strong slope has been established by Azé–Corvellec in* [5, Theorem 5.3].

REMARK 4.3. *Obviously,* (i), (ii), (iii) *are equivalent to the following*:

(iv) *There exist a neighborhood* $U \times V \subseteq X \times Y$ *of* $(\bar{x}, \bar{y})$ *and a real* $\tau \in (0, +\infty)$ *such that for any* $(x,y) \in V$ *with* $y \notin F(x)$ *and any* $\varepsilon > 0$, *then for all sequences* $\{x_n\}_{n\in\mathbb{N}} \subseteq X$ *converging to* x *with*

$$\lim_{n\to\infty} d(y, F(x_n)) = \liminf_{u\to x} d(y, F(u)),$$

there exists a sequence $\{u_n\}_{n\in\mathbb{N}} \subseteq X$ *with* $\lim_{n\to\infty} d(u_n, x) > 0$ *such that* (4.2) *holds.*

Now let X be a complete metric space, and let Y be a normed linear space. Let $F, \Phi : X \rightrightarrows Y$ be set-valued mappings and $(\bar{x}, \bar{y}) \in \text{gph } F \cap \text{gph } \Phi$ be given. To study the perturbation stability of local metric regularity, instead of $\sigma_{F,\Phi}(x, r)$ used in the global case, we will use the following quantity:

(4.5)
$$\sigma_{F,\Phi}(x, t_1, t_2, r) := \sup_{\eta \in \Phi(x) \cap B(\bar{y}, t_1)} \inf_{v \in F(x) \cap B(\bar{y}, t_2)} \sup_{d(u,x) < r, w \in F(u)} \inf_{\xi \in \Phi(u)} \|\eta - v + w - \xi\|,$$

with $x \in X$; $t_1, t_2, r \in (0, +\infty)$.

THEOREM 4.4. *Let X be a complete metric space and Y be a normed linear space. Let $F, \Phi : X \rightrightarrows Y$ be set-valued mappings with closed graphs. Let $(\bar{x}, \bar{y}) \in \text{gph } F \cap \text{gph } \Phi$ be given. Suppose that there exist reals $\alpha, \beta > 0$ and $\tau > 0$ such that*

$$d(x, F^{-1}(y)) \leq \tau d(y, F(x)) \quad \text{for all } (x, y) \in B(\bar{x}, \alpha) \times B(\bar{y}, \beta). \tag{4.6}$$

If there exist positive reals $t_1, t_2, s, \lambda, \delta$ with $t_2 < (0, \beta), \lambda \in (0, \tau^{-1})$ such that

$$\sigma_{F,\Phi}(x, t_1, t_2, r) \leq \lambda r \quad \text{for all } x \in B(\bar{x}, \delta),\ r \in (0, s), \tag{4.7}$$

then Φ is metrically regular around $\bar{x}$ with respect to $\bar{y}$ with modulus $(\tau^{-1} - \lambda)^{-1}$.

Proof. It suffices to show that statement (iii) of Theorem 4.1 applies to the mapping Φ around $(\bar{x}, \bar{y})$. Set

$$\gamma = \min\{t_1/2, \beta - t_2, \delta\tau^{-1}, s\tau^{-1}\}; a := \min\{\alpha, \delta\}; b := \min\{\beta, t_1 - \gamma\}.$$

Let $(x, y) \in B(\bar{x}, a) \times B(\bar{y}, b)$ with $y \notin \Phi(x)$ and $\liminf_{u \to x} d(y, \Phi(u)) < \gamma$. Let $\{x_n\}_{n \in \mathbb{N}}$ be a sequence in X such that $d(x_n, x) \to 0$ and $d(y, \Phi(x_n)) \to \liminf_{u \to x} d(y, \Phi(u))$. Without loss of generality, we can assume that $x_n \in B(\bar{x}, a)$ and $d(y, \Phi(x_n)) < \gamma$ for all $n \in \mathbb{N}$. Pick a sequence $\{\varepsilon_n\}_{n \in \mathbb{N}}$ of positive reals converging to zero and satisfying $(1 + \varepsilon_n) d(y, \Phi(x_n)) < \gamma$ for all $n \in \mathbb{N}$. For each integer n take $\eta_n \in \Phi(x_n)$ such that

$$\|y - \eta_n\| < (1 + \varepsilon_n) d(y, \Phi(x_n)). \tag{4.8}$$

Then,

$$\|\eta_n - \bar{y}\| \leq \|\eta_n - y\| + \|y - \bar{y}\| < \gamma + b \leq t_1.$$

If $r_n := (1 + \varepsilon_n)\tau d(y, \Phi(x_n))$, then $r_n \in (0, s)$. Therefore by (4.7), for each n there exists $v_n \in F(x_n) \cap B(\bar{y}, t_2)$ such that

$$\sup_{d(u,x) < r_n, w \in F(u)} \inf_{\xi \in \Phi(u)} \|\eta_n - v_n + w - \xi\| < (1 + \varepsilon_n)\lambda r_n. \tag{4.9}$$

Set $z_n := y - \eta_n + v_n$. Then

$$\|z_n - \bar{y}\| \leq \|y - \eta_n\| + \|v_n - \bar{y}\| < \gamma + t_2 \leq \beta,$$

that is, $z_n \in B(\bar{y}, \beta)$. According to relation (4.6), we can select $u_n \in F^{-1}(z_n)$ such that

$$\begin{aligned} d(x_n, u_n) &\leq (1 + \varepsilon_n)\tau d(z_n, F(x_n)) \leq (1 + \varepsilon_n)\tau \|z_n - v_n\| \\ &< (1 + \varepsilon_n)\tau d(y, \Phi(x_n)) := r_n < \tau\gamma \leq s. \end{aligned}$$

Therefore, by (4.9),

$$\inf_{\xi\in\Phi(u_n)} \|\eta_n - v_n + z_n - \xi\| < (1+\varepsilon_n)\lambda r_n,$$

i.e., $d(y,\Phi(u_n)) < (1+\varepsilon_n)\lambda r_n$. Next, similar to the argument developed in the proof of Theorem 3.3, we obtain $\lim_{n\to\infty} d(x,u_n) > 0$ and that

$$\limsup_{n\to\infty} \frac{d(y,\Phi(x_n)) - d(y,\Phi(u_n))}{d(x_n,u_n)} > \tau^{-1} - \lambda.$$

By virtue of Theorem 4.1 (and its proof), we derive

$$d(x,\Phi^{-1}(y)) \le \tau d(y,\Phi(x)) \quad \text{for all } (x,y)\in B(\bar{x},2c\tau)\times B(\bar{y},c),$$

where $c = \min\{a, a\tau^{-1}/6, b, \gamma/4\}$. This completes the proof. □

The following corollary generalizes a result established by Dontchev, Lewis, and Rockafellar [15, Theorem 3.3].

COROLLARY 4.5. *Let X be a complete metric space, and let Y be a normed linear space. Let $F, G : X \rightrightarrows Y$ be set-valued mappings such that both F and $\Phi := F + G$ have closed graphs. Let $(\bar{x},\bar{y}) \in \operatorname{gph} F$ and suppose that G is single-valued at $\bar{x}$ with $G(\bar{x}) := \bar{z}$. If F is metrically regular at $\bar{x}$ with respect to $\bar{y}$ with modulus $\tau > 0$ and G is locally Lipschitz around $\bar{x}$ with constant $\lambda \in (0,\tau^{-1})$, then Φ is metrically regular at $\bar{x}$ with respect to $\bar{y}+\bar{z}$ with modulus $(\tau^{-1}-\lambda)^{-1}$.*

Proof. By translation, if necessary, without loss of generality, we can assume that $G(\bar{x}) = 0$. That is, $(\bar{x},\bar{y}) \in \operatorname{gph} F \cap \operatorname{gph} \Phi$. Let $\alpha, \beta > 0$ such that

$$d(x,F^{-1}(y)) \le \tau d(y,F(x)) \quad \text{for all } (x,y)\in B(\bar{x},\alpha)\times B(\bar{y},\beta).$$

Let $\delta \in (0, \beta\lambda^{-1}/4)$ be such that G is Lipschitz on $B(\bar{x},\delta)$ with constant λ, i.e.,

$$\sup_{y\in G(x)} \inf_{u\in G(u)} \|y-u\| \le \lambda d(x,u) \quad \text{for all } x,u \in B(\bar{x},\delta).$$

By considering $u := \bar{x}$, we get $\sup_{y\in G(x)} \|y\| \le \lambda d(x,\bar{x}) < \lambda\delta$ for all $x \in B(\bar{x},\delta)$. Take $x \in B(\bar{x},\delta/2)$ and $r \in (0,\delta/2)$. For any $\eta := \eta_1 + y \in \Phi(x) \cap B(\bar{y},\beta/4)$, with $\eta_1 \in F(x), y \in G(x)$, then

$$\|\eta_1 - \bar{y}\| \le \|y\| + \|\eta - \bar{y}\| < \lambda\delta + \beta/4 < \beta/2.$$

Hence, $\eta_1 \in F(x) \cap B(\bar{y},\beta/2)$ and by virtue of the definition of $\sigma_{F,\Phi}(x,\beta/4,\beta/2,r)$, we obtain

$$\begin{aligned}\sigma_{F,\Phi}(x,\beta/4,\beta/2,r) &\le \sup_{y\in G(x)} \sup_{d(u,x)<r, w\in F(u)} \inf_{\xi\in\Phi(u)} \|y+w-\xi\| \\ &\le \sup_{y\in G(x)} \sup_{d(u,x)<r} \inf_{p\in G(u)} \|y-p\| \le \lambda r.\end{aligned}$$

Thus, all assumptions of Theorem 4.4 are satisfied with $t_1 = \beta/4$; $t_2 = \beta/2$; and $s = \delta/2$. This shows that Φ is metrically regular around $\bar{x}$ with respect to $\bar{y}$ with modulus $(\tau^{-1}-\lambda)^{-1}$. □

REMARK 4.6. *From the proofs of Theorem 4.4 and the preceding corollary, we see that if F is globally metrically regular (that is, $\alpha = \beta = +\infty$ in Theorem 4.5) with modulus τ, then there exists a constant $c > 0$ depending only on τ such that for any*

locally Lipschitzian mapping $G : X \rightrightarrows Y$ on $B(\bar{x}, \delta)$ with constant $\lambda \in (0, \tau^{-1})$ and $G(\bar{x}) := \bar{z}$, one has

$$d(x, \Phi^{-1}(y)) \leq d(y, \Phi(x)) \quad \text{for all } (x, y) \in B(\bar{x}, 2c\tau\delta) \times B(\bar{y} + \bar{z}, c\delta),$$

where $\Phi := F + G$.

Next, we give the following perturbation stability result using the quantity $\sigma_{F,\Phi}(x, r)$ as in the global perturbation stability (section 3), but with an additional suitable condition.

THEOREM 4.7. *Let X be a complete metric space and Y be a normed linear space. Let $F, \Phi : X \rightrightarrows Y$ be set-valued mapping with closed graphs. Let $(\bar{x}, \bar{y}) \in \text{gph } F$ and $(\bar{x}, \bar{z}) \in \text{gph } \Phi$ be given. Suppose that F is metrically regular with modulus $\tau > 0$ and that the following two conditions are satisfied.*

(i) *There exist positive reals s, λ, δ with $\lambda \in (0, \tau^{-1})$ such that*

$$\sigma_{F,\Phi}(x, , r) \leq \lambda r \quad \text{for all } x \in B(\bar{x}, \delta), r \in (0, s); \tag{4.10}$$

(ii) $\lim_{x \to \bar{x}} e(F(x) - \bar{y}, \Phi(x) - \bar{z}) = 0$, *where* $e(F(x) - \bar{y}, \Phi(x) - \bar{z}) = \sup_{u \in F(x) - \bar{y}} d(u, \Phi(x) - \bar{z})$.

Then Φ is metrically regular around $\bar{x}$ with respect to $\bar{z}$ with modulus $(\tau^{-1} - \lambda)^{-1}$.

Proof. By translation, considering $\Phi + \bar{y} - \bar{z}$ instead of Φ, we can assume that $\bar{z} = \bar{y}$. Let $\alpha, \beta > 0$ such that

$$d(x, F^{-1}(y)) \leq \tau d(y, F(x)) \quad \text{for all } (x, y) \in B(\bar{x}, \alpha) \times B(\bar{y}, \beta).$$

By (ii), we can find $\delta_1 \in (0, \delta/2)$ such that

$$e(F(x), \Phi(x)) < \beta/4 \quad \text{for all } x \in B(\bar{x}, \delta_1). \tag{4.11}$$

Set

$$\gamma = \min\{\delta_1/2, \beta/4, s\tau^{-1}, \beta\alpha^{-1}/4\}; a = \min\{\alpha, \delta_1/2\}; b = \beta/4.$$

Similar to the proof of Theorem 4.4, it suffices to show that statement (iii) of Theorem 4.1 is satisfied for the mapping Φ around $(\bar{x}, \bar{y})$. Indeed, let $(x, y) \in B(\bar{x}, a) \times B(\bar{y}, b)$ with $y \notin \Phi(x)$ and $d(y, \Phi(x)) < \gamma$. Let $\{x_n\}_{n \in \mathbb{N}}$; $\{\varepsilon_n\}_{n \in \mathbb{N}}$; $\{r_n\}_{n \in \mathbb{N}}$; $\{\eta_n\}_{n \in \mathbb{N}}$; $\{v_n\}_{n \in \mathbb{N}}$; $\{z_n\}_{n \in \mathbb{N}}$ as in the proof of Theorem 4.4. Then, by relations (4.10) and (4.11), one has

$$\|\eta_n - v_n\| < \sup_{d(u,x) < r_n, w \in F(u)} \inf_{\xi \in \Phi(u)} \|w - \xi\| + (1 + \varepsilon_n)\lambda r_n < \beta/4 + \beta/4 = \beta/2.$$

Consequently, $z_n := y - \eta_n + v_n \in B(\bar{y}, \beta)$. We conclude as in the proof of Theorem 4.4. □

Noticing that if $G : X \rightrightarrows Y$ is locally Lipschitz around $\bar{x}$, then (ii) holds trivially for $\Phi := F + G$. We obtain the following corollary, where, the assumption that G is single-valued at $\bar{x}$ in Corollary 4.5 can be removed.

COROLLARY 4.8. *Let X be a complete metric space, and let Y be a normed linear space. Let $F, G : X \rightrightarrows Y$ be set-valued mappings such that both F and $\Phi := F + G$ have closed graphs. Let $(\bar{x}, \bar{y}) \in \text{gph } F$ and $\bar{z} \in G(\bar{x})$. If F is metrically regular at $\bar{x}$ with respect to $\bar{y}$ with modulus $\tau > 0$ and G is locally Lipschitz around $\bar{x}$ with constant $\lambda \in (0, \tau^{-1})$, then Φ is metrically regular at $\bar{x}$ with respect to $\bar{y} + \bar{z}$ with modulus $(\tau^{-1} - \lambda)^{-1}$.*

In [24, Theorem 1], Izmailov and Solodov established an error bound for so-called strongly 2-regular mappings for which the classical regular condition is not satisfied. Then, this result has been extended by He and Sun [19, Theorems 3.1, 4.2] to the case of cone inclusion constraints. Such proofs are based on the set-valued contraction mapping principle. To end this paper, we use Theorem 4.4 to generalize Theorem 1 in [24] and Theorem 4.2 in [19].

Let X be a Banach space, and let Y be a normed linear space. Let $f : X \to Y$ be a given mapping and $K \subseteq Y$ be a closed convex cone. Consider the following inclusion:

$$S := \{x \in X : f(x) \in K\}. \tag{4.12}$$

Let $\bar{x} \in S$ be given. As in [24, 19], we make use of the following hypotheses:

(H1) f is continuously differentiable on a neighborhood V of $\bar{x}$ in X.

(H2) Denote by $Y_1 \subseteq Y$ the closed subspace spanned by $f'(\bar{x})(X) - K$. Then we suppose that Y_1 has a closed complementary subspace Y_2 in Y (which is not the case in general).

Let P and Q denote the projectors from Y onto Y_2 and Y_1, respectively.

(H3) $Pf' : V \to \mathcal{L}(X, Y)$ is Lipschitzian on V with Lipschitz constant L.

(H4) Pf' is B-directionally differentiable at $\bar{x}$ with derivative $(Pf')'(\bar{x}, \cdot)$, i.e.,

$$Pf'(\bar{x} + h) = Pf'(\bar{x}) + (Pf')'(\bar{x}, h) + o(\|h\|), h \in V - \bar{x}.$$

For $h \in X$, set

$$F(h)(u) = f'(\bar{x})(u) + (Pf')'(\bar{x}, h)(u); \quad F(h, K)(u) := (F(h))(u) - K, \quad u \in X,$$
$$\|F(h, K)^{-1}\| := \sup_{y \in Y, \|y\|=1} d(0, F(h, K)^{-1}(y)).$$

Corollary 4.9. *Suppose that assumptions (H1)–(H4) are satisfied. Let K_1 and K_2 be two closed convex cones such that $K_1 + K_2 = K$. If there exists $\nu > 0$ such that*

$$\sup\{\|F(h, K)^{-1}\| : h \in \mathbb{T}_\nu; \|h\| = 1\} := \mu < +\infty, \tag{4.13}$$

where $\mathbb{T}_\nu = \{h \in X : d(f'(\bar{x})(h), K_1) \le \nu,\ d((Pf')'(\bar{x}, h)h, K_2) \le \nu\}$, then there exist $\tau > 0$, $\delta > 0$ such that

$$d(x, S) \le \tau\Big[d(Q(f(x) - f(\bar{x}), K_1) + \|x - \bar{x}\|^{-1} d(P(f(x) - f(\bar{x})), K_2)\Big] \quad \text{for all } x \in B(\bar{x}, \delta) \setminus \{\bar{x}\}. \tag{4.14}$$

In the proof, we make use of the following lemmas.

Lemma 4.10 (Izmailov–Solodov [24, Lemma 1]). *Suppose that the assumptions (H1)–(H3) are fulfilled. Then for any $\varepsilon > 0$ there exists $\delta > 0$ such that*

$$\|Q(f(x) - f(\bar{x})) - f'(\bar{x})(x - \bar{x})\| \le \varepsilon\|x - \bar{x}\| \quad \text{for all } x \in B(\bar{x}, \delta).$$

In addition, assume that Pf' is B-directionally differentiable at $\bar{x}$ with respect to a cone C, i.e., for all $h \in C$, Pf' has a directional derivative at $\bar{x}$ with respect to every $h \in C$ and

$$Pf'(\bar{x} + h) = Pf'(\bar{x}) + (Pf')'(\bar{x}, h) + o(\|h\|), h \in (V - \bar{x}) \cap C,$$

then one has

$$\left\| P(f(x)-f(\bar{x}))-\frac{1}{2}(Pf')'(\bar{x},x-\bar{x})(x-\bar{x})\right\| \\ \leq \varepsilon\|x-\bar{x}\|^2 \quad \textit{for all } x\in B(\bar{x},\delta)\cap(\bar{x}+C).$$

The following lemma is a refinement of Lemma 2 in [24].

LEMMA 4.11. *Suppose that the hypotheses* $(H1)$–$(H3)$ *are verified. Then for any* $\varepsilon,\gamma>0$, *there exists* $\delta>0$ *such that* $B(\bar{x},(1+\gamma)\delta)\subseteq V$ *and for all* $x\in B(\bar{x},\delta)$, *all* $u,v\in B(0,\gamma\|x-\bar{x}\|)$, *one has*

$$\|Q(f(x+u)-f(x+v))-f'(\bar{x})(u-v)\|\leq\varepsilon\|u-v\|.$$

In addition, assume that Pf' *is* B*-directionally differentiable at* $\bar{x}$ *with respect to a cone* C; *then one has*

$$\|P(f(x+u)-f(x+v))-(Pf')'(\bar{x},x-\bar{x})(u-v)\|\leq(\varepsilon+L\gamma)\|x-\bar{x}\|\|u-v\|$$

for all $x\in B(\bar{x},\delta)\cap(\bar{x}+C)$, *and all* $u,v\in B(0,\gamma\|x-\bar{x}\|)$.

Proof. Let $\varepsilon,\gamma>0$ be given. Let $\delta>0$ such that $B(\bar{x},(1+\gamma)\delta)\subseteq V$ and

$$\|Qf'(x)-Qf'(\bar{x})\|+\|Pf'(x)-Pf'(\bar{x})\|<\varepsilon \quad \text{for all } x\in B(\bar{x},(1+\gamma)\delta).$$

Let $x\in B(\bar{x},\delta)$, u, $v\in B(0,\gamma\|x-\bar{x}\|)$. By virtue of the mean value theorem (M.V.T), there exists $\theta\in[0,1]$ such that

$$Q(f(x+u)-f(x+v))=Qf'(z)(u-v),\ \text{with } z:=x+\theta u+(1-\theta)v.$$

Hence,

$$\|Q(f(x+u)-f(x+v))-Qf'(\bar{x})(u-v)\|=\|Qf'(z)-Qf'(\bar{x})\|\|u-v\|\leq\varepsilon\|u-v\|.$$

For the second part, in addition, let $\delta>0$ such that

$$\|Pf'(\bar{x}+th)-Pf'(\bar{x})-t(Pf')'(\bar{x},h)\|<\varepsilon t, h\in(V-\bar{x})\cap C, t\in[0,\delta]. \tag{4.15}$$

Let $x\in B(\bar{x},\delta)$, $u,v\in B(0,\gamma\|x-\bar{x}\|)$ be given. As above, by (M.V.T), noting that Pf' is Lipschitzian on V, one has

$$\|P(f(x+u)-f(x+v))-Pf'(x)(u-v)\|\leq L\gamma\|x-\bar{x}\|\|u-v\|.$$

On the other hand, by (4.15), note that $Pf'(\bar{x})=0$,

$$\|Pf'(x)(u-v)-(Pf')'(\bar{x},x-\bar{x})(u-v)\|\leq\varepsilon\|x-\bar{x}\|\|u-v\|.$$

Combining this inequality and the previous inequality, we derive the conclusion. □

Proof of Corollary 4.9. By Lemmas 4.10 and 4.11, for any ε with $0<\varepsilon<\min\{\mu^{-1}/2,\nu/2\}$, there exist $\delta,\gamma>0$ such that $B(\bar{x},(1+\gamma)\delta)\subseteq V$ and for all $x\in B(\bar{x},\delta)$, all $u,v\in B(0,\gamma\|x-\bar{x}\|)$, one has

$$\|Q(f(x)-f(\bar{x}))-f'(\bar{x})(x-\bar{x})\|\leq\varepsilon\|x-\bar{x}\|,$$

$$\|P(f(x)-f(\bar{x}))-\frac{1}{2}(Pf')'(\bar{x})(x-\bar{x})(x-\bar{x})\|\leq\varepsilon\|x-\bar{x}\|^2 \tag{4.16}$$

$$\|Q(f(x+u)-f(x+v))-f'(\bar{x})(u-v)\|\leq\varepsilon\|u-v\|, \tag{4.17}$$

and

$$\|P(f(x+u)-f(x+v))-(Pf')'(\bar{x},x-\bar{x})(u-v)\|\le\varepsilon\|x-\bar{x}\|\|u-v\|. \tag{4.18}$$

Let $x\in B(\bar{x},\delta)\setminus\{\bar{x}\}$ and set $h=(x-\bar{x})/\|x-\bar{x}\|$. We distinguish the following two cases.

Case 1. $h\notin\mathbb{T}_\nu$. By (4.16), one obviously derives

$$\begin{aligned}d(Q(f(x)-f(\bar{x})),K_1)&\ge\|x-\bar{x}\|\big(d(f'(\bar{x})(h),K_1)-\varepsilon\big),\\ \|x-\bar{x}\|^{-1}d(P(f(x)-f(\bar{x})),K_2)&\ge\|x-\bar{x}\|\left(\frac{1}{2}d((Pf')'(\bar{x},h)h,K_2)-\varepsilon\right).\end{aligned}$$

Hence

$$\begin{aligned}&d(Q(f(x)-f(\bar{x})),K_1)+\|x-\bar{x}\|^{-1}d(P(f(x)-f(\bar{x})),K_2)\\&\quad\ge(\nu/2-\varepsilon)\|x-\bar{x}\|\ge(\nu/2-\varepsilon)d(x,S).\end{aligned}$$

Case 2. $h\in\mathbb{T}_\nu$. Let $g:X\to Y$ and $\Phi:X\rightrightarrows Y$ defined by, $u\in X$,

$$\begin{aligned}g(u)&:=Q(f(x+u)-f(\bar{x}))+\|x-\bar{x}\|^{-1}P(f(x+u)-f(\bar{x}))-F(h)(u),\ \text{and}\\ \Phi(u)&=F(h,K)(u)+g(u).\end{aligned}$$

One has

$$\begin{aligned}\|g(u)-g(v)\|\le{}&\|P(f(x+u)-f(x+v))-(Pf')'(\bar{x},x-\bar{x})(u-v)\|\\&+\|x-\bar{x}\|^{-1}\|P(f(x+u)-f(x+v))-(Pf')'(\bar{x},x-\bar{x})(u-v)\|.\end{aligned}$$

Then by relations (4.17) and (4.18), g is Lipschitz on $B(0,\gamma\|x-\bar{x}\|)$ with constant 2ε. Note that $F(h,K)(\cdot)$ is a closed convex process, since (4.13) and by virtue of Theorem 1 in [39] due to Robinson, $F(h,K)$ is (globally) metrically regular with modulus μ. Therefore, by Corollary 4.5, Φ is metrically regular around 0 with respect to $g(0)=Q(f(x)-f(\bar{x}))+\|x-\bar{x}\|^{-1}P(f(x)-f(\bar{x}))$ with modulus $(\mu^{-1}-2\varepsilon)^{-1}$. Moreover, according to Remark 4.6, there exists a constant $c>0$ depending only on μ such that

$$\begin{aligned}d(u,\Phi^{-1}(y))\le(\mu^{-1}-2\varepsilon)^{-1}d(y,\Phi(u))\quad&\text{for all }(u,y)\in B(0,2c\tau\gamma\|x-\bar{x}\|)\\&\times B(g(0),c\gamma\|x-\bar{x}\|).\end{aligned} \tag{4.19}$$

Hence, if $d(0,\Phi(0))=d(g(0),K)\ge c\gamma\|x-\bar{x}\|$, then

$$d(Q(f(x)-f(\bar{x})),K_1)+\|x-\bar{x}\|^{-1}d(P(f(x)-f(\bar{x})),K_2)\ge d(g(0),K)\ge c\gamma d(x,S);$$

otherwise, taking a sequence $\{z_n\}_{n\in\mathbb{N}}\subseteq K$ with $\lim_{n\to\infty}\|g(0)-z_n\|=d(g(0),K)$, one has (note that $x+\Phi^{-1}(z_n)\subseteq S$)

$$\begin{aligned}d(x,S)&\le\liminf d(0,\Phi^{-1}(z_n))\le(\mu^{-1}-2\varepsilon)^{-1}\liminf d(z_n,\Phi(0))\\&=(\mu^{-1}-2\varepsilon)^{-1}d(g(0),K).\end{aligned}$$

Consequently, one obtains

$$d(x,S)\le(\mu^{-1}-2\varepsilon)^{-1}\big(d(Q(f(x)-f(\bar{x})),K_1)+\|x-\bar{x}\|^{-1}d(P(f(x)-f(\bar{x})),K_2)\big).$$

Combining the two cases, we establish the proof. □

In [25], Izmailov and Solodov have established a locally covering property at a point of so-called 2-regular mappings. We next give a new simple proof of that result, which is based on Corollary 4.5.

COROLLARY 4.12 (see [25, Theorem 4.1]). *Let X, Y be Banach spaces. Suppose that the assumptions $(H1)$–$(H3)$ hold with $K = \{0\}$ and that the mapping $Pf'(\cdot)$ has a directional derivative at the point $\bar{x}$ in a direction $h \in X$ with $\|h\| = 1$, satisfying $f'(\bar{x})(h) = 0$ and $(Pf')'(\bar{x}, h)(h) = 0$. If the mapping f is 2-regular at $\bar{x}$ with respect to h, i.e., the continuous linear mapping*

$$u \mapsto F(h)(u) := f'(\bar{x})(u) + (Pf')'(\bar{x}, h)(u)$$

from X to Y is surjective, then there exist a neighborhood W of $f(\bar{x})$ in Y and a constant $C > 0$ such that

$$d(\bar{x}, f^{-1}(y)) \le C\|f(\bar{x}) - y)\|^{1/2} \quad \text{for all } y \in W.$$

Proof. We can assume $f(\bar{x}) = 0$ without loss of generality. Since $F(h)(\cdot)$ is surjective, then by the open mapping principle,

$$M := \sup\{d(0, F(h)^{-1}(y)) : y \in Y, \|y\| = 1\} < +\infty.$$

For each $t > 0$, set $x_t := x + th$. Using Lemmas 4.10 and 4.11, for any ε, $\gamma > 0$ with $2\varepsilon + L\gamma \in (0, M^{-1})$, we can find $\delta \in (0, 1)$ such that $B(\bar{x}, (1+\gamma)\delta) \subseteq V$ and that for all $t \in (0, \delta)$, all $u, v \in B(0, \gamma t)$, one has

$$\|Q(f(x_t))\| \le \varepsilon t, \quad \|P(f(x_t))\| \le \varepsilon t^2 \tag{4.20}$$

$$\|Q(f(x_t + u) - f(x_t + v)) - f'(\bar{x})(u - v)\| \le \varepsilon\|u - v\|, \tag{4.21}$$

and

$$\|P(f(x_t + u) - f(x_t + v)) - t(Pf')'(\bar{x}, h)(u - v)\| \le (\varepsilon + L\gamma)t\|u - v\|. \tag{4.22}$$

Similar to the proof of Case 2 in Corollary 4.9, defining $g_t : X \to Y$ and $\Phi_t : X \to Y$ by

$$g_t(u) := Q(f(x_t + u)) + t^{-1}P(f(x_t + u)) - F(h)(u)$$

and

$$\Phi_t(u) = F(h)(u) + g_t(u) = Q(f(x_t + u)) + t^{-1}P(f(x_t + u)),$$

there exists a constant $c > 0$ depending only on M such that

$$d(u, \Phi_t^{-1}(y)) \le (M^{-1} - 2\varepsilon - L\gamma)^{-1}\|y - \Phi_t(u)\|$$
$$\text{for all } (u, y) \in B(0, 2c\tau\gamma t) \times B(g_t(0), c\gamma t). \tag{4.23}$$

Therefore, we obtain (for $\tau = (M^{-1} - 2\varepsilon - L\gamma)^{-1}$)

$$\begin{aligned} d(0, \Phi_t^{-1}(y) &= d(x_t, f^{-1}(y)) \\ &\le \tau\Big[\|Q(f(x_t) - y)\| + t^{-1}\|P(f(x_t) - y)\|\Big] \quad \text{for all } y \in B(g_t(0), c\gamma t). \end{aligned}$$

Take $\varepsilon \in (0, M^{-1}/2)$ such that $\varepsilon < c\gamma/2$. From (4.20) it yields $\|g_t(0)\| \le 2\varepsilon t$. Set $\eta := \min\{\delta, c\gamma - 2\varepsilon\}^2$. Now let $y \in B(0, \eta)$ and take $t = \|y\|^{1/2}$. Then, obviously, $y \in B(g_t(0), c\gamma t)$. Hence, from (4.23), we obtain

$$\begin{aligned} d(\bar{x}, f^{-1}(y)) &\le d(x_t, f^{-1}(y)) + t \le \tau(\|Q(f(x_t) - y)\| + t^{-1}\|P(f(x_t) - y)\|) + t \\ &\le \tau(\|Qy\| + t^{-1}\|Py\|) + 2\varepsilon t + t \le \tau(\|y\| + 2(\varepsilon + 1)\|y\|^{1/2}), \end{aligned}$$

establishing the proof. □

Acknowledgments. The authors would like to warmly thank the two anonymous referees for their careful reading of this paper.

REFERENCES

[1] D. Azé, *A survey on error bounds for lower semicontinuous functions*, in Proceedings of the 2003 MODE-SMAI Conference, ESAIM Proc., 1, EDP Sci., Les Ulis, 2003, pp. 1–17.

[2] D. Azé, *A unified theory for metric regularity of multifunctions*, J. Convex Anal., 13 (2006), pp. 225–252.

[3] D. Azé, J.-N. Corvellec, and R. E. Lucchetti, *Variational pairs and applications to stability in nonsmooth analysis*, Nonlinear Anal., 49 (2002), pp. 643–670.

[4] D. Azé and J.-N. Corvellec, *On the sensitivity analysis of Hoffman constants for systems of linear inequalities*, SIAM J. Optim., 12 (2002), pp. 913–927.

[5] D. Azé and J.-N. Corvellec, *Characterizations of error bounds for lower semicontinuous functions on metric spaces*, ESAIM Control Optim. Calc. Var., 10 (2004), pp. 409–425.

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

[7] J. M. Borwein and A. L. Dontchev, *On the Bartle-Graves theorem*, Proc. Amer. Math. Soc., 131 (2003), pp. 2553–2560.

[8] J. M. Borwein and Q. J. Zhu, *Viscosity solutions and viscosity subderivatives in smooth Banach spaces with applications to metric regularity*, SIAM J. Control Optim., 34 (1996), pp. 1568–1591.

[9] J. M. Borwein and D. M. Zhuang, *Verifiable necessary and sufficient conditions for openness and regularity of set-valued maps*, J. Math. Anal. Appl., 134 (1988), pp. 441–459.

[10] P. Bosch, A. Jourani, and R. Henrion, *Sufficient conditions for error bounds and applications*, Appl. Math. Optim., 50 (2004), pp. 161–181.

[11] R. Cominetti, *Metric regularity, tangent cones, and second-order optimality conditions*, Appl. Math. Optim., 21 (1990), pp. 265–287.

[12] E. De Giorgi, A. Marino, and M. Tosques, *Problemi di evoluzione in spazi metrici e curve di massima pendenza (Evolution problems in metric spaces and curves of maximal slope)*, Atti. Accad. Naz. Lincei rend. Cl. Sci. fis. Mat. Natur., 68 (1980), pp. 180–187.

[13] A. V. Dmitruk, A. A. Milyutin, and N. P. Osmolovskii, *Lusternik's theorem and the theory of extremum*, Uspekhi Mat. Nauk, 35 (1980), pp. 11–46.

[14] A. L. Dontchev, *The Graves theorem revisited*, J. Convex Anal., 3 (1996), pp. 45–53.

[15] A. L. Dontchev, A. S. Lewis, and R. T. Rockafellar, *The radius of metric regularity*, Trans. Amer. Math. Soc., 355 (2003), pp. 493–517.

[16] A. L. Dontchev and A. S. Lewis, *Perturbation and metric regularity*, Set-Valued Anal., 13 (2005), pp. 417–438.

[17] I. Ekeland, *On the variational principle*, J. Math. Anal. Appl., 47 (1974), pp. 324–353.

[18] L. M. Graves, *Some mapping theorems*, Duke Math. J., 17 (1950), pp. 111–114.

[19] Y. He and J. Sun, *Error bounds for degenerate cone inclusion problems*, Math. Oper. Res., 30 (2005), pp. 701–717.

[20] A. Ioffe, *Regular points of Lipschitz functions*, Trans. Amer. Math. Soc., 251 (1979), pp. 61–69.

[21] A. D. Ioffe, *Metric regularity and subdifferential calculus (in Russian)*, Uspeckhi Math. Nauk, 55 (2000), pp. 103–162; English translation in Russian Math. Survey, 55 (2000), pp. 501–558.

[22] A. D. Ioffe, *On perturbation stability of metric regularity*, Set-Valued Anal., 9 (2001), pp. 101–109.

[23] A. D. Ioffe, *Towards metric theory of metric regularity*, in Approximation, Optimization and Mathematical Economics (Pointe à Pitre, 1999), Physica, Heidelberg, 2001, pp. 165–176.

[24] A. F. Izmailov and M. V. Solodov, *Error bounds for* 2*-regular mappings with Lipschitzian derivatives and their applications*, Math. Program., 89 (2001), pp. 413–435.

[25] A. F. Izmailov and M. V. Solodov, *The theory of* 2*-regularity for mappings with Lipschitzian derivatives and their applications to optimality conditions*, Math. Oper. Res., 27 (2001), pp. 614–635.

[26] A. Jourani, *Hoffman's error bound, local controllability, and sensivity analysis*, SIAM J. Control Optim., 38 (2000), pp. 947–970.

[27] A. Jourani and L. Thibault, *Metric regularity and subdifferential calculus in Banach spaces*, Set-Valued Anal., 3 (1995), pp. 87–100.

[28] A. Jourani and L. Thibault, *Coderivatives of multivalued mappings, locally compact cones and metric regularity*, Nonlinear Anal., 35 (1994), pp. 925–945.
[29] L. A. Lyusternik, *Sur les extrêmes relatives de fonctionnelles*, Math. Sbornik, 41 (1934), pp. 390–401.
[30] B. S. Mordukhovich and Y. Shao, *Stability of set-valued mappings in infinite dimensions: Point criteria and applications*, SIAM J. Control Optim., 35 (1997), pp. 285–314.
[31] B. S. Mordukhovich, *Variational analysis and generalized differentiation.* I. *Basic theory*, in Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences] 330, Springer-Verlag, Berlin, 2006.
[32] B. S. Mordukhovich, *Variational analysis and generalized differentiation.* II. *Applications*, in Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences] 331, Springer-Verlag, Berlin, 2006.
[33] B. S. Mordukhovich and B. Wang, *Restrictive metric regularity and generalized differential calculus in Banach spaces*, Int. J. Math. Math. Sci., 50 (2004), pp. 2650–2683.
[34] H. Ngai and M. Théra, *Metric inequality, subdifferential calculus and applications*, Set-Valued Anal., 9 (2001), pp. 187–216.
[35] H. Ngai and M. Théra, *Error bounds and implicit multifunctions in smooth Banach spaces and applications to optimization*, Set-Valued Anal., 12 (2004), pp. 195–223.
[36] H. Ngai and M. Théra, *Error bounds for systems of lower semicontinuous functions in Asplund spaces*, accepted to Math. Program.
[37] K. F. Ng and X. Y. Zheng, *Error bounds for lower semicontinuous functions in normed spaces*, SIAM J. Optim., 12 (2001), pp. 1–17.
[38] J.-P. Penot, *Regularity, openness and Lipschitzian behavior of multifunctions*, Nonlinear Anal., 13 (1989), pp. 629–643.
[39] S. M. Robinson, *Normed convex processes*, Trans. Amer. Math. Soc., 174 (1972), pp. 127–140.
[40] Z. Wu and J. Ye, *On error bounds for lower semicontinuous functions*, Math. Program., 92 (2002), pp. 301–314.
[41] X. Y. Zheng and K. F. Ng, *Metric regularity and constraint qualifications for convex inequalities on Banach spaces*, SIAM J. Optim., 14 (2003), pp. 757–772.

SIAM J. OPTIM.
Vol. 19, No. 1, pp. 21–38

BRANCH-AND-CUT FOR THE MAXIMUM FEASIBLE SUBSYSTEM PROBLEM*

MARC E. PFETSCH†

Abstract. This paper presents a branch-and-cut algorithm for the NP-hard maximum feasible subsystem problem: For a given infeasible linear inequality system, determine a feasible subsystem containing as many inequalities as possible. The complementary problem, where one has to remove as few inequalities as possible in order to make the system feasible, can be formulated as a set covering problem. The rows of this formulation correspond to irreducible infeasible subsystems, which can be exponentially many. It turns out that the main issue of a branch-and-cut algorithm for the maximum feasible subsystem problem (MAX FS) is to efficiently find such infeasible subsystems. We present three heuristics for the corresponding NP-hard separation problem and discuss cutting planes from the literature, such as set covering cuts of Balas and Ng, Gomory cuts, and $\{0, \frac{1}{2}\}$-cuts. Furthermore, we compare a heuristic of Chinneck and a simple greedy algorithm. The main contribution of this paper is an extensive computational study on a variety of instances arising in a number of applications.

Key words. infeasible linear inequality system, irreducible infeasible subsystem (IIS), maximum feasible subsystem problem, minimum IIS-cover, branch-and-cut

AMS subject classification. 90C27

DOI. 10.1137/050645828

1. Introduction.

In the *maximum feasible subsystem problem* (MAX FS), we are given an infeasible linear inequality system $\Sigma : \{A\boldsymbol{x} \leq \boldsymbol{b}\}$, with $A \in \mathbb{R}^{m\times n}$, $\boldsymbol{b} \in \mathbb{R}^m$, and have to find a feasible subsystem containing as many inequalities as possible. This NP-hard combinatorial optimization problem has a number of interesting applications in a wide range of fields, for instance, in linear programming [29, 31, 36], statistical discriminant analysis and machine learning [4, 19, 43], telecommunications [54], and computational biology [61]. Additional applications and a survey can be found in [4] and [5], respectively.

The complementary problem of MAX FS amounts to removing as few inequalities of Σ as possible so that the resulting system is feasible. To achieve feasibility, one has to remove at least one inequality from each *irreducible infeasible subsystem* (IIS), i.e., an infeasible subsystem of Σ for which every proper subsystem is feasible. Introducing a binary variable y_i for each inequality of Σ, the complementary problem can be formulated as a set covering problem and is therefore called MIN IIS COVER:

$$\begin{array}{lll} \min & \sum_{i=1}^m y_i & \\ \text{such that} & \sum_{i\in I} y_i \geq 1 & \text{for all IISs } I, \\ & \boldsymbol{y} \in \{0,1\}^m. & \end{array} \tag{1}$$

Since the number of IISs can be exponential in the size of the system Σ (see Chakravarti [28] and Pfetsch [53]), IISs have to be generated dynamically in order to solve this formulation of MIN IIS COVER efficiently.

*Received by the editors November 22, 2005; accepted for publication (in revised form) July 12, 2007; published electronically February 6, 2008.
http://www.siam.org/journals/siopt/19-1/64582.html
†Zuse Institute Berlin, Takustr. 7, 14195 Berlin, Germany (pfetsch@zib.de). This author's research was supported by the DFG Research Center MATHEON "Mathematics for key technologies" in Berlin.

Clearly, the set of all inequalities not contained in a solution of MAX FS form a solution of MIN IIS COVER and vice versa. Hence, these two problems are equivalent when solving to optimality and are both strongly NP-hard; see Johnson and Preparata [39], Sankaran [58], and Chakravarti [28]. In terms of approximability, however, they differ: MAX FS does not admit a polynomial-time approximation scheme, unless P = NP, but there exists a 2-approximation; see Amaldi and Kann [9]. MIN IIS COVER is harder to approximate: Unless P = NP, it cannot be approximated in polynomial time within any constant factor; see Amaldi and Kann [10].

In this paper, we present a branch-and-cut approach for MAX FS via formulation (1) for MIN IIS COVER. A key issue of this approach is to find violated *IIS-inequalities*, i.e., the inequalities arising from IISs in (1). The corresponding separation problem is NP-hard, and we present three heuristics for it (see section 3.2). Two of these methods generate either a feasible solution for MIN IIS COVER or a (hopefully violated) IIS-inequality. As long as no feasible solution has been generated, the process is iterated, which often produces many useful IIS-inequalities. The additional benefit is reasonably good primal solutions, which can be improved by a simple greedy algorithm. This combination leads to an effective primal heuristic. Additionally, we examine the application of inequalities of Balas and Ng [18] for set covering problems, $\{0, \frac{1}{2}\}$-cuts, and Gomory cuts.

The emphasis of this paper is on an extensive computational study of the branch-and-cut implementation. Our aim is to show the potential and the limits of such an approach by performing tests on three problem sets: random infeasible inequality systems (section 4.2), problems arising in digital video broadcasting (section 4.3), and classification problems (section 4.4).

The theoretical foundation for our approach appears in Amaldi, Pfetsch, and Trotter [12], where algorithmic and geometric questions concerning IISs are studied and the feasible subsystem polytope is investigated. (The polyhedral results carry over to the polytope for MIN IIS COVER by a simple affine transformation.) The work presented here is an improved version of part of the author's Ph.D. thesis [53].

In the literature to date, only two exact approaches towards MIN IIS COVER have appeared. Parker and Ryan [52] discuss an iterative approach that generates IISs in each step and then solves an integer program. This approach turns out to be impractical for harder instances. Codato and Fischetti [33] present a branch-and-cut algorithm for MIN IIS COVER in a more general context. We discuss these approaches in more detail in the next section. Our algorithm improves upon both methods and is currently the best available exact approach (see section 4).

The outline of this paper is as follows. In section 2 we review solution approaches for MAX FS. In section 3 we describe the main ingredients of our branch-and-cut implementation. We discuss a way to check the feasibility of solutions for MIN IIS COVER, three methods to separate IIS-inequalities, primal heuristics, preprocessing, branching, inequalities by Balas and Ng, and other used cutting planes. In section 4 we extensively test the implementation on the abovementioned problem sets. We close with some conclusions in section 5.

We use the following notation. We define $[n] := \{1, \dots, n\}$ for $n \in \mathbb{N}$ and typeset vectors in bold font. For a set $S \subseteq [n]$ and a vector $\boldsymbol{x} \in \mathbb{R}^n$, define

$$\boldsymbol{x}(S) = \sum_{i \in S} x_i \,.$$

The *support* of a vector $\boldsymbol{x} \in \mathbb{R}^n$ is $\operatorname{supp}(\boldsymbol{x}) := \{i \in [n] \,:\, x_i \neq 0\}$. By $\mathbb{1}$ we denote a vector of all 1's of appropriate dimension.

2. Alternative solution approaches. In this section we give a short overview of solution approaches for MAX FS and MIN IIS COVER.

In the context of linear programming, attention was first devoted to the problem of identifying IISs with a small and possibly minimum number of inequalities (see Greenberg and Murphy [36], Chinneck [30], and Chinneck and Dravnieks [32]). The goal is to help the modeler resolve infeasibility of large linear programs (LPs). Since minimum cardinality covers of IISs reveal essential information about infeasibility of the model and are often smaller than IISs, emphasis has shifted towards their identification. Chinneck [29, 31] developed extended greedy heuristics for MAX FS/MIN IIS COVER and provided computational results; see section 4.4.

For the application of MIN IIS COVER to classification problems (see section 4.4), several heuristics were proposed, based on nonlinear programming formulations of MAX FS (Bennett and Bredensteiner [19], Bennett and Mangasarian [20], and Mangasarian [43]).

An exact integer programming approach for MIN IIS COVER appeared in Parker and Ryan [52] and Parker [51]. Their idea is to consider the formulation in (1) with a partial list of IISs. If there exist IISs that are not covered by a solution to this formulation, they are added and the process is iterated. Otherwise, an optimal solution to MIN IIS COVER is found. Parker and Ryan discuss several methods of generating IISs at each step and consider heuristics for solving the set covering problem (only the last instance has to be solved exactly).

We reimplemented a basic version of their algorithm, where the set covering problems are solved to optimality. This implementation turned out to be inferior to our branch-and-cut implementation: It could not solve within one hour instances solved by our branch-and-cut approach within a few minutes. We therefore refrained from performing further experiments.

There is a straightforward mixed integer programming formulation for MIN IIS COVER containing a binary variable with a "big-M" for each of the inequalities of Σ, so that an inequality is relaxed when the corresponding binary variable is 1. This formulation has the typical numerical problems of big-M formulations and is in general inefficient for MAX FS; see Parker [51]. If there are fixed bounds on the variables, however, one can obtain a tight formulation. This leads to a quite efficient approach; see Rossi, Sassano, and Smriglio [54] and Codato and Fischetti [33]. In fact, Codato and Fischetti propose a general way of removing the "big-M" from this type of formulation and apply it to classification instances. In this context, it leads to the formulation (1), and their solution method is, in fact, a branch-and-cut method for MIN IIS COVER, independent from our approach. Computational results show that their approach is faster compared to the big-M formulation. In section 4.4 we compare our implementation with their approach.

Versions of the classical *relaxation method* of Agmon [3] and Motzkin and Schoenberg [47] for solving linear inequality systems can be applied to minimize the sum of violations in infeasible linear inequality systems. Randomized variants of this method were proposed by Amaldi [4] to solve MAX FS. Amaldi and Hauser [8] and Amaldi, Belotti, and Hauser [6] establish probabilistic convergence guarantees to an optimal solution of MAX FS under appropriate conditions. Computational results for digital video broadcasting data, classification instances, and huge systems arising in computational biology are given in [6].

Amaldi, Bruglieri, and Casale [7] propose a two-step heuristic in which first a linearization of an exact bilinear formulation of MAX FS is used to derive a feasible

subsystem. In the second step, a reduced problem is solved to optimality in order to identify inequalities that can be added to the first system while preserving feasibility. This turns out to be competitive with respect to the method of Codato and Fischetti and an integer programming solver applied to the "big-M" formulation for the whole system.

3. Ingredients for branch-and-cut. In the following we assume that the reader is familiar with the branch-and-cut approach. More information can be found in Nemhauser and Wolsey [48], Padberg and Rinaldi [50], Thienel [60], and Caprara and Fischetti [26]. A description and computational study of Gomory cuts is given in Balas et al. [17].

Recall that we are given the infeasible system $\Sigma : \{A\boldsymbol{x} \leq \boldsymbol{b}\}$, where $A \in \mathbb{R}^{m\times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$. Depending on the application, *mandatory* variable bounds can be present; i.e., these bounds may not be removed for obtaining a feasible system (see sections 4.3 and 4.4). This can easily be dealt with in the branch-and-cut approach. Furthermore, weighted versions of Min IIS Cover are easy to handle, too.

Without loss of generality we can restrict attention to inequality systems in the form of Σ: Clearly, bounds on variables and "greater or equal" inequalities can be transformed to this format. Equations can be replaced by a pair of opposing inequalities. Since any point satisfies at least one inequality out of each pair, an optimal solution to the new instance contains $m^* + m_E$ inequalities if and only if an optimal solution to the original instance with m^* linear relations exists; here m_E is the number of equations. Thus, from a computational point of view, it suffices to handle systems in the form of Σ. Polyhedral results for the two cases, however, may differ; see [12, 53] for more information.

To simplify notation, we identify an inequality of Σ with its index. Then $S(\Sigma) := [m]$ is the set of constraints of Σ. With this notation, $I \subseteq S(\Sigma)$ is an IIS of Σ if and only if all proper subsets of I are feasible. We call a set $C \subseteq S(\Sigma)$ an *IIS-cover* if it intersects every IIS of Σ.

In the rest of this section we give a more detailed account of the main aspects of our implementation: the recognition problem for IIS-covers, the separation problem of IIS-inequalities, pool handling, primal heuristics, preprocessing, branching, and other cutting planes.

3.1. Recognition problem for IIS-covers. We consider the following fundamental problem: Given a subset $C \subseteq S(\Sigma)$, check whether it is an IIS-cover and if this is not the case, generate a witness, i.e., an IIS which is not covered. Our approach is based on the following theorem.

Theorem 1 (Gleeson and Ryan [35]). *Let $\Sigma : \{A\boldsymbol{x} \leq \boldsymbol{b}\}$ be an infeasible system. Then the IISs of Σ are in one-to-one correspondence with the supports of the vertices of the polyhedron*

$$P(\Sigma) := \{\, \boldsymbol{y} \in \mathbb{R}^m \,:\, \boldsymbol{y}^{\mathrm{T}}A = \boldsymbol{0},\ \boldsymbol{y}^{\mathrm{T}}\boldsymbol{b} = -1,\ \boldsymbol{y} \geq \boldsymbol{0} \,\}.$$

Note that the vertices of $P(\Sigma)$ are uniquely defined by their supports. This theorem is strongly related to the Farkas lemma, which states that $P(\Sigma) \neq \varnothing$ if and only if Σ is infeasible; see, e.g., Schrijver [59]. The polyhedron $P(\Sigma)$ is called the *alternative polyhedron* of Σ.

To apply Theorem 1, we define for $S \subseteq S(\Sigma)$ the polyhedron

$$P_S(\Sigma) := \{\, \boldsymbol{y} \in P(\Sigma) \,:\, y_i = 0,\ i \in S \,\},$$

which might be empty. We need the following fact.

LEMMA 2 (Parker and Ryan [52]). *The set $C \subseteq S(\Sigma)$ is an IIS-cover if and only if $P_C(\Sigma) = \varnothing$.*

Proof. The system defining $P(\Sigma)$, in which all variables indexed by C are removed, has no solution if and only if $P_C(\Sigma) = \varnothing$. By the Farkas lemma, the former is the case if and only if Σ with inequalities indexed by C removed is feasible, i.e., C is an IIS-cover. □

Recognizing whether $C \subseteq S(\Sigma)$ is an IIS-cover is now easy: If $P_C(\Sigma) = \varnothing$, by Lemma 2, C is an IIS-cover. Otherwise, let $\boldsymbol{v}$ be a vertex of $P_C(\Sigma)$. Then we have $\operatorname{supp}(\boldsymbol{v}) \cap C = \varnothing$, showing that $\operatorname{supp}(\boldsymbol{v})$ is an IIS that is uncovered (by Theorem 1). This provides a polynomial-time algorithm for the problem, since finding a vertex of a polyhedron can be done in polynomial time; see Grötschel, Lovász, and Schrijver [37]. Note that by Theorem 1 and Lemma 2, $P_C(\Sigma)$ always has a vertex if it is nonempty.

This recognition test in fact suffices for a rudimentary branch-and-cut algorithm, since we can now test feasibility of a vector $\boldsymbol{y} \in \{0,1\}^m$ for (1) by testing whether $\operatorname{supp}(\boldsymbol{y})$ is an IIS-cover.

3.2. Separation of IIS-inequalities. IIS-inequalities play a prominent role in the formulation (1) for MIN IIS COVER. In fact, it can be shown that the inequality arising from the IIS I defines a facet of the polytope

$$P_{IISC} = \operatorname{conv}\{\, \boldsymbol{y} \in \{0,1\}^m \,:\, y(S) \geq 1 \text{ for all IISs } S \,\},$$

as long as $|I| > 1$; see Amaldi, Pfetsch, and Trotter [12]. Therefore, the following *separation problem for IIS-inequalities* is crucial: Given a vector $\boldsymbol{y}^* \in [0,1]^m$, check whether there exists an IIS I so that its corresponding inequality is violated by $\boldsymbol{y}^*$, i.e., $\boldsymbol{y}^*(I) < 1$. The recognition problem for IIS-covers is a special case, where $\boldsymbol{y}^*$ is the incidence vector of the set to be tested. In the general case, however, we have the following.

PROPOSITION 3 (Amaldi, Pfetsch, and Trotter [12]). *The separation problem for IIS-inequalities is* NP*-hard.*

In this section, we therefore present three heuristics for the separation problem. All of these heuristics may fail to produce a violated IIS-inequality.

The heuristics build on the following reformulation of the separation problem: Compute

$$\lambda := \min\{\, \boldsymbol{y}^*(S) \,:\, S = \operatorname{supp}(\boldsymbol{v}),\ \boldsymbol{v} \text{ vertex of } P(\Sigma) \,\}. \tag{2}$$

If $\lambda < 1$, by Theorem 1, $\operatorname{supp}(\boldsymbol{v})$ provides an IIS whose IIS-inequality is violated; otherwise no such IIS exists (we define $\lambda = \infty$ if $P(\Sigma) = \varnothing$).

3.2.1. Method 1: "Single." The first quite intuitive idea for separating an IIS-inequality, already used by Parker and Ryan [52], is to approximate (2) by the following LP:

$$\min\{\, (\boldsymbol{y}^*)^{\mathrm{T}} \boldsymbol{p} \,:\, \boldsymbol{p} \in P(\Sigma) \,\}.$$

A vertex solution provides an IIS, whose corresponding inequality is not necessarily violated, but in practice often is.

This method generates only one IIS at a time. We also experimented with solving the above LP by the simplex algorithm and then testing whether the support of each vertex on the path to the optimum is an IIS whose inequality is violated. In our experiments this variant was inefficient and will not be considered further.

3.2.2. Method 2: "Extend." We extend Method 1 as follows. Let S be the support of $\boldsymbol{y}^*$. Applying Lemma 2, we can check whether S is an IIS-cover by finding a vertex solution of

$$\min\{\, (\boldsymbol{y}^*)^{\mathrm{T}}\boldsymbol{p} \,:\, \boldsymbol{p} \in P_S(\Sigma)\,\}$$

if one exists. If the LP is feasible, the result gives us a vertex which corresponds to an IIS; otherwise we found an IIS-cover, i.e., a primal solution for Min IIS Cover.

This approach can be iterated when S is not an IIS-cover. Let I be the IIS obtained in this case. We enlarge S greedily by an element of I and iterate. In our implementation, we choose an element of I that is contained in the maximal number of IISs we have found so far. At termination this yields an IIS-cover. This procedure is related to a primal heuristic proposed by Ryan [57].

The IISs found by this approach have several nice properties. First, the new IISs are different from all IISs that were known before the run if the current solution $\boldsymbol{y}^*$ of the LP-relaxation satisfies $y^*(I) \geq 1$ for each previously found IIS I. This follows since at least one element of each I is contained in S, and hence I cannot be generated again. Second, the corresponding inequalities are always violated, since they have empty intersection with $S \supseteq \operatorname{supp}(\boldsymbol{y}^*)$; i.e., $\boldsymbol{y}^*(I) = 0 < 1$ for each produced IIS I. Third, by construction of the set S, the generated IISs are pairwise different.

This method turns out to be quite effective for generating many violated IIS-inequalities. Furthermore, we obtain a primal solution in each run, which can be improved to very good solutions; see section 3.4. When the current LP-relaxation contains many cuts, however, the support of $\boldsymbol{y}^*$ tends to be large and often is already an IIS-cover or close to one, and the method cannot produce new IISs; this often happens in the deeper regions of the branch-and-bound tree. This might even be desirable, since this saves time for high depths. Nevertheless, this situation can be changed, as indicated by the next method.

3.2.3. Method 3: "Round." The idea of Method 2 can be further extended by using the fact that an arbitrary set S can be used at the start. In the extension, we choose $\alpha \in [0,1]$ and initially let $S := \{i : y^*_i \geq \alpha\}$. In the implementation we start with $\alpha = 0.1$ and then increase α by 0.1 until S is not an IIS-cover (in this case the above procedure is started). We terminate with a failure if α exceeds 0.6.

The fact that S is smaller for larger α has two effects: First, the number of steps needed to greedily obtain an IIS-cover is larger, and hence the number of generated IISs is increased. Second, the method also computes IISs in the deeper regions of the tree.

Again, in each step an IIS is generated, which is not covered by S, except in the last step where we obtain an IIS-cover. In contrast to the method "extend," the generated IISs are not necessarily new, and their corresponding inequalities may not be violated by $\boldsymbol{y}^*$.

3.3. Pool for IIS-inequalities. The above three methods tend to produce many IISs, which we store in a pool. It turns out that the best performance of the algorithm is achieved by checking the pool for violated inequalities in *every* node of the tree. Of course, the pool should be as small as possible without losing important inequalities. Therefore, the pool is equipped with an aging mechanism which removes IISs whose inequality has not been active for some time.

The computational results presented in section 4 indicate that only a small fraction of the total number of IISs needs to be generated by our branch-and-cut implementation; indeed, for larger problems there are far too many IISs to be enumerated

completely (cf. Table 2 in section 4.2). Hence, the size of the pool can be relatively small.

3.4. Primal heuristics. Chinneck [31] proposed a greedy heuristic for MIN IIS COVER, which we use as an initial primal heuristic. The basic tool is a so-called *elastic LP* in which the inequalities $\Sigma : \{A\boldsymbol{x} \leq \boldsymbol{b}\}$ are relaxed by adding slack variables and the sum of violations is minimized:

$$\begin{aligned} &\min \mathbb{1}^{\mathrm{T}}\boldsymbol{s} \\ &\text{s.t. } A\boldsymbol{x} - \boldsymbol{s} \leq \boldsymbol{b}, \\ &\boldsymbol{s} \geq 0. \end{aligned}$$

Starting with $S = \varnothing$, in each iteration $S \subseteq S(\Sigma)$ is enlarged by an inequality that yields the largest drop in the elastic LP objective if its objective coefficient is set to 0. The method stops once the objective is 0, i.e., S is a MIN IIS COVER. To speed up the solution, in each iteration only inequalities from a candidate set are checked. Chinneck proposes a measure based on the violation and dual variables to generate the candidate set. We refer to [31] for details.

For a heuristic running in the tree, we use a primal heuristic that greedily decreases the size of a given IIS-cover until a minimal one is obtained. We start this heuristic from IIS-covers produced by the separation methods in section 3.2, if available (otherwise we use a simple rounding heuristic). We start with C being an IIS-cover to be improved. We consider each element from C in the order of increasing fractional value of the current LP-solution $\boldsymbol{y}^*$. We remove an element if the remaining set is an IIS-cover (which is checked by the method in section 3.1).

3.5. Preprocessing. In a preprocessing step we search for small IISs. Such small IISs are of interest since their corresponding IIS-inequalities provide "strong" cuts and are hard to find by other methods.

We first check for IISs of cardinality one, e.g., $\boldsymbol{0x} \leq -1$. Then we check for IISs that involve one inequality and bounds on the variables (if present). Such IISs often occur when variable bounds are mandatory; see, e.g., section 4.4. In this case, a single inequality might be infeasible with the bounds and counts as an IIS. Furthermore, we look for IISs of cardinality two, which are easy to find by comparing their normal vectors and right-hand sides. Identifying other types of IISs would require higher computational effort.

3.6. Branching. As a branching rule, we apply *reliability branching*, introduced by Achterberg, Koch, and Martin [2]. It performs strong branching on a subset of the variables, which are chosen based on their so-called pseudocosts during branching. If in strong branching one of the child nodes turns out to be infeasible, the corresponding variable is fixed to the complementary value; if both children are infeasible, the current node can be pruned.

We also experimented with constraint branching rules. For instance, we used the well-known rule of Ryan and Foster [56]. This rule was superior to a simple variable branching, but inferior to reliability branching both in terms of computation time and the number of branch-and-bound nodes. We therefore selected reliability branching for all tests.

3.7. Inequalities for set covering. Many facet-defining inequalities for the set covering polytope have been investigated; see Ceria, Nobili, and Sassano [27] and

Borndörfer [22]. However, few (problem-specific) polynomial-time separable inequalities for set covering are known. For many classes of inequalities the complexity status is unknown, but is likely to be NP-hard.

We experimented with the aggregated cycle cuts of Borndörfer and Weismantel [23, 24]. Unfortunately, on our test problems their separation heuristic almost never found a violated inequality. Furthermore, it remains an interesting open problem to identify problem-specific inequalities for MIN IIS COVER.

A class of inequalities for set covering that we use in our implementation were proposed by Balas and Ng [18]. To describe these inequalities, consider the set covering polytope $P_{SC}(D) = \operatorname{conv}\{\boldsymbol{y} \in \{0,1\}^m : D\boldsymbol{y} \geq \mathbb{1}\}$, where $D = (d_{ij}) \in \{0,1\}^{k\times m}$. Assume $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{y} \geq \beta$, with $\boldsymbol{a} \in \mathbb{Z}^m$ and $\beta \in \mathbb{Z}$, defines a facet of $P_{SC}(D)$. It is well known that if $\beta > 0$, then $\boldsymbol{a} \geq \boldsymbol{0}$, and if $\beta = 1$, then $\boldsymbol{a}$ is a row of D (see, e.g., [18]).

Balas and Ng showed that for every facet defining inequality $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{y} \geq 2$ with $\boldsymbol{a} \in \mathbb{Z}^n$, there exists a set $S \subseteq [k]$ such that $\boldsymbol{a} = \boldsymbol{a}^S$, where

$$a_j^S = \begin{cases} 0 & \text{if } d_{ij} = 0 \text{ for all } i \in S, \\ 2 & \text{if } d_{ij} = 1 \text{ for all } i \in S, \\ 1 & \text{otherwise} \end{cases} \qquad \text{for } j = 1, \ldots, m.$$

These inequalities can also be obtained by a Chvátal–Gomory rounding procedure. Furthermore, Balas and Ng discuss conditions under which $\boldsymbol{a}^{S\mathrm{T}}\boldsymbol{y} \geq 2$ defines a facet of $P_{SC}(D)$.

The separation problem for the above inequalities is NP-hard; see Amaldi and Pfetsch [11]. However, when the size of S is fixed, the separation problem can be solved in polynomial time by enumeration. In our implementation we enumerate sets S of *cardinality three* and check whether the inequalities $\boldsymbol{a}^{S\mathrm{T}}\boldsymbol{y} \geq 2$ are violated by the current LP-solution. Note that sets S of cardinality two are uninteresting, since in this case $\boldsymbol{a}^{S\mathrm{T}}\boldsymbol{y} \geq 2$ is the sum of two IIS-inequalities and hence is never violated if the IIS-inequalities are satisfied.

Additionally, we try to strengthen these cuts: If an inequality is violated, we greedily enlarge the set S as long as the violation of the resulting inequality increases. See section 4 for computational results.

3.8. General purpose inequalities. In our computational experiments we used Gomory (mixed integer) cuts as implemented in SCIP (see section 4); see the books of Nemhauser and Wolsey [48] or Schrijver [59] for a description.

Further, we used $\{0, \frac{1}{2}\}$-cuts introduced by Caprara and Fischetti [25]. Codato and Fischetti [33] identified these cuts as important for solving MIN IIS COVER. We implemented these cuts along the lines of Hansen, Labbé, and Schindl [38]. See also Andreello, Caprara, and Fischetti [13] for a computational study of $\{0, \frac{1}{2}\}$-cuts. Note that in our implementation $\{0, \frac{1}{2}\}$-cuts are produced only for set covering and nonnegativity inequalities; in particular, they do not depend on $\{0, \frac{1}{2}\}$-cuts produced earlier.

We also experimented with mixed integer rounding cuts (CMIR) (see Marchand and Wolsey [44]) and strengthened Chvátal–Gomory cuts (see Letchford and Lodi [42]) as they are implemented in SCIP. The results were, however, discouraging, and we therefore do not present them.

4. Computational results. In this section we discuss computational results of our branch-and-cut implementation for MIN IIS COVER. The algorithm was implemented in `C++` and uses version 0.90 of the framework SCIP by Achterberg [1].

CPLEX 10.11 is used as the basic LP solver. The computations were performed on a 3.4 GHz Pentium 4 machine with 3 GB of main memory and 1 MB cache running Linux. All instances used in the following can be obtained from the web page [45].

We use best-first search as a node selection scheme and the branching rule explained in section 3.6. All separation routines are called only every tenth level of the tree, except that the pool of IIS-inequalities is checked in every node of the tree. In nodes in which cuts are separated, we proceed until no more violated cuts can be found. SCIP chooses among the generated cuts according to an orthogonality measure; see, for instance, Andreello, Caprara, and Fischetti [13]. We perform reduced cost fixing at every node of the tree.

Before presenting computational results, we want to discuss the influence of the limited precision used for solving LPs. The basic question that has to be repeatedly answered in our context is whether a given system is infeasible or not. Today's LP solvers are tuned towards quickly finding an optimal solution of a feasible LP. Sometimes their bases are not really optimal, but this has only a negligible effect on the objective function value; see Koch [41]. When checking infeasibility, however, small errors can lead to completely wrong decisions. The answer depends on the particular instance, the solution method of the LP solver, its parameters, e.g., the precision (usually around 10^{-6}), and often also the preprocessing and starting basis. Being aware of the possibility that we might produce wrong results, as a safeguard, we confirmed that the final solution is really an IIS-cover for the original system.

Currently, using exact LP solvers, like the ones included in `lrs` [15] or `cdd` [34] is computationally too expensive. In the future, codes that use dynamically adjusted precision might help; see Applegate et al. [14].

4.1. The Netlib problems. The Netlib library [49] contains a well-known set of 29 infeasible linear inequality systems. We do not report results on these data since these instances all can be solved within seconds, except for numerical difficulties with the problem `gran`. They were also solved to optimality by Parker [51] and Parker and Ryan [52]; for more computational results on these problems, see Chinneck [31] and Pfetsch [53].

4.2. Random problems. We consider random inequality systems to compare different cut strategies in the branch-and-cut implementation. We used difficult random instances that nevertheless can be solved within approximately one hour of computation time. In contrast, the instances discussed in the following sections vary highly in size and complexity: Most are either solved within seconds or cannot be solved to optimality in reasonable time.

The infeasible random inequality systems are generated as follows: Each coefficient and the right-hand side were chosen to be a random integer in the range -100 to 100. We generated five instances for each of the combinations $(5, 100)$, $(10, 80)$, $(15, 80)$, $(20, 90)$, $(25, 90)$, where the first component is the dimension n of the space and the second one is the number m of inequalities. Each system turned out to be infeasible (this almost always happens as soon as $m > 2 \cdot n$; see Motzkin [46]) and is almost completely dense. Note that all the instances in the following sections are dense as well.

In Theorem 1, the alternative polyhedra of these random systems are nondegenerate with high probability. It is currently unknown whether Max FS and Min IIS Cover restricted to such systems are NP-hard.

TABLE 1

Results of the branch-and-cut algorithm on random inequality systems *for different IIS separation strategies. The numbers are averages over five instances of each size. The last line gives the averages over each column.*

		Single			Extend			Round		
n	m	Nodes	Time	IISs	Nodes	Time	IISs	Nodes	Time	IISs
5	100	70473.0	1050.64	8781.0	120371.4	1808.71	5281.4	16913.8	564.44	11034.8
10	80	167970.8	1226.45	10298.4	174302.6	1689.26	8450.4	79086.6	996.51	14491.8
15	80	214004.0	1509.72	53419.8	255933.0	1984.60	44825.8	106119.0	1465.16	62151.0
20	90	50029.0	276.05	22354.8	59117.8	337.11	15869.0	28699.0	317.22	23418.0
25	90	169868.2	1185.81	99728.6	243568.6	1534.17	80400.4	77147.0	1235.41	155331.4
	∅:	134469.0	1049.73	38916.5	170658.7	1470.77	30965.4	61593.1	915.75	53285.4

We first compare the three different strategies to separate IIS-inequalities of section 3.2. Table 1 provides a comparison of methods "single" (section 3.2.1), "extend" (section 3.2.2), and "round" (section 3.2.3). Columns labeled "nodes" give the average number of nodes in the branch-and-bound tree, those labeled "time" are the average CPU times in seconds, and those labeled "IISs" give the average number of IISs found during the optimization; here averages are taken over the five instances of each size. To eliminate the influence of primal heuristics we initialized all runs with the optimal solution.

Among the three IIS-inequality separation versions, method "round" outperforms methods "single" and "extend" in the number of nodes and in the total computation time, although method "single" is sometimes a bit faster. Method "round" also generates the highest number of IISs. Based on this result, we decided to use method "round" in the following experiments.

Table 2 shows the total number of IISs and the number of IISs found by method "round" for small random instances generated in the same manner as above. By Theorem 1, the IISs correspond to vertices of the alternative polyhedron. We enumerated the vertices with `lrs` [15]. Since the alternative polyhedra are nondegenerate, the IISs can be generated in time polynomial in the input and output size; see Avis and Fukuda [16]. Note that for general polyhedra this is not possible unless P = NP; see Khachiyan et al. [40].

We could not enumerate or count the IISs for larger instances. From Table 2, however, it can be expected that the total number of IISs for the instances used in Table 1 is much higher. We conclude that the branch-and-cut implementation needs only a small part of the total set of IISs (the number of IISs for instance $(5, 70)$ is two orders of magnitudes larger than the average number of IISs found by any of the variants in Table 1).

TABLE 2

The number of IISs found by method "round" for random problems and the total number of IISs.

n	m	Found	Total
5	30	11	1986
5	40	101	44816
5	50	520	204833
5	60	526	614853
5	70	453	1818718

TABLE 3

Results of the branch-and-cut algorithm on random inequality systems *for different cut generation strategies; all variants use method "round" as a basis. Given are the average values over all* 25 *instances.*

Type	Nodes	Time	Root	# BaNg	# Gom.	#$\{0, \frac{1}{2}\}$
round	61593.1	915.75	6.54	0.0	0.0	0.0
BaNg	58796.4	1054.39	6.80	6134.0	0.0	0.0
Gom.	58434.7	1164.56	7.00	0.0	10440.6	0.0
$\{0, \frac{1}{2}\}$	61479.1	957.37	6.54	0.0	0.0	43.0
BaNg & Gom.	57911.9	1298.49	7.22	6955.3	10234.8	0.0
BaNg & $\{0, \frac{1}{2}\}$	60197.0	1080.89	6.78	5738.6	0.0	31.0
Gom. & $\{0, \frac{1}{2}\}$	58852.8	1158.42	7.01	0.0	10441.2	56.8
all	60092.7	1365.63	7.19	6699.5	10335.6	46.2

Table 3 lists computational results for all combinations of method "round" with Balas/Ng cuts (BaNg), Gomory cuts (Gom.), and $\{0, \frac{1}{2}\}$-cuts. The values are averages over all 25 instances. Column "root" gives the dual bound after the root node. The last three columns list the number of cuts found for the respective methods. Again, we initialize the algorithms with the optimal solution. All cuts are separated every ten levels of the tree.

The studied combinations on average reduce the number of nodes with respect to the method "round" alone; the best combination in this respect are Balas/Ng and Gomory cuts. Furthermore, all combinations, except $\{0, \frac{1}{2}\}$-cuts, improve the root dual bound with respect to the basic version. The studied methods, however, increase the CPU time needed. The main slowdown comes from the fact that the intermediate LPs become harder to solve. The corresponding separation times are acceptable, however. The average separation times for the version that uses all three methods are 1.8% (BaNg), 17.0% (Gomory), 1.0% ($\{0, \frac{1}{2}\}$). We conclude that the basic version "round" alone is fastest on random systems.

Table 4 shows average results for method "round" on random instances with $m = 80$ inequalities. It can be observed that the optimal values of the random problems tend to decrease when increasing the dimension. This often makes the problems more tractable. But of course, the solution of the intermediate LPs over the alternative polyhedron is more time consuming.

4.3. Digital video broadcasting problems. In this section we present results for problems arising in an application of MAX FS in telecommunications, which is described by Rossi, Sassano, and Smriglio [54]. Here, to plan the digital video broadcasting (DVB) network of Italy, transmitters have to be placed and their emission

TABLE 4

Results of method "round" for random instances with $m = 80$ *inequalities. Column "Opt" gives the average optimal solution values. All entries are averages over five instances.*

n	Nodes	Time	IISs	Root	Opt
5	2029.4	32.26	3527.8	12.23	21.8
10	79086.6	996.51	14491.8	6.88	15.8
15	106119.0	1465.16	62151.0	4.56	11.8
20	7408.0	56.18	5743.4	2.69	5.8
25	16472.6	132.79	20884.0	2.43	6.8
∅:	42223.1	536.58	21359.6	5.76	12.4

TABLE 5
Results for the DVB instances in section 4.3 with method "round." The column labeled "[6]" lists the names of the instances as used in Amaldi, Belotti, and Hauser [6].

Name	[6]	m	Nodes	Time	IISs	Root	Dual	Best	Gap
dvb1	dvb2	1044	503	103.6	3064	166.4	174.0	174	0.0
mfs_UHF_P4_1	dvb1	642	1	2.3	86	104.0	104.0	104	0.0
mfs_UHF_P4_3	dvb3	1717	539	599.72	5414	174.2	183.0	183	0.0
mfs_UHF_P4_4	–	1174	68049	196514.41	1002912	90.3	115.2	124	7.6

frequency and power have to be chosen as to maximize the area coverage, subject to quality constraints. A subproblem of this can be modeled as a linear inequality system. Interference of the signals leads to areas where the digital signal cannot be received, resulting in an infeasible system. Maximizing the total weight of satisfied inequalities then amounts to maximizing the area coverage.

Linearizing the model leads to numerically challenging problems. The coefficients take values between 10^{-11} and 10^{11}, and the resulting LPs are very instable. We tackled the problems by scaling the original instances before starting the branch-and-cut algorithm. This helps but nevertheless leaves hard problems. Without scaling, however, the algorithm terminated early with a completely wrong solution.

We could compute optimal solutions for the smallest instances used in Amaldi, Belotti, and Hauser [6] and Amaldi, Bruglieri, and Casale [7]; see Table 5. Here, column "dual" gives the final lower bound, "best" denotes the value of the best primal solution obtained (i.e., the primal bound), and "gap" is the gap between the dual bound and primal bound in percent, i.e., (best − dual)/dual · 100.0. The dimension of these instances is always 487, and the variable bounds ($0 \le \boldsymbol{x} \le 1$) are mandatory. We separate $\{0, \frac{1}{2}\}$-cuts every 10th level of the tree. Our primal heuristic of section 3.4 is run every 40th level. Note that these instances can be solved faster using the "big-M" formulation (resulting in the same optimal solution values); see [6, 7].

4.4. Classification problems. One of the historically first applications of MIN IIS COVER is the design of linear classifiers; see Amaldi [4], Mangasarian [43], Bennett and Bredensteiner [19], and Rubin [55].

In this application, one is given m points $\boldsymbol{p}_1, \dots, \boldsymbol{p}_m$ in $\mathbb{R}^N$, each belonging to one of two possible *classes* P_1 and P_2; i.e., P_1 and P_2 partition the set $\{\boldsymbol{p}_1, \dots, \boldsymbol{p}_m\}$. Each of the N components of the points stores a measurement of an attribute (or feature) relevant for the concrete application. The goal is to strictly separate these points in $\mathbb{R}^N$ by an oriented hyperplane defined by $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} \le \beta$, with $\boldsymbol{a} \in \mathbb{R}^N$ and $\beta \in \mathbb{R}$. The points in P_1 should satisfy the inequality $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} < \beta$, and the points in P_2 should satisfy $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} > \beta$. Hence, we are looking for $(\boldsymbol{a}, \beta) \in \mathbb{R}^n$, with $n := N + 1$, so that the number of misclassified points

$$|\{\boldsymbol{p} \in P_1 \,:\, \boldsymbol{a}^{\mathrm{T}}\boldsymbol{p} \ge \beta\}| + |\{\boldsymbol{p} \in P_2 \,:\, \boldsymbol{a}^{\mathrm{T}}\boldsymbol{p} \le \beta\}|$$

is minimized. This minimization is performed in order to maximize the chance that a new point can be correctly classified. Note that with this formulation points in $\{\boldsymbol{x} \,:\, \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} = \beta\}$ are counted twice (the models can be modified to eliminate this).

In the following we will discuss two equivalent ways to model this problem via MIN IIS COVER and present computational results for different data sets. In the first model no bounds on the variables are present, while in the second all variables are bounded except one.

For the first model we use variables $(\boldsymbol{a}, \beta) \in \mathbb{R}^n$ and the following inequalities:

$$\boldsymbol{p}^{\mathrm{T}}\boldsymbol{a} - \beta \begin{cases} < 0 & \text{if } \boldsymbol{p} \in P_1 \\ > 0 & \text{if } \boldsymbol{p} \in P_2 \end{cases} \qquad \text{for each } \boldsymbol{p} \in \{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_m\}.$$

Since $(\boldsymbol{a}, \beta)$ are unbounded we can scale them to obtain

$$\boldsymbol{p}^{\mathrm{T}}\boldsymbol{a} - \beta \begin{cases} \leq -1 & \text{if } \boldsymbol{p} \in P_1 \\ \geq 1 & \text{if } \boldsymbol{p} \in P_2 \end{cases} \qquad \text{for each } \boldsymbol{p} \in \{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_m\}.$$

Of course, any other positive value instead of 1 can be taken in order to obtain a numerically more stable system.

The second model is due to Rubin [55]. It uses variables $\boldsymbol{a} \in \mathbb{R}^N$ and β, $\gamma \in \mathbb{R}$ in the following system:

$$\begin{aligned} \boldsymbol{p}^{\mathrm{T}}\boldsymbol{a} - \beta + \gamma \leq 0 \quad & \text{if } \boldsymbol{p} \in P_1, \\ \boldsymbol{p}^{\mathrm{T}}\boldsymbol{a} - \beta - \gamma \geq 0 \quad & \text{if } \boldsymbol{p} \in P_2, \\ -\mathbb{1} \leq \boldsymbol{a} \leq \mathbb{1}, & \\ \gamma \geq 0.001. & \end{aligned}$$

Hence, the coefficients of the normal vector $\boldsymbol{a}$ are bounded to lie within the interval $[-1, 1]$, while β is unbounded. Of course, the lower bound 0.001 for γ can be replaced by any suitably small positive number. For instances arising from this model the variable bounds are mandatory.

Note that in both models it might happen that the systems are feasible, i.e., the points are completely separable (in which case we need only solve one LP).

In our first test we use the first model and classification data from the UCI Repository of Machine Learning Databases (Blake and Merz [21]). The problem characteristics are given in Table 6. For some instances we had to remove incomplete data sets. A complete description of the instances is available at the UCI Repository.

TABLE 6

Characteristics of the classification instances. *Column "N" lists the number of attributes. The column labeled "$m^\star$" gives the number of original data sets and column "m" gives the number of data sets remaining after removing incomplete ones. The rightmost column gives additional notes, e.g., the name of the instance in the UCI database.*

Name	N	m	$m^\star$	Notes
breast-cancer	9	683	699	breast-cancer-wisconsin
bupa	6	345	345	liver-disorders
echo	8	61	132	echocardiogram
glass	9	214	214	type 2 vs. others
heart	13	297	303	heart-disease (Cleveland)
ionosphere	34	351	351	
iris.1	4	150	150	Versicolor vs. others
iris.2	4	150	150	Virginica vs. others
new-thyroid	5	215	215	normal vs. others
pima	8	768	768	Pima-indians-diabetes
tic-tac-toe	9	958	958	
wpbc	32	194	198	Wisconsin breast-cancer database

TABLE 7
Results of the branch-and-cut algorithm for the classification *instances.*

Name	Nodes	Time	IISs	Root	Dual	Best	Gap	Chi
breast-cancer	313	2.88	359	7.2	11.0	11	0.0	11
bupa	9669	18000.11	179562	43.2	59.6	83	39.3	83
echo	2	0.05	89	6.0	6.0	6	0.0	6
glass	36859	18000.00	99833	18.5	32.7	36	10.0	41
heart	51274	18000.02	122000	12.8	23.5	29	23.6	30
ionosphere	2465	38.59	3967	2.4	6.0	6	0.0	6
iris.1	845	12.45	623	19.1	25.0	25	0.0	25
iris.2	1	0.01	2	0.0	1.0	1	0.0	1
new-thyroid	2	0.09	147	11.0	11.0	11	0.0	11
pima	1522	18000.18	64166	68.2	75.6	148	95.7	148
tic-tac-toe	50691	5167.03	19850	60.9	86.0	86	0.0	93
wpbc	56657	18000.00	739494	3.5	8.7	13	48.7	13

Most of these twelve instances are also used by Chinneck [31] for testing his heuristic for MAX FS/MIN IIS COVER.

Table 7 lists the results of the branch-and-cut implementation on these instances with method "round" of section 3.2.3. The computation time was limited to *five hours* (18000 sec.). The columns have the same meaning as in sections 4.2 and 4.3.

Column "Chi" gives results obtained by the heuristic of Chinneck (see section 3.4); its running times are negligible and therefore not listed. Our implementation found the same solutions as Chinneck [31], except for the instances `glass` and `wpbc`, for which Chinneck obtained solutions of size 39 and 10, respectively. Our primal heuristic described in section 3.4 is run every tenth level. It could improve the initial solutions for models `glass`, `heart`, and `tic-tac-toe`. We conclude that the heuristic of Chinneck generates very good starting solutions, while our primal heuristic sometimes helps to find better solutions.

The results of Table 7 show that most instances are quite hard to solve and about half of them could not be solved within the time bound of five hours. Because of their size, only few nodes could be processed.

We also conducted experiments with the same data but using the second model instead of the first. Intuitively this should result in better numerical properties of the LPs that have to be solved during the algorithm. The results are, however, comparable to the ones shown in Table 7, and we therefore do not present them here.

Table 8 compares the gaps of the different cut strategies. The table displays only instances for which the optimal solutions could not be found within five hours. It turns

TABLE 8
Classification problems: *Comparison of the* gaps *of different variants of cutting planes. Only instances for which a positive gap remains after five hours are shown. The notation is as in Table* 3. *The last line contains the averages over each column.*

Name	Round	BaNg	Gom.	$\{0, \frac{1}{2}\}$	BaNg Gom.	BaNg $\{0, \frac{1}{2}\}$	Gom. $\{0, \frac{1}{2}\}$	all
bupa	39.3	46.7	40.0	41.9	44.0	45.0	41.5	45.5
glass	10.0	12.7	10.0	10.6	12.2	12.7	9.8	12.4
heart	23.6	23.8	22.6	24.2	25.4	25.2	27.8	26.0
pima	95.7	101.8	95.0	98.6	103.2	101.4	94.1	105.4
wpbc	48.7	47.8	44.6	49.8	49.0	49.0	45.6	50.5
∅:	43.5	46.6	42.4	45.0	46.7	46.7	43.8	48.0

out that all variants find the same final primal solutions, although at different times during the computation. Note that this actually compares the interplay of cutting strategies and our primal heuristic. On the average, the smallest gaps are produced by taking Gomory cuts, then method "round," Gomory and $\{0, \frac{1}{2}\}$-cuts, $\{0, \frac{1}{2}\}$-cuts alone, Balas/Ng cuts, Balas/Ng cuts and Gomory cuts, Balas/Ng cuts and $\{0, \frac{1}{2}\}$-cuts, and finally all cuts together. The main reason why all cuts together produce the worst results (on average) is that this combination could explore the fewest number of nodes. We conclude that the additional cutting planes do not yield a big improvement over method "round" alone. Although Gomory cuts produce the smallest gaps, the studied cutting planes do not seem to be crucial for solving these instances.

Our second test set consists of data from Codato and Fischetti [33] and uses the second model. The data again originate from the UCI Repository of Machine Learning Databases, but are preprocessed in a way we could not reconstruct. Hence, the results for these instances and the instances of Table 6 may not be comparable (there are three instances which seem to arise from the same original data: `breast-cancer` ↔ `breast-cancer-2`, `iris.1` ↔ `iris-150`, `wpbc` ↔ `WPBC194`). Instances `Breast-Cancer-2` and `Breast-Cancer-400` seem to be different from those used in Codato and Fischetti [33].

Table 9 shows the results of method "round" on these instances. The notation is as in Table 3. Note that here the dimension is $n = N + 2$, because we use the second model. Most of the instances could be solved within a few seconds. This is the first time that the complete set could be solved to optimality: No optimal solution to the harder instances (`Flags-169`, `Horse-colic-185`, `Horse-colic-253`, and `Solar-flare-1066`) was previously available. Our implementation solves all instances except these four in under a minute. Although we worked on a faster computer, it nevertheless seems fair to say that our code considerably improves upon the results of Codato and Fischetti [33].

5. Conclusions. In this paper we described a branch-and-cut implementation for the Max FS/Min IIS Cover problem, which is the best exact method currently available. The findings of the extensive computational results can be roughly summarized as follows: With respect to the implementation, the best cutting plane strategy is to find as many (violated) IIS-inequalities as possible. Additionally applying Balas/Ng, Gomory, or $\{0, \frac{1}{2}\}$-cuts does not significantly help to improve the performance: On random instances they do not improve the running time, but usually help to reduce the number of nodes. Gomory cuts only slightly help to reduce the gaps for classification instances, and the other cuts do not improve the gap.

With respect to the problem data, the considered instances vary greatly in their properties and difficulty. Depending on the particular data, quite large instances can be solved to optimality, but there are also relatively small instances which turn out to be extremely hard to solve. As shown by the DVB problems, one has to be careful with numerically instable instances.

An interesting open issue is the existence of problem-specific cutting planes and whether they can be efficiently separated. Another question is whether other valid inequalities for the set covering problem could help improve the performance of the implementation.

Acknowledgments. The author thanks Tobias Achterberg for help with the SCIP implementation and Edoardo Amaldi and Les Trotter for helpful discussions. Furthermore, he thanks Edoardo Amaldi and Pietro Belotti for providing the DVB instances of section 4.3, and Gianni Codato and Matteo Fischetti for the data used in section 4.4.

TABLE 9
Classification problems: *Results of the branch-and-cut algorithm for the problems of Codato and Fischetti with method "round."*

Name	n	m	Nodes	Time	IISs	Root	Opt
Balloons-76	7	76	1	0.02	59	10.0	10
BCW-367	12	367	110	0.97	252	5.5	8
BCW-683	12	683	71	1.70	235	6.8	10
Breast-Cancer-2	11	683	352	2.21	322	7.0	11
Breast-Cancer-400	20	400	2	0.08	116	24.0	24
Bridges-132	14	132	299	3.44	1563	20.2	23
BusVan-437	20	437	237	1.72	353	3.0	6
BusVan-445	20	445	605	5.53	750	3.3	8
BusVan-447	20	447	2334	37.65	4187	4.4	10
BV-OS-282	20	282	214	1.39	338	3.0	6
BV-OS-376	20	376	969	12.03	1361	4.2	9
Chorales-107	8	107	951	9.57	1187	21.4	27
Chorales-116	8	116	1022	19.85	1981	17.2	24
Chorales-134	8	134	1198	50.99	4008	20.8	30
Credit-300	17	300	13	0.93	222	5.9	8
Flag-169	31	169	7621	209.63	17276	3.5	9
Glass-163	12	163	15	0.64	158	10.9	13
Horse-Colic-151	28	151	231	2.25	540	2.2	5
Horse-Colic-185	28	183	69155	886.10	61414	3.6	10
Horse-Colic-253	28	253	273389	7938.84	308862	4.8	13
House-Votes84-435	18	435	56	0.68	200	4.0	6
Iris-150	7	150	1017	6.58	1011	11.7	18
Lymphography-142	20	142	21	0.24	131	2.9	5
Mech-analysis-107	10	107	1	0.04	83	7.0	7
Mech-analysis-137	9	137	757	5.83	890	11.6	18
Mech-analysis-152	10	152	900	32.05	3042	13.0	21
Monks-tr-115	8	115	917	16.24	1570	20.9	27
Monks-tr-122	8	122	4	0.45	267	11.2	13
Monks-tr-124	8	124	489	5.91	1187	18.1	24
Opel-Saab-76	20	76	1111	9.28	1756	2.9	7
Opel-Saab-80	20	80	241	2.01	512	3.0	6
Opel-Saab-83	20	83	2113	25.05	3904	3.2	8
Opel-Saab-84	20	84	572	7.06	1318	3.3	7
Pb-gr-txt-198	12	198	147	1.09	267	7.7	11
Pb-hl-pict-277	12	277	178	1.61	314	6.7	10
Pb-pict-txt-444	12	444	2	0.12	79	7.0	7
Postoperative-88	10	88	1	0.12	209	16.0	16
Solar-flare-323	14	323	3	0.71	478	37.2	38
Solar-flare-1066	14	1066	2292	787.64	14960	227.3	243
Water-treat-206	40	206	41	1.43	204	1.7	4
Water-treat-213	40	213	288	8.04	845	2.2	5
WPBC-194	36	194	172	3.21	468	2.2	5

REFERENCES

[1] T. ACHTERBERG, *SCIP—A Framework to Integrate Constraint and Mixed Integer Programming*, Report 04–19, Zuse Institute Berlin, Berlin, 2004, http://www.zib.de/Publications/abstracts/ZR-04-19/.

[2] T. ACHTERBERG, T. KOCH, AND A. MARTIN, *Branching rules revisited*, Oper. Res. Lett., 33 (2005), pp. 42–54.

[3] S. AGMON, *The relaxation method for linear inequalities*, Canad. J. Math., 6 (1954), pp. 382–392.

[4] E. AMALDI, *From Finding Maximum Feasible Subsystems of Linear Systems to Feedforward Neural Network Design*, Ph.D. thesis, EPF-Lausanne, Lausanne, Switzerland, 1994.

[5] E. Amaldi, *The maximum feasible subsystem problem and some applications*, in Modelli e Algoritmi per l'Ottimizzazione di Sistemi Complessi, A. Agnetis and G. D. Pillo, eds., Pitagora Editrice, Bologna, Italy, 2003, pp. 31–69.

[6] E. Amaldi, P. Belotti, and R. Hauser, *Randomized relaxation methods for the maximum feasible subsystem problem*, in Integer Programming and Combinatorial Optimization, Lecture Notes in Comput. Sci. 3509, M. Jünger and V. Kaibel, eds., Springer-Verlag, Berlin, Heidelberg, 2005, pp. 249–264.

[7] E. Amaldi, M. Bruglieri, and G. Casale, *A two-phase relaxation-based heuristic for the maximum feasible subsystem problem*, Comput. Oper. Res., 35 (2008), pp. 1377–1756.

[8] E. Amaldi and R. Hauser, *Boundedness theorems for the relaxation method*, Math. Oper. Res., 30 (2005), pp. 1–17.

[9] E. Amaldi and V. Kann, *The complexity and approximability of finding maximum feasible subsystems of linear relations*, Theoret. Comput. Sci., 147 (1995), pp. 181–210.

[10] E. Amaldi and V. Kann, *On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems*, Theoret. Comput. Sci., 209 (1998), pp. 237–260.

[11] E. Amaldi and M. E. Pfetsch, *Separation Problems for Set Covering*, manuscript, 2005.

[12] E. Amaldi, M. E. Pfetsch, and L. E. Trotter, Jr., *On the maximum feasible subsystem problem, IISs, and IIS-hypergraphs*, Math. Program., 95 (2003), pp. 533–554.

[13] G. Andreello, A. Caprara, and M. Fischetti, *Embedding* $\{0, \frac{1}{2}\}$*-cuts in a branch-and-cut framework: A computational study*, INFORMS J. Comput., 19 (2007), pp. 229–238.

[14] D. L. Applegate, W. Cook, S. Dash, and D. G. Espinoza, *Exact solutions to linear programming problems*, Oper. Res. Lett., 35 (2007), pp. 693–699.

[15] D. Avis, `lrs` *Home Page*, http://cgm.cs.mcgill.ca/∼avis/C/lrs.html (2005).

[16] D. Avis and K. Fukuda, *Reverse search for enumeration*, Discrete Appl. Math., 65 (1996), pp. 21–46.

[17] E. Balas, S. Ceria, G. Cornuéjols, and N. Natraj, *Gomory cuts revisited*, Oper. Res. Lett., 19 (1996), pp. 1–9.

[18] E. Balas and S. M. Ng, *On the set covering polytope*. I. *All the facets with coefficients in* $\{0, 1, 2\}$, Math. Programming, 43 (1989), pp. 57–69.

[19] K. P. Bennett and E. J. Bredensteiner, *A parametric optimization method for machine learning*, INFORMS J. Comput., 9 (1997), pp. 311–318.

[20] K. P. Bennett and O. L. Mangasarian, *Neural network training via linear programming*, in Advances in Optimization and Parallel Computing, P. M. Pardalos, ed., North–Holland, Amsterdam, 1992, pp. 56–67.

[21] C. L. Blake and C. J. Merz, *UCI Repository of Machine Learning Databases*, http://www.ics.uci.edu/∼mlearn/MLRepository.html (1998).

[22] R. Borndörfer, *Aspects of Set Packing, Partitioning, and Covering*, Ph.D. thesis, TU Berlin, Berlin, 1998.

[23] R. Borndörfer and R. Weismantel, *Set packing relaxations of some integer programs*, Math. Program., 88 (2000), pp. 425–450.

[24] R. Borndörfer and R. Weismantel, *Discrete relaxations of combinatorial programs*, Discrete Appl. Math., 112 (2001), pp. 11–26.

[25] A. Caprara and M. Fischetti, $\{0, \frac{1}{2}\}$*-Chvátal-Gomory cuts*, Math. Programming, 74 (1996), pp. 221–235.

[26] A. Caprara and M. Fischetti, *Branch-and-cut algorithms*, in Annotated Bibliographies in Combinatorial Optimization, M. Dell'Amico, F. Maffioli, and S. Martello, eds., John Wiley & Sons, Chichester, UK, 1997, pp. 45–63.

[27] S. Ceria, P. Nobili, and A. Sassano, *Set covering problem*, in Annotated Bibliographies in Combinatorial Optimization, M. Dell'Amico, F. Maffioli, and S. Martello, eds., John Wiley & Sons, Chichester, UK, 1997, pp. 415–428.

[28] N. Chakravarti, *Some results concerning post-infeasibility analysis*, European J. Oper. Res., 73 (1994), pp. 139–143.

[29] J. W. Chinneck, *An effective polynomial-time heuristic for the minimum-cardinality IIS set-covering problem*, Ann. Math. Artif. Intell., 17 (1996), pp. 127–144.

[30] J. W. Chinneck, *Finding a useful subset of constraints for analysis in an infeasible linear program*, INFORMS J. Comput., 9 (1997), pp. 164–174.

[31] J. W. Chinneck, *Fast heuristics for the maximum feasible subsystem problem*, INFORMS J. Comput., 13 (2001), pp. 210–223.

[32] J. W. Chinneck and E. W. Dravnieks, *Locating minimal infeasible constraint sets in linear programs*, ORSA J. Comput., 3 (1991), pp. 157–168.

[33] G. Codato and M. Fischetti, *Combinatorial Benders' cuts*, in Integer Programming and Combinatorial Optimization, Lecture Notes in Comput. Sci. 3064, D. Bienstock and G. Nemhauser, eds., Springer-Verlag, Berlin, Heidelberg, 2004, pp. 178–195.

[34] K. Fukuda, `cdd` *Home Page*, http://www.cs.mcgill.ca/∼fukuda/soft/cdd_home/cdd.html (25 August 2005).
[35] J. Gleeson and J. Ryan, *Identifying minimally infeasible subsystems of inequalities*, ORSA J. Comput., 2 (1990), pp. 61–63.
[36] H. J. Greenberg and F. H. Murphy, *Approaches to diagnosing infeasible linear programs*, ORSA J. Comput., 3 (1991), pp. 253–261.
[37] M. Grötschel, L. Lovász, and A. Schrijver, *Geometric Algorithms and Combinatorial Optimization*, Algorithms Combin. 2, 2nd ed., Springer-Verlag, Heidelberg, 1993.
[38] P. Hansen, M. Labbé, and D. Schindl, *Set Covering and Packing Formulations of Graph Coloring: Algorithms and First Polyhedral Results*, Technical report, GERAD, Montreal, Canada, 2005.
[39] D. S. Johnson and F. P. Preparata, *The densest hemisphere problem*, Theoret. Comput. Sci., 6 (1978), pp. 93–107.
[40] L. Khachiyan, E. Boros, K. Borys, K. Elbassioni, and V. Gurvich, *Generating all vertices of a polyhedron is hard*, in Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2006), ACM, New York, SIAM, Philadelphia, 2006, pp. 758–765.
[41] T. Koch, *The final* Netlib-*LP results*, Oper. Res. Lett., 32 (2004), pp. 138–142.
[42] A. N. Letchford and A. Lodi, *Strengthening Chvátal-Gomory cuts and Gomory fractional cuts*, Oper. Res. Lett., 30 (2002), pp. 74–82.
[43] O. L. Mangasarian, *Misclassification minimization*, J. Global. Optim., 5 (1994), pp. 309–323.
[44] H. Marchand and L. Wolsey, *Aggregation and mixed integer rounding to solve MIPS*, Oper. Res., 49 (2001), pp. 363–371.
[45] *Maximum Feasible Subsystem Problem Home Page*, http://risorse.del.polimi.it/maxfs/ (1 July 2006).
[46] T. S. Motzkin, *The probability of solvability of linear inequalities*, in Selected Papers, Contemp. Mathematicians, D. Cantor, B. Gordon, and B. Rothschild, eds., Birkhäuser Boston, Boston, 1983, pp. 116–120.
[47] T. S. Motzkin and I. J. Schoenberg, *The relaxation method for linear inequalities*, Canad. J. Math., 6 (1954), pp. 393–404.
[48] G. L. Nemhauser and L. A. Wolsey, *Integer and Combinatorial Optimization*, John Wiley & Sons, New York, 1988.
[49] *Netlib*, http://www.netlib.org.
[50] M. Padberg and G. Rinaldi, *A branch-and-cut algorithm for the resolution of large-scale symmetric traveling salesman problems*, SIAM Rev., 33 (1991), pp. 60–100.
[51] M. Parker, *A Set Covering Approach to Infeasibility Analysis of Linear Programming Problems and Related Issues*, Ph.D. thesis, University of Colorado at Denver, Denver, 1995.
[52] M. Parker and J. Ryan, *Finding the minimum weight IIS cover of an infeasible system of linear inequalities*, Ann. Math. Artif. Intell., 17 (1996), pp. 107–126.
[53] M. E. Pfetsch, *The Maximum Feasible Subsystem Problem and Vertex-Facet Incidence of Polyhedra*, Ph.D. thesis, TU Berlin, Berlin, 2002.
[54] F. Rossi, A. Sassano, and S. Smriglio, *Models and algorithms for terrestrial digital broadcasting*, Ann. Oper. Res., 107 (2001), pp. 267–283.
[55] Rubin, *Solving mixed integer classification problems by decomposition*, Ann. Oper. Res., 74 (1997), pp. 51–64.
[56] D. M. Ryan and B. A. Foster, *An integer programming approach to scheduling*, in Computer Scheduling of Public Transport: Urban Passenger Vehicle and Crew Scheduling, A. Wren, ed., North–Holland, Amsterdam, 1981, pp. 269–280.
[57] J. Ryan, *Transversals of IIS-hypergraphs*, in Proceedings of the 22nd Southeast Conference on Combinatorics, Graph Theory, and Computing (Baton Rouge, 1991), Congr. Numer., 81, 1991, pp. 17–22.
[58] J. K. Sankaran, *A note on resolving infeasibility in linear programs by constraint relaxation*, Oper. Res. Lett., 13 (1993), pp. 19–20.
[59] A. Schrijver, *Theory of Linear and Integer Programming*, John Wiley & Sons, Chichester, UK, 1986.
[60] S. Thienel, *ABACUS—A Branch-And-CUt System*, Ph.D. thesis, Universität zu Köln, Köln, Germany, 1995.
[61] M. Wagner, J. Meller, and R. Elber, *Solving huge linear programming problems for the design of protein folding potentials*, Math. Program., 101 (2004), pp. 301–318.

SIAM J. OPTIM.
Vol. 19, No. 1, pp. 39–61

LAGRANGIAN-DUAL FUNCTIONS AND MOREAU–YOSIDA REGULARIZATION*

FANWEN MENG†, GONGYUN ZHAO‡, MARK GOH§, AND ROBERT DE SOUZA†

Abstract. In this paper, we consider the Lagrangian-dual problem of a class of convex optimization problems. We first discuss the semismoothness of the Lagrangian-dual function φ. This property is then used to investigate the second-order properties of the Moreau–Yosida regularization η of the function φ, e.g., the semismoothness of the gradient g of the regularized function η. We show that φ and g are piecewise C^2 and semismooth, respectively, for certain instances of the optimization problem. We establish a relationship between the original problem and the Fenchel conjugate of the regularization of the corresponding Lagrangian dual problem. We also find some instances of the optimization problem whose Lagrangian-dual function φ is not piecewise smooth. However, its regularized function still possesses nice second-order properties. Finally, we provide an alternative way to study the semismoothness of the gradient under the structure of the epigraph of the dual function.

Key words. Lagrangian dual, Moreau–Yosida regularization, piecewise C^k functions, semismoothness, Fenchel conjugate

AMS subject classifications. 90C25, 65K10, 52A41

DOI. 10.1137/060673746

1. Introduction.

Consider the following convex program:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & Ax = a, \\ & f_i(x) \le 0, \quad i \in \hat{I} = \{1, 2, \ldots, \theta\}, \end{aligned} \tag{1}$$

where f, f_i, $i = 1, 2, \ldots, \theta$, are smooth and convex on $\mathbb{R}^n$, and where $A \in \mathbb{R}^{m\times n}$ with $\operatorname{rank}(A) = m$ and $0 < m < n$.

It is known that many practical problems can be converted to problem (1) above. For instance, some recently studied multistage stochastic programming models can be formulated as (1). See [19, Chapter 1] for the detailed modeling in this regard.

Let $\mathcal{F} := \{x \in \mathbb{R}^n : f_i(x) \le 0, i \in \hat{I}\}$. In many circumstances, particularly in multistage stochastic programming, f and $\mathcal{F}$ are separable, while the constraint $Ax = a$ is nonseparable. Thus, we seek to relax the constraint $Ax = a$ using the Lagrangian dual of problem (1) as follows:

$$\min\{\varphi(v) \mid v \in \mathbb{R}^m\}, \tag{2}$$

*Received by the editors October 30, 2006; accepted for publication (in revised form) August 7, 2007; published electronically February 6, 2008. The work of the first, third, and fourth authors was supported in part by the Economic Development Board of Singapore Research grant R-385-000-019-414.

http://www.siam.org/journals/siopt/19-1/67374.html

†The Logistics Institute—Asia Pacific, National University of Singapore, 7 Engineering Drive 1, Singapore 117574 (tlimf@nus.edu.sg, tlirbrtm@nus.edu.sg).

‡Department of Mathematics, National University of Singapore, 2 Science Drive 2, Singapore 117543 (matzgy@nus.edu.sg). The work of this author was supported in part by the National University of Singapore Academic Research grant R-146-000-057-112.

§Department of Decision Sciences, National University of Singapore, 1 Business Link, Singapore 117592, and The Logistics Institute—Asia Pacific, National University of Singapore, 7 Engineering Drive 1, Singapore 117574 (bizgohkh@nus.edu.sg, tligohkh@nus.edu.sg).

where

$$\varphi(v) = \sup\{-f(x) + v^T(Ax - a) \mid x \in \mathcal{F}\}. \tag{3}$$

In these circumstances, the subproblem in (3) is separable, and then is solvable through the well-developed parallel algorithms. This makes the evaluation of φ much easier in general. However, an obstacle in solving problem (2) is that the function φ is nondifferentiable. To overcome this and noticing that the underlying function φ is convex on $\mathbb{R}^m$, we then use the well-known regularization of Moreau [14] and Yosida [23] to convert (2) into a smooth problem as follows:

$$\min\{\eta(v) \mid v \in \mathbb{R}^m\}, \tag{4}$$

where η is the Moreau–Yosida regularization of φ as defined below,

$$\eta(v) = \min_{w \in \mathbb{R}^m} \left\{ \varphi(w) + \frac{1}{2}||w - v||_M^2 \right\}, \quad v \in \mathbb{R}^m, \tag{5}$$

M is a symmetric positive definite $m \times m$ matrix, and $\|v\|_M^2 = v^T M v$ for any $v \in \mathbb{R}^m$.

It is well known that the set of minimizers of problem (4) is exactly the set of minimizers of (2). It can be shown that η is continuously differentiable and that its gradient $g = \nabla\eta$ is globally Lipschitz continuous with modulus $||M||$. For the properties of the Moreau–Yosida regularization, the reader is referred to [8, 7]. For the problem discussed in the present paper, there are some advantages of using the Moreau–Yosida regularization, as given next.

Fukushima and Qi [7] have shown that superlinear convergence can be guaranteed by using approximate solutions of the problem (5) to construct search directions for minimizing η. While finding an exact solution for a nonsmooth function φ is difficult, the computation of an approximate solution is relatively easier. We can, e.g., consider a parameterized function $\varphi(w, \mu)$, where $\varphi(w, \mu) \to \varphi(w)$ as $\mu \to 0$ and $\varphi(w, \mu)$ is smooth for any $\mu > 0$ as in the case of the barrier function method. This method was utilized for solving multistage stochastic nonlinear problems recently in [24], in which the underlying stochastic problem was formulated as problem (1). For any prescribed accuracy, we can now choose an appropriate $\mu > 0$ such that the minimizer of $\varphi(w, \mu) + (1/2)\|w - v\|_M^2$ is a desirable approximate solution to (5).

It is interesting that both the parameterized function $\varphi(\cdot, \mu)$ and the regularized function η (without parameter) are used to smooth the nonsmooth function φ. However, they function in different ways and have different properties: the former is successful in global convergence, while the latter can speed up local convergence. Incorporating the parameterizations into the Moreau–Yosida regularization can be a way to combine advantages in both approaches.

Besides the parameterizations mentioned above, there are many other methods for computing approximate minimizers of φ. Each of these methods can be incorporated into the Moreau–Yosida regularization, giving rise to an enhanced method for minimizing the nonsmooth function φ. Hence establishing the theoretical framework of the Moreau–Yosida regularization can benefit a variety of algorithms.

For the problem under consideration, one of the most important properties about the Moreau–Yosida regularization is the semismoothness of the gradient of the regularized function, which has played a key role in establishing the superlinear convergence of the generalized Newton method for nonsmooth convex problems by combining the Moreau–Yosida regularization scheme in (5) [7].

The concept of semismooth functions, an important subclass of Lipschitz functions, was first introduced by Mifflin [12]. In order to study the superlinear convergence of Newton method for solving nondifferentiable equations, Qi and Sun [16] extended the definition of semismoothness to vector-valued functions. After the work of Qi and Sun, semismoothness was extensively used to establish superlinear/quadratic convergence of Newton's method for solving the convex best interpolation problem [4, 5], nondifferentiable equations in which the underlying functions are slant differentiable functions [1], and complementarity problems and variational inequalities [6], for instance.

In this paper, we will focus on a special case of semismooth functions, *piecewise C^k functions*, which is a large class of locally Lipschitz continuous functions, found in most practical problems [20, 17]. In the past few years, many people have studied the piecewise smoothness of nonsmooth functions and designed algorithms based on Newton's method for solving the associated nonsmooth equations or nonlinear optimization problems. For example, the analysis was mainly focused on the concept of piecewise C^k functions in [10, 13, 21], where the authors have considered properties of g for some specific classes of φ. Specifically, Sun and Han [21] showed the semismoothness of g if φ is the maximum of several twice continuously differentiable convex functions under a constant rank constraint qualification (CRCQ). Later, Meng and Hao [10] derived the same result for the case of unconstrained problem (1) with the objective function f being a piecewise C^2 function under a weaker sequential constant rank constraint qualification. In [13], Mifflin, Qi, and Sun investigated the case where φ is piecewise C^2 which is a generalization of the maximum of convex C^2 functions under a so-called affine independence preserving constraint qualification (AIPCQ).

Having motivated the importance of the notions of semismoothness and the Moreau–Yosida regularization in nonsmooth analysis, in this paper we will investigate properties of the Lagrangian-dual function φ and the gradient of its Moreau–Yosida regularization η. Further, studying the properties of Lagrangian-dual function φ has its own interest as well; see [22] and the references therein, for instance. Since piecewise smooth functions as a special class of semismooth functions possess more enjoyable properties than semismooth functions [12, 16, 17, 20], we will concentrate on the study of piecewise smoothness of φ and the gradient g of the regularized function η in the context. We have adopted two different methods in analyzing properties of g. In terms of the first method, the main tool used in this study is based on Proposition 1 (see section 2), which was established by Mifflin, Qi, and Sun [13] using the notion of piecewise smoothness. We will first study the piecewise smoothness of φ. This property will then be used to show the semismoothness of g. For the problem with the linear objective function $f(x) = c^T x$, we can show that the function φ is piecewise C^2 and satisfies AIPCQ, and thus g is semismooth by Proposition 1 if all f_i's are affine functions or all $\nabla^2 f_i$'s are positive definite. We also present an example whose region $\mathcal{F}$ is defined by a linear constraint and a strictly convex constraint. In this example, the function φ is, surprisingly, *not* piecewise C^2, and, equally surprisingly, the gradient g of the regularization of φ *is* still semismooth. For general convex objective functions f and constraint functions f_j, it is completely unknown how smooth φ and g should be. This issue is considered by analyzing some special cases where the objective function possesses a positive definite Hessian. The second method is mainly based on the metric projection operator under the structure of the epigraph of the Lagrangian-dual function. Using the projection mapping, the study of the properties of g is equivalently converted to the study of the properties of solutions

to a system of nonsmooth equations. The analysis is basically based on the framework established by Meng, Sun, and Zhao [11] recently. The results obtained complement and enrich the framework of piecewise smooth functions [20, 17] and also enhance the recent results on the Moreau–Yosida regularization [11].

Another topic of interest is the study of the duality of the original problem (1). It is well known that the duality theory is a fundamental issue in optimization both theoretically and numerically. For problem (1) with a linear objective, we derive an interesting result regarding the original problem and the Fenchel conjugate of Moreau–Yosida regularization of its Lagrangian-dual function, characterizing a relationship between the conjugate and the Lagrangian-dual. This provides a new way to look at the Lagrangian-dual and the Moreau–Yosida regularization. We believe that the established results complement the dual theory in optimization, particularly the theory of Magnanti [9] to some extent.

The rest of the paper is organized as follows. In section 2, basic definitions and properties are collected. The analysis of problems with the linear objective functions covers the next two sections. Section 3 investigates the piecewise smoothness of the function φ. Section 4 studies the semismoothness of the gradient g and the conjugate of the Moreau–Yosida regularization. Illustrative examples are presented in sections 3 and 4. Section 5 discusses the case of general convex objective functions. Section 6 concludes.

2. Preliminaries. In this section, we briefly recall some concepts, such as semismoothness, piecewise smoothness, and AIPCQ, which will be used in the rest of this paper.

It is known that the regularized function η is a continuously differentiable convex function defined on $\mathbb{R}^m$, even though φ may be nondifferentiable. The gradient of η at v (see [8]) is

$$g(v) \equiv \nabla\eta(v) = M(v - p(v)), \qquad v \in \mathbb{R}^m, \tag{6}$$

where $p(v)$ represents the unique solution of the minimization problem in (5). In order to use Newton method or modified Newton's methods for solving (4), it is important to study the Hessian of η, i.e., the Jacobian of g. Note that, in general, g may not be differentiable. To extend the definition of Jacobian to certain classes of nonsmooth functions, Qi and Sun [16] introduced the definition of semismoothness [12] for vector-valued functions. See [16] for details.

A remarkable feature of semismoothness is that superlinear or quadratic convergence of a generalized Newton method for solving nonsmooth equations can be obtained under the assumption of semismoothness. See [7, 15, 16] for the relevant discussions. Note that in general a direct verification of semismoothness is difficult. Some equivalent definitions of semismooth functions and further studies on semismoothness can be found in [11, 15] and the references therein. As for the underlying Lagrangian dual function φ, it has a special feature; i.e., φ is piecewise smooth. We shall make use of this special feature to investigate the semismoothness of g in the subsequent analysis. We now give a definition of piecewise smooth functions below, which is slightly different from the one given in [20].

DEFINITION 1. *A continuous function $\psi : \mathbb{R}^n \to \mathbb{R}^l$ is said to be a* piecewise C^k *function on a set $D \subseteq \mathbb{R}^n$ if there exist a finite index set $I = \{1, \ldots, q\}$, closed sets $D_1, \ldots, D_q$, open sets $U_1, \ldots, U_q$ (or relatively open with respect to the affine hull of D), and functions $\psi_1, \ldots, \psi_q$ such that*

(i) *$D \subseteq \cup_{j=1}^{q} D_j$ and $D_j \subseteq U_j$ for each $j \in I$,*

(ii) $\psi_j \in C^k(U_j)$ *for each* $j \in I$,
(iii) $\psi(u) = \psi_j(u)$ *for any* $u \in D \cap D_j$ *and each* $j \in I$.
We refer to $\{(D_j, U_j, \psi_j)\}_{j\in I}$ *as a* representation *of* ψ.

Remark 1. If the closure of D is contained by every U_j, then Definition 1 can simply be stated as follows. A continuous function ψ is a *piecewise* C^k function on the set $D \subseteq \mathbb{R}^n$ if there exists a finite set of functions $\psi_j \in C^k(U_j)$ for $j = 1, \ldots, q$ such that for any $u \in D$, $\psi(u) \in \{\psi_1(u), \ldots, \psi_q(u)\}$.

Note that for the Moreau–Yosida regularization of a piecewise smooth function to be smooth, the pieces ψ_j must be joined together properly. Mifflin, Qi, and Sun [13] introduced the following constraint qualification—AIPCQ. For any $u \in D$, we write

$$I(u) = \{i \in I : u \in D_i\}.$$

Definition 2. *The* AIPCQ *is said to hold for a piecewise smooth function* ψ *at* u *if for every subset* $K \subseteq I(u)$ *for which there exists a sequence* $\{u^k\}$ *with* $\{u^k\} \to u$, $K \subseteq I(u^k)$, *and the vectors*

$$\left\{ \begin{pmatrix} \nabla\psi_i(u^k) \\ 1 \end{pmatrix} : i \in K \right\} \tag{7}$$

being linearly independent, it follows that the vectors

$$\left\{ \begin{pmatrix} \nabla\psi_i(u) \\ 1 \end{pmatrix} : i \in K \right\} \tag{8}$$

are linearly independent.

Remark 2. The set $I(u)$ defined in this paper and the corresponding set in [13], denoted by $I'(u)$, are slightly different. In [13], they define

$$I'(u) = \{j \in I : \psi_j(u) = \psi(u)\}.$$

Since $u \in D_j$ implies $\psi_j(u) = \psi(u)$, we have $I(u) \subseteq I'(u)$. For $u \in U_j \setminus D_j$, $\psi_j(u)$ can be set to any value (as long as $\psi_j \in C^k(U_j)$); hence we can assume, without loss of generality, that $\psi_j(u) \neq \psi(u)$ for all $u \in U_j \setminus D_j$. Under this assumption,

$$I(u) = I'(u) \qquad \forall\, u \in D.$$

By virtue of the AIPCQ, Mifflin, Qi, and Sun [13] derived the following result, which will be used in the analysis of this paper.

Proposition 1. *Suppose that the convex function* φ *is piecewise* C^2 *on* $\mathbb{R}^m$ *and that the AIPCQ holds at the proximal point* $p(v)$ *for a given* $v \in \mathbb{R}^m$. *Then there exists an open neighborhood* $\mathcal{N}(v)$ *about* v *such that the gradient* g *of the function* η, *the Moreau–Yosida regularization of* φ, *is piecewise* C^1 *(smooth) on* $\mathcal{N}(v)$. *Hence* g *is semismooth at* v.

3. Piecewise smoothness of φ. In this section, we will study the piecewise smoothness of the Lagrangian-dual function φ for the case $f(x) = c^T x$ in (3), which is defined by

$$\varphi(v) = \sup\{-c^T x + v^T(Ax - a) \mid x \in \mathcal{F}\}. \tag{9}$$

The piecewise smoothness is an important characteristic of the Lagrangian-dual function φ. The investigation of this characteristic is helpful to optimization methods

which use the Lagrangian dual. Hence the results in this section are significant in their own right. In the next section, the piecewise smoothness of φ will then be used to prove the semismoothness of the gradient of the Moreau–Yosida regularization. Denote

$$\Omega := \{u = A^T v - c : v \in \mathbb{R}^m\}.$$

Clearly, Ω is an m-dimensional affine set in $\mathbb{R}^n$ since $\mathrm{rank}(A) = m$. We make the following assumptions throughout the paper.

Assumption 1. $c \notin \{A^T s : s \in \mathbb{R}^m\}$.

Assumption 2. $f_i \in C^2(\mathbb{R}^n)$ for all $i \in \hat{I}$.

Assumption 3. $\mathcal{F} \neq \emptyset$ and $\Omega \cap \mathcal{F}^b \neq \emptyset$.

Here, $\mathcal{F}^b$ denotes the *barrier cone* of the convex set $\mathcal{F}$ defined by

$$\mathcal{F}^b = \{y \in \mathbb{R}^n \mid \exists\, \beta \in \mathbb{R} \text{ such that } y^T x \leq \beta \ \ \forall\, x \in \mathcal{F}\}.$$

Remark 3. If $c = A^T s$ for some $s \in \mathbb{R}^m$, then $Ax = a$ implies $c^T x = s^T Ax = s^T a$. This means that any feasible solution of (1) is an optimal solution. Assumption 1 should rule out this degenerate case. Assumption 1 can also be written as $0 \notin \Omega$. Assumption 2 is a natural assumption of smoothness. The motivation of Assumption 3 is to guarantee the properness of the function φ, as shown by Lemma 1 below.

Define ζ, the *support function* of $\mathcal{F}$ in $\mathbb{R}^n$, as follows

$$\zeta(u) = \delta^*(u \mid \mathcal{F}) := \sup\{\langle u, x\rangle \mid x \in \mathcal{F}\}, \quad u \in \mathbb{R}^n. \tag{10}$$

Then the Lagrangian-dual function φ defined in (9) can be rewritten as

$$\varphi(v) = \zeta(A^T v - c) - a^T v. \tag{11}$$

We now define some notation which will be used in the paper.

(i) Q is said to be a *facet* of $\mathcal{F}$ if there exists an index subset $I_Q \subset \hat{I}$ such that $Q = \{x \in \mathcal{F} : f_i(x) = 0, \forall i \in I_Q\}$. I_Q is referred to as the *index set of the facet* Q.

(ii) For a convex function $h : \mathbb{R}^s \to \bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$, the *domain* of h, denoted by $\mathrm{dom}h$, is defined by $\mathrm{dom}h := \{z \in \mathbb{R}^s : h(z) < +\infty\}$.

Lemma 1. *The Lagrangian-dual function φ is a proper convex function on $\mathbb{R}^m$ if and only if Assumption* 3 *holds. One also has*

$$\mathrm{dom}\varphi = \{v \in \mathbb{R}^m \mid A^T v - c \in \mathcal{F}^b\}.$$

Proof. It is evident that $\varphi(v)$ can never be $-\infty$ if $\mathcal{F} \neq \emptyset$, and if $\mathcal{F} = \emptyset$, then $\varphi \equiv -\infty$.

By (11), we have

$$\mathrm{dom}\varphi = \{v \in \mathbb{R}^m \mid A^T v - c \in \mathrm{dom}\zeta\}.$$

Hence, $\mathrm{dom}\varphi \neq \emptyset$ if and only if $\Omega \cap \mathrm{dom}\zeta \neq \emptyset$. Since ζ is the support function of $\mathcal{F}$, it is easy to see that

$$\mathrm{dom}\zeta = \mathcal{F}^b.$$

Therefore, the second condition in Assumption 3 is a necessary and sufficient condition for $\mathrm{dom}\varphi \neq \emptyset$. ∎

PROPOSITION 2. *If ζ is piecewise C^2 on a set $D \subseteq \mathbb{R}^n \setminus \{0\}$, then φ is piecewise C^2 on the set $E := \{v : A^T v - c \in D\} \subseteq \mathbb{R}^m$ under Assumption* 1.

Proof. Since ζ is piecewise C^2, there exist closed sets D_i, open sets U_i, and functions $\zeta_i \in C^2(U_i)$, $i \in l$, where l is a finite index set, which satisfy Definition 1. Let

$$\varphi_i(v) := \zeta_i(A^T v - c) - a^T v, \quad E_i := \{v : A^T v - c \in D_i\}, \quad V_i := \{v : A^T v - c \in U_i\}.$$

Then it is evident that $\varphi_i \in C^2(V_i)$, E_i is closed, and V_i is open. Furthermore, $\varphi_i (i \in l)$ satisfy (i)–(iii) in Definition 1. Hence φ is a piecewise C^2 function. ∎

PROPOSITION 3. *Suppose that f_i is an affine function on $\mathbb{R}^n$ for every $i \in \hat{I}$. Then the function φ defined in* (9) *is a piecewise C^2 function on its domain. Especially, φ is piecewise affine on its domain.*

Proof. By Proposition 2, it suffices to show that ζ is piecewise C^2 on $\Omega \cap \mathrm{dom}\zeta$. According to the remark after Definition 1, it suffices to show that there exist twice continuously differentiable functions ζ_j on $\mathbb{R}^n (= U_j)$, $j \in \hat{J}$ a finite index set, such that for any $u \in \Omega \cap \mathrm{dom}\zeta$

$$\zeta(u) \in \{\zeta_j(u) : j \in \hat{J}\}. \tag{12}$$

It is known that the polyhedral $\mathcal{F}$ can be represented by its vertices $\{x_1, \ldots, x_p\}$ and extreme rays $\{r_1, \ldots, r_q\}$ in the form

$$\mathcal{F} = \left\{ x = \sum_{i=1}^{p} \alpha_i x_i + \sum_{i=1}^{q} \lambda_i r_i \ : \ \alpha_i \geq 0, \sum_{i=1}^{p} \alpha_i = 1, \lambda_i \geq 0 \right\}.$$

Define

$$\bar{\mathcal{F}} = \left\{ x = \sum_{i=1}^{p} \alpha_i x_i + \sum_{i=1}^{q} \lambda_i r_i \ : \ \alpha_i \geq 0, \sum_{i=1}^{p} \alpha_i = 1, \, 0 \leq \lambda_i \leq 1 \right\}.$$

We claim that, for any $u \in \mathrm{dom}\zeta$, $\sup\{u^T x : x \in \mathcal{F}\} = \sup\{u^T x : x \in \bar{\mathcal{F}}\}$. Assume by contradiction that there exist a $u \in \mathrm{dom}\zeta$ and a $\bar{x} \in \mathcal{F} \setminus \bar{\mathcal{F}}$ such that $u^T \bar{x} > \sup\{u^T x : x \in \bar{\mathcal{F}}\}$. Denote $J := \{i : \lambda_i > 1\}$, where the λ_i's are the coefficients in the representation of $\bar{x}$. Let $\hat{x} \in \bar{\mathcal{F}}$ be defined by the same representation of $\bar{x}$ except for changing the λ_i, $i \in J$, to 1. Then $\bar{x} - \hat{x} = \sum_{i \in J} (\lambda_i - 1) r_i$. Since $\hat{x} \in \bar{\mathcal{F}}$, we have $u^T \hat{x} < u^T \bar{x}$, i.e.,

$$\sum_{i \in J} (\lambda_i - 1) u^T r_i > 0.$$

Thus there exists at least an $\bar{i} \in J$ with $u^T r_{\bar{i}} > 0$. For any fixed $x_0 \in \mathcal{F}$ and any $\lambda \geq 0$, $x_0 + \lambda r_{\bar{i}} \in \mathcal{F}$. Thus $\zeta(u) \geq u^T x_0 + \lambda u^T r_{\bar{i}} \to +\infty$ as $\lambda \to +\infty$, which contradicts the fact $u \in \mathrm{dom}\zeta$. This shows that for any $u \in \mathrm{dom}\zeta$

$$\zeta(u) = \sup\{u^T x \mid x \in \bar{\mathcal{F}}\}.$$

Note that $\bar{\mathcal{F}}$ is a bounded polytope. Without loss of generality, let $\{\bar{x}_1, \ldots, \bar{x}_k\}$ be all vertices of $\bar{\mathcal{F}}$, and define $\zeta_j(u) = \bar{x}_j^T u$. Then $\zeta_j \in C^2(\mathbb{R}^n)$ (here $U_j = \mathbb{R}^n$). For any $u \in \Omega \cap \mathrm{dom}\zeta$, because $u \neq 0$ by Assumption 1, the set of maximizers of $\zeta(u)$ must contain at least a vertex, say $\bar{x}_j$, of $\bar{\mathcal{F}}$. It follows that

$$\zeta(u) = \bar{x}_j^T u = \zeta_j(u),$$

which shows (12). Thus, $\zeta(u)$ is piecewise C^2 on its domain. Evidently, $\zeta(u)$ is also piecewise affine on its domain, and so is φ. □

Next, we consider the case where all $\nabla^2 f_i$ ($i \in \hat{I}$) are positive definite. Our analysis will proceed as follows. For each facet Q (of any dimension) of $\mathcal{F}$, we will define an open set U and a C^2 function on U. Roughly speaking, we first define a mapping from x-space to an open set in u-space (actually the mapping is defined on enlarged spaces), prove that this mapping is bijective, and then use the inverse of this mapping to define a function on the open set in u-space. For any facet Q of $\mathcal{F}$ with the index set I_Q, we define

$$W := \left\{ (x,\lambda) \in \mathbb{R}^n \times \mathbb{R}^{|I_Q|} : f_i(x) = 0,\ i \in I_Q,\ \sum_{i\in I_Q} \lambda_i \nabla^2 f_i(x) \succ 0 \right\}, \tag{13}$$

where $B \succ 0$ means that the matrix B is symmetric positive definite,

$$U := \left\{ u = \sum_{i\in I_Q} \lambda_i \nabla f_i(x) \in \mathbb{R}^n : (x,\lambda) \in W \right\}. \tag{14}$$

Note that for $(x,\lambda) \in W$, x is not required to be in Q. Actually, x need not be in $\mathcal{F}$. Without loss of generality, let $I_Q = \{1,\dots,k\}$. Denote $\tilde{f} = (f_1,\dots,f_k)^T$, and define a mapping $\Gamma : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^{n+k}$ by

$$\Gamma(x,\lambda) := \begin{pmatrix} \sum_{i=1}^k \lambda_i \nabla f_i(x) \\ \tilde{f}(x) \end{pmatrix}. \tag{15}$$

Note that the Karush–Kuhn–Tucker (KKT) conditions for problem (10) can be written as

$$\Gamma(x,\lambda) = (u;0).$$

The following lemma plays a fundamental role in our analysis.

LEMMA 2. *Let W, U be defined by* (13), (14), *respectively. Suppose that for any $x \in \mathbb{R}^n$ all $\nabla^2 f_i(x)$ ($i \in I_Q$) are positive definite and $\{\nabla f_i(x)\}_{i\in I_Q}$ are linearly independent. Then* (i) *U is an open set in $\mathbb{R}^n$; and* (ii) *there exists a continuously differentiable bijective mapping $\xi = (\xi_x, \xi_\lambda) : U \to W$ such that for all $u \in U$, $\Gamma(\xi_x(u), \xi_\lambda(u)) = (u;0)$; i.e., ξ is the inverse mapping of Γ restricted on U.*

Proof. (i) For any $(\bar{x},\bar{\lambda}) \in W$, let $(\bar{u};\bar{v}) = \Gamma(\bar{x},\bar{\lambda})$. Then $\bar{u} = \sum_{j=1}^k \bar{\lambda}_j \nabla f_j(\bar{x})$ and $\bar{v} = 0$. In the following, we seek to show that $\bar{u}$ is an interior point of U. Let us denote $\nabla\tilde{f} := (\nabla f_1,\dots,\nabla f_k) \in \mathbb{R}^{n\times k}$. Then, $\nabla\tilde{f}(x)$ has full column rank, i.e., $\operatorname{rank}(\nabla\tilde{f}(x)) = k$, by assumption. By the continuity of $\nabla^2 f_i$, $i \in \hat{I}$, there exists a neighborhood of $(\bar{x},\bar{\lambda})$, denoted by $\mathcal{N}_x$, such that $\sum_{i=1}^k \lambda_i \nabla^2 f_i(x) \succ 0$ for all $(x,\lambda) \in \mathcal{N}_x$. Thus, the Jacobian of Γ,

$$\nabla\Gamma(x,\lambda) = \begin{pmatrix} \sum_{i=1}^k \lambda_i \nabla^2 f_i(x) & \nabla\tilde{f}(x) \\ \nabla\tilde{f}(x)^T & 0 \end{pmatrix},$$

is nonsingular on $\mathcal{N}_x$. By the inverse function theorem, there exists a neighborhood of $(\bar{u},\bar{v})$, denoted by $\mathcal{N}_u$, such that there exists an inverse mapping Ψ of Γ defined on

$\mathcal{N}_u$, and for any $(u,v) \in \mathcal{N}_u$, $\Psi(u,v) \in \mathcal{N}_x$ and $\Gamma(\Psi(u,v)) = (u;v)$. In particular, for any $(u,v) \in \mathcal{N}_u$ with $v = \bar{v} = 0$, $(x,\lambda) = \Psi(u,0)$ satisfies

$$u = \sum_{i=1}^{k} \lambda_i \nabla f_i(x), \qquad \tilde{f}(x) = 0, \qquad \sum_{i=1}^{k} \lambda_i \nabla^2 f_i(x) \succ 0. \tag{16}$$

This implies $(x,\lambda) \in W$ and thus $u \in U$ for all $(u,0) \in \mathcal{N}_u$. Since $U_0 := \{u : (u,0) \in \mathcal{N}_u\}$ is an open set in $\mathbb{R}^n$, and $\bar{u} \in U_0 \subset U$, $\bar{u}$ is an interior point of U. Thus U is open.

(ii) Since the Jacobian $\nabla\Gamma(x,\lambda)$ is nonsingular and continuous on the entire set W and since Γ maps W onto $U \times \{0\}$, the inverse mapping Ψ of Γ defined in (i) is a continuously differentiable bijective mapping from $U \times \{0\}$ onto W. Define a mapping $\xi : U \to W$ by $\xi(u) = \Psi(u,0)$. Then ξ is continuously differentiable and bijective, and $\Gamma(\xi(u)) = (u;0)$. □

As a consequence of Lemma 2, we obtain the following result.

LEMMA 3. *Let $\zeta_Q(u) = u^T \xi_x(u)$, where ξ_x is defined in Lemma 2. Then $\zeta_Q \in C^2(U)$, and for any $u \in U$*

$$\nabla\zeta_Q(u) = \xi_x(u). \tag{17}$$

Proof. From Lemma 2 and the first equation of $\Gamma(\xi(u)) = (u;0)$, it follows that

$$u = \sum_{i \in I_Q} \xi_{\lambda_i}(u) \nabla f_i(\xi_x(u)).$$

Thus,

$$\begin{aligned} \nabla\zeta_Q(u) &= \xi_x(u) + \nabla\xi_x(u)u \\ &= \xi_x(u) + \nabla\xi_x(u) \textstyle\sum_{i \in I_Q} \xi_{\lambda_i}(u) \nabla f_i(\xi_x(u)) \\ &= \xi_x(u) + \textstyle\sum_{i \in I_Q} \xi_{\lambda_i}(u) \nabla\xi_x(u) \nabla f_i(\xi_x(u)). \end{aligned}$$

According to the second equation in $\Gamma(\xi(u)) = (u;0)$, we have $f_i(\xi_x(u)) = 0$ for all $u \in U$ and $i \in I_Q$. Differentiating these functions, we obtain $\nabla\xi_x(u)\nabla f_i(\xi_x(u)) = 0$ for all $i \in I_Q$. Hence,

$$\sum_{i \in I_Q} \xi_{\lambda_i}(u) \nabla\xi_x(u) \nabla f_i(\xi_x(u)) = 0.$$

Thus, it follows that

$$\nabla\zeta_Q(u) = \xi_x(u).$$

By Lemma 2, $\xi_x(u)$ is continuously differentiable on U. Therefore, ζ_Q is twice continuously differentiable on U. □

The following proposition is one of the main results in this paper, showing the piecewise smoothness of the function φ.

PROPOSITION 4. *For φ defined by (9), suppose that, for all $i \in \hat{I}$, $\nabla^2 f_i(x)$ are positive definite, and for any facet Q of $\mathcal{F}$ with the index set I_Q, $\{\nabla f_i(x)\}_{i \in I_Q}$ are linearly independent. Then φ is piecewise C^2 on its domain.*

Proof. Let us first consider the function ζ defined by (10) on the set D, where $D = \Omega \cap \mathrm{dom}\zeta$. Let $\{Q_1, \ldots, Q_q\}$ be the set of all facets of $\mathcal{F}$. Let W_i, U_i, and ξ_i be

defined in (13), (14), and Lemma 2 for the facet $Q_i = \{x \in \mathcal{F} : f_l(x) = 0,\ l \in I_i\}$. Define

$$D_i := \Omega \cap \left\{ u = \sum_{l \in I_i} \lambda_l \nabla f_l(x) : x \in Q_i,\ \lambda_l \geq 0 \right\}, \tag{18}$$

which is evidently a closed set. By Lemma 2, U_i is open. Define $\zeta_i(u) := u^T \xi_{ix}(u)$. In what follows, we show that (i), (ii), and (iii) in Definition 1 hold.

(i) For any $u \in \mathrm{ri}D$, by [18, Theorem 23.4 and Corollary 23.5.3], there exists an optimal solution x^* to problem (10), which together with a Lagrangian multiplier $\bar{\lambda}^*$ satisfies the KKT conditions:

$$\begin{aligned} & u = \textstyle\sum_{i=1}^{\theta} \bar{\lambda}_i^* \nabla f_i(x^*), \\ & \bar{\lambda}_i^* \geq 0, \\ & f_i(x^*) \leq 0, \\ & \bar{\lambda}_i^* f_i(x^*) = 0, \quad i = 1, 2, \ldots, \theta. \end{aligned} \tag{19}$$

Because $u \neq 0$ by Assumption 1, x^* must lie on some facets of $\mathcal{F}$. Let $Q_j = \{x \in \mathcal{F} : f_i(x) = 0,\ i \in I_j\}$ be the *smallest* facet at x^*. By "smallest" we mean that for any $i \notin I_j$, $f_i(x^*) \neq 0$. Then $\bar{\lambda}_i^* = 0$ for all $i \notin I_j$. Let λ^* denote the subvector of $\bar{\lambda}^*$ consisting of components in I_j. Then $u = \sum_{i \in I_j} \lambda_i^* \nabla f_i(x^*)$, which together with $\lambda^* \geq 0$ and $x^* \in Q_j$ implies $u \in D_j$. This shows that $\mathrm{ri}D \subseteq \cup_{i=1}^q D_i$. Thereby, $D \subseteq \cup_{i=1}^q D_i$ since each D_i is closed.

For any $u \in D_j$, let $\bar{x} \in Q_j$ and $\bar{\lambda} \geq 0$ represent u as in (18). $\bar{\lambda} \neq 0$, since $u \neq 0$ by Assumption 1. This implies that $\sum_{i \in I_j} \bar{\lambda}_i \nabla^2 f_i(\bar{x}) \succ 0$, since all $\nabla^2 f_i$ are positive definite. Thus $(\bar{x}, \bar{\lambda}) \in W_j$ and $u \in U_j$. This shows $D_j \subseteq U_j$.

(ii) By Lemma 3, $\zeta_i \in C^2(U_i)$ for $i = 1, \ldots, q$.

(iii) For any $u \in D \cap D_j$, let $\bar{x} \in Q_j$ and $\bar{\lambda} \geq 0$ represent u as in (18). Let $\bar{\lambda}^*$ be defined by $\bar{\lambda}_i^* = \bar{\lambda}_i$ for $i \in I_j$ and $\bar{\lambda}_i^* = 0$ for $i \notin I_j$. Then $(\bar{x}, \bar{\lambda}^*)$ satisfies the KKT conditions (19). Thus $\zeta(u) = u^T \bar{x}$. On the other hand, the second part of (i) shows that $(\bar{x}, \bar{\lambda}) \in W_j$. Using the relation in (18), we have $(u; 0) = \Gamma_j(\bar{x}, \bar{\lambda})$, where Γ_j is the mapping defined in (15). Since ξ_j is the inverse of Γ_j restricted on U_j, $\xi_j(u) = (\bar{x}, \bar{\lambda})$. By definition, we have

$$\zeta_j(u) = u^T \xi_{jx}(u) = u^T \bar{x}.$$

Thus $\zeta(u) = \zeta_j(u)$ for any $u \in D \cap D_j$.

The above shows that ζ is piecewise C^2 on $D(= \Omega \cap \mathrm{dom}\zeta)$. By virtue of Proposition 2, φ is piecewise C^2 on its domain. □

Remark 4. In Propositions 3 and 4 we conclude that φ is piecewise C^2 convex under the assumption that the constraints for $\mathcal{F}$ are either all linear or all have positive definite Hessian matrices. A natural question arises: Can φ be piecewise C^2 for more general $\mathcal{F}$? The following example considers an $\mathcal{F}$ which is defined by a linear constraint and a strictly convex constraint with a positive definite Hessian, and gives a negative answer to the above question.

Example 1. Let

$$\varphi(v) = \sup\{-c^T x + v^T(Ax - a) \mid x \in \mathcal{F}\},$$

where

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \qquad c = (0, 0, -1)^T, \qquad a = (0, 0)^T,$$

and

$$\mathcal{F} = \{x \in \mathbb{R}^3 : x_1^2 + x_2^2 + x_3^2 - 1 \le 0,\ x_3 \le 0\}.$$

Since $\mathcal{F}$ is bounded, $\mathcal{F}^b = \mathbb{R}^3$ and $\mathrm{dom}\varphi = \mathbb{R}^2$. One can verify that Assumptions 1, 2, and 3 are satisfied. It is easy to see that

$$\varphi(v) = \sup\{(v_1, v_2, 1)^T x \mid x \in \mathcal{F}\},$$

and that the maximizer is $x = (v_1/\|v\|, v_2/\|v\|, 0)^T$ if $v \neq 0$ and is any point on $\mathcal{F} \cap \{(x_1, x_2, x_3) \mid x_3 = 0\}$ if $v = 0$. It follows that

$$\varphi(v) = \sqrt{v_1^2 + v_2^2}.$$

Obviously, φ is smooth at any point $v \neq 0$. So for any nonzero $v \in \mathbb{R}^2$, the gradient and the Hessian of φ can be written as

$$\nabla\varphi(v) = \begin{pmatrix} v_1/\sqrt{v_1^2 + v_2^2} \\ v_2/\sqrt{v_1^2 + v_2^2} \end{pmatrix},$$

$$\nabla^2\varphi(v) = \begin{pmatrix} v_2^2/(v_1^2 + v_2^2)^{3/2} & -v_1v_2/(v_1^2 + v_2^2)^{3/2} \\ -v_1v_2/(v_1^2 + v_2^2)^{3/2} & v_1^2/(v_1^2 + v_2^2)^{3/2} \end{pmatrix}.$$

It is evident that $\nabla^2\varphi(v)$ is unbounded as $v \to 0$ ($v \neq 0$), (either $\frac{\partial^2\varphi(v)}{\partial v_1^2} \to \infty$ if $|v_1| \le |v_2|$, or $\frac{\partial^2\varphi(v)}{\partial v_2^2} \to \infty$ if $|v_1| \ge |v_2|$).

To show φ is not piecewise C^2 on its domain $\mathbb{R}^2$, let (E_j, V_j, φ_j) be any piece representing φ in a neighborhood of $v = 0$, namely, $0 \in E_j \subset V_j$ and φ_j is a function on V_j satisfying $\varphi_j(v) = \varphi(v)$ for all $v \in E_j$. Since $\nabla^2\varphi_j(v) = \nabla^2\varphi(v)$ wherever φ is twice differentiable, we have $\nabla^2\varphi_j(v) \to \infty$ as $v \to 0$ ($E_j \ni v \neq 0$). Since the origin is an interior point of V_j, $\varphi_j \notin C^2(V_j)$. Therefore, φ is not a piecewise C^2 function on its domain.

4. Semismoothness of the gradient of η and its conjugate. In this section, we will study the semismoothness of the gradient of the Moreau–Yosida regularization of φ as discussed in section 3, where $f(x) = c^T x$. We will also investigate the properties of the conjugate of η and explore its relations with the original problem.

4.1. Semismoothness of the gradient of η. Our study on the semismoothness of g is based on the theory established by Mifflin, Qi, and Sun [13]. In their paper, they assume that φ is piecewise C^2 convex on the whole space $\mathbb{R}^m$, i.e., $\mathrm{dom}\varphi = \mathbb{R}^m$. We follow this assumption in this section. By Lemma 1, $\mathrm{dom}\varphi = \mathbb{R}^m$ if and only if $\Omega \subset \mathcal{F}^b$. Therefore, we make the following assumption in this section to replace Assumption 3.

Assumption 4. $\mathcal{F} \neq \emptyset$ and $\Omega \subset \mathcal{F}^b$.

To show the semismoothness of g, we shall first show that φ defined in (3) satisfies AIPCQ in the following two cases: (i) all f_i are affine, and (ii) all f_i possess positive

definite Hessian matrices. Suppose that $\varphi(v)$ is a piecewise smooth function with the representation $\{(E_i, V_i, \varphi_i)\}_{i\in I}$. For any $v \in \mathbb{R}^m$, define

$$I(v) := \{i \in I : v \in E_i\}.$$

LEMMA 4. *Suppose for every $i \in \hat{I}$ that f_i is an affine function on $\mathbb{R}^n$. Then for the piecewise affine function $\varphi(v)$ defined by* (9)*, the AIPCQ holds at every v on $\mathbb{R}^m$.*

Proof. Suppose that φ is represented by $\{\varphi_i\}_{i\in I}$, where $\varphi_i(v) = \beta_i^T v - \alpha_i$. For any $w \in \mathbb{R}^m$ and any index set $K \subseteq I(w)$,

$$\left\{ \begin{pmatrix} \nabla\varphi_i(w) \\ 1 \end{pmatrix} : i \in K \right\} = \left\{ \begin{pmatrix} \beta_i \\ 1 \end{pmatrix} : i \in K \right\}$$

is a set of constant vectors. Therefore the AIPCQ holds at any $v \in \mathbb{R}^m$. ∎

Now we consider the case that the set $\mathcal{F}$ is defined by all convex functions f_j with positive definite Hessian matrices. In the proof of Proposition 4, we have defined a representation $\{(D_j, U_j, \zeta_j)\}_{j\in I}$ of ζ. This representation induces a representation $\{(E_j, V_j, \varphi_j)\}_{j\in I}$ of φ as defined in the proof of Proposition 2; we will use these notations below.

Because the value of $\varphi_j(v)$, $v \in V_j \setminus E_j$, does not affect the representation of φ, it therefore can be set to any value. For simplicity, in what follows, we assume that

$$\varphi_j(v) \neq \varphi(v) \qquad \forall j \in I,\ v \in V_j \setminus E_j \tag{20}$$

(see also Remark 2).

LEMMA 5. *Suppose that the conditions of Proposition* 4 *are satisfied. Let $\{(E_j, V_j, \varphi_j)\}_{j\in I}$ be a representation of φ. Then, for any $v \in \mathrm{dom}\varphi$ and any $i, j \in I$, $\nabla\varphi_i(v) = \nabla\varphi_j(v)$ if $\varphi_i(v) = \varphi_j(v) = \varphi(v)$.*

Proof. It suffices to show that for any $u \in \Omega \cap \mathrm{dom}\zeta$ and any $i, j \in I$, $\nabla\zeta_i(u) = \nabla\zeta_j(u)$ if $\zeta_i(u) = \zeta_j(u) = \zeta(u)$, where $\{(D_j, U_j, \zeta_j)\}_{j\in I}$ is the corresponding representation of ζ. Let $u \in U_i$ and $\xi_i : U_i \to W_i$ be defined as in Lemma 2. If we can show that $\zeta_i(u) = \zeta(u)$ implies that $\xi_{ix}(u)$ is indeed the unique maximizer x^* of problem (9) for the given u, then the fact $\nabla\zeta_i(u) = \xi_{ix}(u)$ (Lemma 3) leads readily to $\nabla\zeta_i(u) = x^* = \nabla\zeta_j(u)$, provided that $\zeta_i(u) = \zeta_j(u) = \zeta(u)$.

Now, if $\zeta_i(u) = \zeta(u)$, then (20) (applying to ζ) and $u \in U_i$ imply $u \in D_i$. By definition of (18), $x = \xi_{ix}(u) \in Q_i \subset \mathcal{F}$; i.e., $\xi_{ix}(u)$ is a feasible solution of problem (9). So, $u^T\xi_{ix}(u) = \zeta_i(u) = \zeta(u)$ implies that $\xi_{ix}(u)$ is a unique optimal solution x^* of problem (9) for a given u. ∎

The above lemma actually holds true for φ with $\mathrm{dom}\varphi \neq \mathbb{R}^m$. This lemma will be used to prove Lemma 6. In addition to it, we obtain a property of the function φ as a by-product, namely, φ is indeed differentiable on the relative interior of $\mathrm{dom}\varphi$, because the subdifferential $\partial\varphi(v)$ at any point $v \in \mathrm{ri}(\mathrm{dom}\varphi)$ is a singleton.

LEMMA 6. *Suppose that the conditions of Proposition* 4 *are satisfied. Then, for the piecewise C^2 function φ, the AIPCQ holds at each $v \in \mathbb{R}^m$.*

Proof. Let $v \in \mathbb{R}^m$ and $K \subseteq I(v)$. If $|K| = 1$, the vectors in the set in (8) are evidently linearly independent (actually, the set is a singleton). So the conditions for AIPCQ are satisfied. If $|K| \geq 2$, then for any $w \neq v$ with $K \subseteq I(w)$ and for any $i \neq j \in K$, $\varphi_i(w) = \varphi_j(w) = \varphi(w)$ implies that $\nabla\varphi_i(w) = \nabla\varphi_j(w)$ by Lemma 5. Thus the set of vectors in (7) can never be linearly independent. This means that the conditions for AIPCQ hold automatically. ∎

From the piecewise C^2 smoothness of φ shown in section 3 and the qualification AIPCQ verified in this section, we have the semismoothness of g, as stated below.

PROPOSITION 5. *Let φ be defined by* (9). *Suppose that Assumptions* 1, 2, *and* 4 *are satisfied. Suppose that f_i, $i \in \hat{I}$, are either all affine or all possess positive definite Hessian matrices. In the latter case suppose that for any facet Q of $\mathcal{F}$ with the index set I_Q, $\{\nabla f_i(x)\}_{i\in I_Q}$ are linearly independent. Then the gradient $g(v)(= \nabla\eta(v))$ of the Moreau–Yosida regularization η is piecewise smooth, and thereby semismooth, on $\mathbb{R}^m$.*

Proof. The proof follows directly from Propositions 1, 3, and 4, and Lemmas 4 and 6. □

Remark 5. The above proposition shows that g is semismooth if constraints defining $\mathcal{F}$ either are all linear or all possess positive definite Hessian matrices. In Example 1 of section 3, we found that, for some simple mixed constraints, the Lagrangian-dual function φ is not piecewise C^2. Actually, the second-order derivatives of φ tend to infinity at some point. Since the semismoothness of g is closely related to the piecewise C^2 smoothness of φ, we might expect that for this example g is not semismooth either. However, the gradient g of the Moreau–Yosida regularization of this function φ is semismooth, as shown below.

Example 2 (Example 1 (continued)). It is known that $\varphi(v) = \sqrt{v_1^2 + v_2^2}$. For convenience in description we set $M = I$, so we have

$$\eta(v) = \min\left\{\sqrt{w_1^2 + w_2^2} + \frac{1}{2}\|w - v\|^2 \mid w \in \mathbb{R}^2\right\}.$$

It is easy to verify that, for $\|v\| \le 1$,

$$\eta(v) = (v_1^2 + v_2^2)/2, \qquad p(v) = (0,\ 0)^T,$$

and for $\|v\| \ge 1$,

$$\eta(v) = \sqrt{v_1^2 + v_2^2} - 1/2, \qquad p(v) = (1 - 1/\|v\|)v.$$

Let $\hat{V}_1 = \{v \in \mathbb{R}^2 : \|v\| \le 1\}$ and $\hat{V}_2 = \{v \in \mathbb{R}^2 : \|v\| \ge 1\}$. By (6), it suffices to study the semismoothness of p. For $v \in \mathrm{int}\hat{V}_1$, the Jacobian of p is

$$J(p(v)) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \tag{21}$$

and for $v \in \mathrm{int}\hat{V}_2$,

$$J(p(v)) = \begin{pmatrix} 1 - v_2^2/\|v\|^3 & v_1v_2/\|v\|^3 \\ v_1v_2/\|v\|^3 & 1 - v_1^2/\|v\|^3 \end{pmatrix}. \tag{22}$$

From the Jacobian of p above, we can see that p is smooth on the interior of $\hat{V}_i$ $(i = 1, 2)$. Thus we need only to investigate the semismoothness of p on the region where the two sets meet, namely, $\{v \in \mathbb{R}^2 : \|v\| = 1\}$. Let $\bar{v} = (\bar{v}_1, \bar{v}_2)^T$ be any point on this region; we will show that p is semismooth at $\bar{v}$. By the definition of semismoothness [16], it suffices to show that

$$\lim_{h'\to h, t\to 0+} \{Vh' : V \in \partial p(\bar{v} + th')\} \tag{23}$$

exists for any $h \in \mathbb{R}^2$. Let $S_1(\bar{v}) = \{h \in \mathbb{R}^2 : h^T\bar{v} < 0\}$, $S_2(\bar{v}) = \{h \in \mathbb{R}^2 : h^T\bar{v} > 0\}$, $S_3(\bar{v}) = \{h \in \mathbb{R}^2 : h^T\bar{v} = 0\}$. Write $v' = \bar{v}+th'$. Then $\|v'\|^2 = \|\bar{v}\|^2+2t\bar{v}^Th'+t^2\|h'\|^2$.

If $h \in S_1(\bar{v})$ or $h \in S_2(\bar{v})$, then for any sufficiently small $t > 0$ and h' close to h, $v' \in \mathrm{int}\hat{V}_1$ or $v' \in \mathrm{int}\hat{V}_2$. It is evident that the limit in (23) exists.

If $h \in S_3(\bar{v})$, then for any sufficiently small $t > 0$ and h' close to h there are the following three cases: if $\|v'\| < 1$, we have

$$\lim_{h'\to h,t\to 0+} Vh' = \lim_{h'\to h,t\to 0+} \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} h'_1 \\ h'_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{24}$$

If $\|v'\| > 1$, we have

$$\begin{aligned} \lim_{h'\to h,t\to 0+} Vh' &= \lim_{h'\to h,t\to 0+} \begin{pmatrix} 1 - {v'_2}^2/\|v'\|^3 & v'_1v'_2/\|v'\|^3 \\ v'_1v'_2/\|v'\|^3 & 1 - {v'_1}^2/\|v'\|^3 \end{pmatrix} \begin{pmatrix} h'_1 \\ h'_2 \end{pmatrix} \\ &= \lim_{h'\to h,t\to 0+} \begin{pmatrix} (1 - {v'_2}^2/\|v'\|^3)h'_1 + (v'_1v'_2/\|v'\|^3)h'_2 \\ (v'_1v'_2/\|v'\|^3)h'_1 + (1 - {v'_1}^2/\|v'\|^3)h'_2 \end{pmatrix}. \end{aligned} \tag{25}$$

Since

$$\begin{aligned} &\lim_{h'\to h,t\to 0+}[(1 - {v'_2}^2/\|v'\|^3)h'_1 + (v'_1v'_2/\|v'\|^3)h'_2] \\ &= (1 - \bar{v}_2^2/\|\bar{v}\|^3)h_1 + (\bar{v}_1\bar{v}_2/\|\bar{v}\|^3)h_2 = (1 - \bar{v}_2^2)h_1 + \bar{v}_1\bar{v}_2h_2 \\ &= \bar{v}_1^2h_1 + \bar{v}_1\bar{v}_2h_2 = \bar{v}_1(h_1\bar{v}_1 + h_2\bar{v}_2) = 0, \end{aligned}$$

and similarly, $\lim_{h'\to h,t\to 0+}[(v'_1v'_2/\|v'\|^3)h'_1 + (1 - {v'_1}^2/\|v'\|^3)h'_2] = 0$, by (24), we have

$$\lim_{h'\to h,t\to 0+} Vh' = (0, 0)^T. \tag{26}$$

Hence, Vh' tends to the same limit in these two cases by (24) and (26).

If $\|v'\| = 1$, by the definition of the generalized Jacobian, V is a convex combination of the Jacobians in (21) and (22) (with v replaced by v'). Thus, Vh' tends to the same limit, namely 0, as the above two cases.

Thereby, the limit in (23) exists if $h \in S_3(\bar{v})$. The above shows that p is semismooth on $\mathbb{R}^2$. Therefore, g is semismooth on $\mathbb{R}^2$ as well.

4.2. Conjugate of the Moreau–Yosida regularization. In this subsection, we investigate the relationship between the original problem with the linear objective and the Fenchel conjugate of Moreau–Yosida regularization of its Lagrangian-dual function.

First, recall the notion of Fenchel conjugate. Let ϕ be a real-valued convex function on $\mathbb{R}^l$. The *Fenchel conjugate*, denoted by ϕ^*, of ϕ is defined by (see [18])

$$\phi^*(x) := \sup\{\langle x^*, x\rangle - \phi(x) \mid x^* \in \mathbb{R}^l\} \quad \forall\, x \in \mathbb{R}^l.$$

Note that η, the Moreau–Yosida regularization of φ defined in (9), can be rewritten as

$$\eta(v) = (\pi_1 \square \pi_2)(v) := \inf\{\pi_1(v - w) + \pi_2(w) : w \in \mathbb{R}^m\}, \quad v \in \mathbb{R}^m, \tag{27}$$

where "$\square$" denotes the *infimal convolution* operation [18], $\pi_1(v) := \frac{1}{2}\|v\|_M^2, \pi_2(v) := \varphi(v)$, as defined in (9). Evidently, both π_1 and π_2 are proper convex functions; then by [18, Theorem 5.4], η is a convex function.

Using the conjugate operator, it is not hard to derive that

$$\pi_1^*(v) = \tfrac{1}{2}\|v\|^2_{M^{-1}} \quad \forall\, v \in \mathbb{R}^m.$$

Hence, we have $\mathrm{dom}\pi_1^* = \mathbb{R}^m$. Thereby, it follows from [8, Corollary 2.1.3] that

$$\eta^*(v) = \pi_1^*(v) + \pi_2^*(v) \quad \forall\, v \in \mathbb{R}^m.$$

Next, we study the conjugate of π_2. To ease notation, we define a mapping $\mathcal{A} : \mathbb{R}^m \to \mathbb{R}^n$ by

$$\mathcal{A}(v) = A^T v - c.$$

Then we have

$$\delta^*(A^T v - c \mid \mathcal{F}) = \zeta \circ \mathcal{A}(v), \quad v \in \mathbb{R}^m,$$

where ζ is defined in (10). Since $\mathrm{dom}\zeta = \mathcal{F}^b$, so $\zeta \circ \mathcal{A}$ is a closed convex function on $\mathbb{R}^m$ under Assumption 3. Thus, by [18, Theorem 16.3], it follows that

$$(\zeta \circ \mathcal{A})^*(v) = \mathrm{cl} \inf_{x \in \mathbb{R}^n} \{\zeta^*(x) - \langle -c, x\rangle \mid Ax = v\}.$$

Since $\mathcal{F}$ is closed, we then have

$$\begin{aligned}(\zeta \circ \mathcal{A})^*(v) &= \mathrm{cl} \inf_{x \in \mathbb{R}^n} \{(\delta^*(x \mid \mathcal{F}))^* + \langle c, x\rangle \mid Ax = v\} \\ &= \mathrm{cl} \inf_{x \in \mathbb{R}^n} \{\delta(x \mid \mathcal{F}) + \langle c, x\rangle \mid Ax = v\} \\ &= \mathrm{cl}\inf\{\langle c, x\rangle \mid Ax = v, x \in \mathcal{F}\}.\end{aligned}$$

On the other hand, by definition of conjugate, we have

$$\begin{aligned}\pi_2^*(v) &= \sup\{\langle v + a, v'\rangle - \sup\{\langle A^T v' - c, x\rangle \mid x \in \mathcal{F}\} \mid v' \in \mathrm{dom}\pi_2\} \\ &= \sup\{\langle v + a, v' - \zeta \circ \mathcal{A}(v')\rangle \mid v' \in \mathrm{dom}\pi_2\} \\ &= \sup\{\langle v + a, v' - \zeta \circ \mathcal{A}(v')\rangle \mid v' \in \mathrm{dom}(\zeta \circ \mathcal{A})\} \\ &= (\zeta \circ \mathcal{A})^*(v + a).\end{aligned}$$

Thus, we obtain the conjugate of π_2 as follows:

$$\pi_2^*(v) = \mathrm{cl}\inf\{\langle c, x\rangle \mid Ax = v + a, x \in \mathcal{F}\}.$$

We now derive an interesting result on the conjugate of Moreau–Yosida regularization of the Lagrangian-dual function as follows.

PROPOSITION 6. *Assume $\Omega \cap \mathcal{F}^b \neq \emptyset$. Then, for any $v \in \mathbb{R}^m$,*

$$\begin{aligned}&\mathrm{cl}\inf\{\langle c, x\rangle \mid Ax - a = v,\ x \in \mathcal{F}\} \\ &= \eta^*(v) - \tfrac{1}{2}\|v\|^2_{M^{-1}}.\end{aligned}$$

From Proposition 6, we can see that the optimal value function of the underlying parametric optimization problem can be represented by the conjugate function of

the regularized dual function of the (unperturbed) original problem, together with a quadratic function in terms of the perturbation parameter v. Note that the expression is taken under the *closure* and *infimal* operations on the set of objective values due to the fact that the minimum of the set of objective values of the corresponding feasible points might not exist in general.

Next we investigate under which situations these two operations can be replaced by the usual minimization operator so as to simplify the analysis on conventional minimization problems. We need the following assumption in the rest of this subsection.

Assumption 5. $\Omega \cap \mathrm{ri}\mathcal{F}^b \neq \emptyset$.

Note that under Assumption 5 and by virtue of [8, Theorem 2.2.3], we have

$$(\zeta \circ \mathcal{A})^*(v) = \min\{\langle c, x\rangle \mid Ax = v,\ x \in \mathcal{F}\} \quad \forall\, v \in \mathrm{dom}(\zeta \circ \mathcal{A})^*.$$

Let (P_v) denote the perturbed problem

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & \langle c, x\rangle \\ \text{s.t.} \quad & Ax - a = v, \\ & x \in \mathcal{F}, \end{aligned}$$

where v serves as the perturbation parameter. We refer to the original problem (1) where the objective function is taken as an affine function, denoted by (P_0), to the unperturbed problem. We denote the optimal value function of (P_v) by $f_{\mathrm{val}}(v)$. Accordingly, $f_{\mathrm{val}}(0)$ denotes the optimal value of the original problem (1) or (P_0).

Then, we derive the following result immediately by virtue of Proposition 6.

Proposition 7. *Suppose that Assumption* 5 *holds. Then*

$$f_{\mathrm{val}}(v) = \eta^*(v) - \tfrac{1}{2}\|v\|^2_{M^{-1}}$$

for any $v \in dom(\zeta \circ \mathcal{A})^* - a$.

Note that the above result enhances Proposition 6. It provides a new and interesting characteristic of convex conjugates in perturbation analysis. Note that the result is valid only if the parameter v belongs to the set $\mathrm{dom}(\zeta \circ \mathcal{A})^* - a$. Also, this result has a potential role in studying sensitivity analysis and some stochastic programs, both theoretically and numerically.

The next immediate question is about the nonemptiness of the domain of $(\zeta \circ \mathcal{A})^*$. Consider the case when the original problem (P_0) is bounded below; by definition, it follows that

$$\begin{aligned} & (\zeta \circ \mathcal{A})^*(a) \\ &= \min\{\langle c, x\rangle \mid Ax = a,\ x \in \mathcal{F}\} < \infty. \end{aligned}$$

Thus, $a \in \mathrm{dom}(\zeta \circ \mathcal{A})^*$. This implies that $\mathrm{dom}(\zeta \circ \mathcal{A})^*$ is nonempty, and so is $\mathrm{dom}(\zeta \circ \mathcal{A})^* - a$.

Before ending this section, we derive the following result based on the above arguments.

Proposition 8. *Suppose that the original problem, namely,*

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & \langle c, x\rangle \\ \text{s.t.} \quad & Ax = a, \\ & f_i(x) \leq 0, \quad i \in \hat{I} = \{1, \ldots, \theta\}, \end{aligned}$$

is bounded below. Then, dom$(h \circ \mathcal{A})^* \neq \emptyset$ *and* $a \in dom(h \circ \mathcal{A})^*$.

Furthermore, let $\{v^k\}$ *be a sequence in dom*$(h \circ \mathcal{A})^* - a$ *satisfying* $v^k \to 0$ *as* $k \to \infty$*; then*

$$\lim_{v^k \to 0} \eta^*(v^k) = \lim_{v^k \to 0} \left(f_{\text{val}}(v^k) + \tfrac{1}{2}\|v^k\|^2_{M^{-1}} \right)$$
$$= f_{\text{val}}(0) = \min\{\langle c, x\rangle \mid Ax - a = 0,\ x \in \mathcal{F}\}.$$

Remark 6. Note that Assumption 4 used in section 4.1 is obviously stronger than Assumption 5. In other words, the former implies the latter, but not vice versa. Hence, the results obtained in Propositions 6–8 will be valid under Assumption 4. In Proposition 8, we assume that problem (P_0) is bounded below. This assumption is natural and reasonable in optimization. Proposition 8 tells us that the optimal value of the unperturbed optimization problem (the original problem) can be achieved by solving a sequence of the conjugates which corresponds to the perturbed problems, in which affine equality constraints are perturbed on the right-hand side, and setting the perturbation parameters driven to zero. This result helps us to better understand the conjugate and Lagrange dual, and it might serve to study multistage stochastic nonlinear convex programs.

Also, this kind of perturbation problem is closely related to the perturbation problems discussed in [3]. In [9], Magnanti showed the equivalence between Fenchel dual and Lagrangian dual problems where the convex conjugate was employed. We believe that the results established in this subsection complement his theory to some extent. In addition, note that η is originally obtained from the Moreau–Yosida regularization by relaxing the original problem using the Lagrangian dual. Its conjugate η^*, as shown in Propositions 6–8, is related to the parametric (or perturbed) problem of the original problem. From this observation, we see that the perturbation analysis and Lagrangian dual are closely linked under the conjugate operation and Moreau–Yosida regularization. Besides the usual optimization methods, it also provides another possible option for solving some optimization problems, i.e., by solving the induced conjugate.

5. General convex objectives functions. In this section, we investigate the piecewise smoothness and semismoothness of the Lagrangian-dual function φ and the gradient g for the case of the general convex objective functions in (1). We will also provide an alternative way to study the semismoothness of the gradient g based on the structure of the epigraph of φ.

5.1. Convex objective functions with positive definite Hessian. We now discuss the case for the general convex objective functions in (1). Consider the following Lagrangian-dual function φ in (3):

$$\varphi(v) = \sup\{-f(x) + v^T(Ax - a) \mid x \in \mathcal{F}\}.$$

When analyzing the piecewise smoothness of φ in section 3, we frequently use the fact that the optimal solutions of problem (9) lie on the boundary (or facets) of the set $\mathcal{F}$. This fact is guaranteed by Assumption 1, namely $u = A^T v - c \neq 0$, for the linear objective function $f(x) = c^T x$. For nonlinear objective functions, Assumption 1 cannot be made. Thus, multiple optimal solutions of problem (3) may appear in the interior of $\mathcal{F}$, and the piecewise C^2 smoothness of φ may probably be destroyed. This conjecture is confirmed by the following example where $\nabla^2\varphi$ is unbounded in some area, and thus φ is not piecewise C^2.

Example 3. Let

$$f(x) = \begin{cases} 0 & \text{if} \quad \rho \le 1, \\ (\rho - 1)^4 & \text{if} \quad \rho > 1, \end{cases}$$

where $x \in \mathbb{R}^2$ and $\rho = \sqrt{x_1^2 + x_2^2}$, and let

$$A = \begin{pmatrix} 1 & 1 \end{pmatrix}, \quad a = 0, \quad \mathcal{F} = \{x \in \mathbb{R}^2 : x_1^2 + x_2^2 \le 4\}.$$

Obviously, $f(x)$ is convex and twice continuously differentiable on $\mathbb{R}^2$. After some manipulations we obtain

$$\varphi(v) = \sqrt{2}|v| + (3/4)|v|^{4/3}$$

for v in a neighborhood of zero, namely $\mathcal{N} = \{v \in \mathbb{R} : |v| < 2\sqrt{2}\}$. Since $\mathcal{F}$ is bounded, the effective domain of φ is the whole space $\mathbb{R}$. On $\mathbb{R} \setminus \mathcal{N}$, the function φ has a different form. For our purpose, the investigation of φ within $\mathcal{N}$ suffices. Thus we do not elaborate φ outside $\mathcal{N}$. For any $0 \ne v \in \mathcal{N}$, the first- and second-order derivatives of φ are

$$\varphi'(v) = \sqrt{2}\,\mathrm{sign}(v) + v^{1/3}, \qquad \varphi''(v) = (1/3)v^{-2/3}.$$

Now for any nonzero $v \to 0$, we have $\varphi''(v) \to \infty$. Using the same arguments as in Example 1, we can see that φ cannot be piecewise C^2 in the neighborhood $\mathcal{N}$.

This example shows that we cannot extend the results in sections 3 and 4 to problems with arbitrary convex objective functions. However, if the objective function $f(x)$ possesses a positive definite Hessian, we can obtain results similar to those in sections 3 and 4. Also, in this case, the constraints need not be strictly convex.

PROPOSITION 9. *Let φ be defined by* (3)*, where f and f_i, $i \in \hat{I}$, are C^2 convex functions on $\mathbb{R}^n$. Suppose that the Hessian of f is positive definite, and for any facet Q of $\mathcal{F}$ with the index set I_Q and for any $x \in Q$, $\{\nabla f_i(x)\}_{i \in I_Q}$ are linearly independent. Suppose also that $\mathcal{F}$ is nonempty and bounded. Then the Lagrangian-dual function φ is piecewise C^2, and the gradient g of the Moreau–Yosida regularization η is piecewise smooth, and thereby semismooth, on $\mathbb{R}^m$.*

Proof. Similar to the analysis in section 3, we shall construct a piece corresponding to each facet of $\mathcal{F}$. There is one major difference we should highlight. For the problem with a nonlinear objective function, maximizers of the problem (3) can lie on the boundary as well as in the interior of $\mathcal{F}$, while in the case of linear objective functions, Assumption 1 prohibits interior maximizers. Thus, in the present case, an additional piece corresponding to the interior of $\mathcal{F}$ is needed.

Here we define the function ζ slightly differently from the approach in (10):

$$\zeta(u) = \sup\{u^T x - f(x) \mid x \in \mathcal{F}\}. \tag{28}$$

Then $\varphi(v) = \zeta(A^T v) - a^T v$. For each facet Q (on the boundary) of $\mathcal{F}$, we still construct a piece by a slightly different definition:

$$W := \left\{ (x, \lambda) \in \mathbb{R}^n \times \mathbb{R}^{|I_Q|} : f_i(x) = 0,\ i \in I_Q,\ \nabla^2 f(x) + \sum_{i \in I_Q} \lambda_i \nabla^2 f_i(x) \succ 0 \right\},$$

$$U := \left\{ u = \nabla f(x) + \sum_{i \in I_Q} \lambda_i \nabla f_i(x) \in \mathbb{R}^n : (x, \lambda) \in W \right\},$$

and

$$\Gamma(x,\lambda) := \begin{pmatrix} \nabla f(x) + \sum_{i\in I_Q} \lambda_i \nabla f_i(x) \\ \tilde{f}(x) \end{pmatrix}.$$

Then the result of Lemma 2 can be analogously proved, and a piece can be constructed.

In addition, a piece corresponding the interior of $\mathcal{F}$ will be constructed as follows. Since $\nabla^2 f(x) \succ 0$, for any $u \in \mathbb{R}^n$,

$$\nabla f(x) = u$$

has a unique solution, denoted by $\xi_{0x}(u)$. In other words, $\xi_{0x}(u)$ is the unique maximizer of the unconstrained problem

$$\max_{x\in\mathbb{R}^n} \{u^T x - f(x)\}.$$

Now, this piece is defined by $U_0 = \mathbb{R}^n$, $\zeta_0(u) = u^T \xi_{0x}(u) - f(\xi_{0x}(u))$, and $D_0 = \mathrm{cl} D_{\mathrm{int}}$, where $D_{\mathrm{int}} = \{u \mid \xi_{0x}(u) \in \mathrm{int}\mathcal{F}\}$. For any $u \in D_{\mathrm{int}}$, since the unique maximizer $\xi_{0x}(u)$ of the objective function $u^T x - f(x)$ is in the interior of the set $\mathcal{F}$, $\xi_{0x}(u)$ is the optimal solution to the constrained problem (28), too. Thus

$$\zeta(u) = u^T \xi_{0x}(u) - f(\xi_{0x}(u)) = \zeta_0(u). \tag{29}$$

Since ζ and ζ_0 are continuous, thus $\zeta(u) = \zeta_0(u)$ also holds for all $u \in D_0$. It is also easy to verify that

$$\nabla\zeta_0(u) = \xi_{0x}(u) \qquad \forall\, u \in U_0. \tag{30}$$

Now an analogue of the proof of Proposition 4 is valid to prove the piecewise-C^2 smoothness of ζ with the representation $\{(D_0, U_0, \zeta_0), (D_1, U_1, \zeta_2), \ldots, (D_q, U_q, \zeta_q)\}$. (The only difference is that now the nonnegative vector $\bar{\lambda}$ need not be nonzero since $u \neq 0$ is not assumed. Still, $\nabla^2 f(x) + \sum \bar{\lambda}_i \nabla^2 f_i(x) \succ 0$ because $\nabla^2 f(x) \succ 0$. This implies that the Jacobian of Γ is invertible.) Therefore, φ is piecewise C^2 on its domain.

The proof of the piecewise smoothness of g follows from Lemmas 5 and 6 and Proposition 5. The proofs of Lemmas 5 and 6 do not directly rely on Assumption 1, and thus they can be extended without changing to the representation $\{(E_0, V_0, \varphi_0), (E_1, V_1, \varphi_1), \ldots, (E_q, V_q, \varphi_q)\}$ of the Lagrangian-dual function φ of the present problem. □

5.2. Piecewise smoothness under the structure of the epigraph. In this subsection, we investigate the piecewise smoothness and the semismoothness of g using a different approach. In the analysis we will employ the piecewise smoothness or the semismoothness of the metric projection mapping under the structure of the epigraph of the underlying function. Our analysis is based on the framework of [11].

Recently, Meng, Sun, and Zhao [11] investigated the Moreau–Yosida regularization of a lower semicontinuous convex function, $\gamma : Z \to \mathbb{R} \cup \{+\infty\}$, and derived the semismoothness of the solution to the Moreau–Yosida regularization under the structure of the epigraph of γ. Here, Z is a finite dimensional vector space equipped with a scalar product, and the Moreau–Yosida regularization of γ is defined in the form of

$$\begin{aligned} \hat{\gamma}_\epsilon(u) := \min & \left\{\gamma(z) + \tfrac{\epsilon}{2}\langle u - z, u - z\rangle\right\} \\ \text{s.t. } & z \in Z, \end{aligned} \tag{31}$$

where ϵ is a positive number. Let Υ be the epigraph of γ; i.e., $\Upsilon := \mathrm{epi}(\gamma) = \{(u,t) \in Z \times \mathbb{R} \mid t \geq \gamma(u)\}$. Noticing that Υ is a closed convex set, problem (31) then can be written as

$$\begin{array}{ll} \min & \left\{ \frac{1}{\epsilon} t + \frac{1}{2} \langle u - z, u - z \rangle \right\} \\ \text{s.t.} & (z,t) \in \Upsilon. \end{array} \tag{32}$$

For any closed convex set D of Z and $z \in Z$, let $\Pi_D(z)$ denote the metric projection of z onto D, namely,

$$\Pi_D(z) := \mathrm{argmin} \left\{ \tfrac{1}{2} \|d - z\|^2 \mid d \in D \right\}.$$

Let $(z(u), t(u))$ be the unique optimal solution of (32), where $t(u) := \gamma(z(u))$. Define the mapping H by

$$H(z,t,u) := \begin{pmatrix} z \\ t \end{pmatrix} - \Pi_\Upsilon(G(z,t,u)),$$

where $G(z,t,u) := (u^T \;\; t - 1/\epsilon)^T$. Then, it follows from [11] that

$$H(z(u), t(u), u) = 0, \quad G(z(u), t(u), u) \notin \Upsilon \quad \forall\, u \in Z.$$

The following proposition is taken from [11, Theorem 4].

PROPOSITION 10. *For $u_0 \in Z$, let $z_0 := z(u_0)$ and $t_0 := \gamma(z(u_0))$. Then, $(z(\cdot), t(\cdot))$ is semismooth at u_0 if $\Pi_\Upsilon(G(z_0,t_0,u_0))_z \in \mathrm{int}(\mathrm{dom}\gamma)$ and $\Pi_\Upsilon(\cdot)$ is semismooth at $G(z_0,t_0,u_0)$.*

Here we consider the case where $M = \lambda I$ in the Moreau–Yosida regularization as defined in (5), where I is the identity matrix of $\mathbb{R}^{m \times m}$ and $\lambda > 0$. For $v \in \mathbb{R}^m$, let $w(v)$ denote the unique solution of (5), $s(v) := \varphi(w(v))$, and $\mathrm{epi}(\varphi)$ denote the epigraph of φ. Evidently, $(w(v), s(v))$ is the unique solution of

$$\begin{array}{ll} \min & \left\{ \frac{1}{\lambda} s + \frac{1}{2} \langle v - w, v - w \rangle \right\} \\ \text{s.t.} & (w,s) \in \mathrm{epi}(\varphi), \end{array}$$

which is a reformulation of (5). Note that

$$g(v) = \nabla \eta(v) = \lambda(v - w(v)).$$

Hence to study the semismoothness of g, we need only to study the properties of $w(\cdot)$. Set

$$\Phi(w,s,v) := \begin{pmatrix} w \\ s \end{pmatrix} - \begin{pmatrix} w - v \\ 1/\lambda \end{pmatrix} = \begin{pmatrix} v \\ s - 1/\lambda \end{pmatrix}.$$

According to Proposition 10 and following the arguments as in [11], we then have the following result.

PROPOSITION 11. *For $\bar{v} \in \mathbb{R}^m$, let $\bar{w} := w(\bar{v})$, $\bar{s} := \varphi(w(\bar{v}))$. Suppose that $\Pi_{\mathrm{epi}(\varphi)}(\Phi(\bar{w}, \bar{s}, \bar{v}))_w \in \mathrm{int}(\mathrm{dom}\varphi)$ and $\Pi_{\mathrm{epi}(\varphi)}(\cdot)$ is semismooth at $\Phi(\bar{w}, \bar{s}, \bar{v})$. Then $(w(\cdot), s(\cdot))$ is semismooth at $\bar{v}$. Thereby, g is semismooth at $\bar{v}$.*

Furthermore, if φ is finite valued everywhere and $\Pi_{\mathrm{epi}(\varphi)}(\cdot)$ is semismooth on $\mathbb{R}^m \times \mathbb{R}$, then g is semismooth on $\mathbb{R}^m$.

Similar to the mapping H above, we define a mapping Ξ corresponding to the regularization (5),

$$\Xi(w,s,v) := \begin{pmatrix} w \\ s \end{pmatrix} - \Pi_{\mathrm{epi}(\varphi)}(\Phi(w,s,v)).$$

Thus, for any $v \in \mathbb{R}^m$

$$\Xi(w(v), s(v), v) = 0. \tag{33}$$

We now obtain the following result concerning the piecewise smoothness of g.

PROPOSITION 12. *Let $\bar{v} \in \mathbb{R}^m$. Suppose that* (i) $\Pi_{\mathrm{epi}(\varphi)}(\Phi(\bar{w},\bar{s},\bar{v}))_w \in \mathrm{int}(\mathrm{dom}\varphi)$, *and* (ii) $\Pi_{\mathrm{epi}(\varphi)}(\cdot)$ *is piecewise C^k on a neighborhood $\mathcal{N}_1$ of $(\bar{v}, \varphi(w(\bar{v})) - 1/\lambda)$, where $\bar{w} = w(\bar{v})$ and $\bar{s} = \varphi(w(\bar{v}))$. Then, $(w(\cdot), s(\cdot))$ is piecewise C^k on a neighborhood $\mathcal{N}_2$ of $\bar{v}$. Thereby, g is piecewise C^k on $\mathcal{N}_2$. In particular, g is semismooth on a neighborhood of $\bar{v}$.*

Proof. Define a mapping $\aleph : \mathbb{R}^m \times \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^m \times \mathbb{R} \times \mathbb{R}^m$ by

$$\aleph(w,s,v) = \begin{pmatrix} \Xi(w,s,v) \\ v - \bar{v} \end{pmatrix}.$$

By assumption, since $\Pi_{\mathrm{epi}(\varphi)}(\cdot)$ is piecewise C^k on $\mathcal{N}_1$, it is easy to see that $\aleph(\cdot)$ is piecewise C^k on some neighborhood of $(\bar{w},\bar{s},\bar{v})$, and

$$\aleph(\bar{w},\bar{s},\bar{v}) = 0. \tag{34}$$

Next, we show that every matrix in $\partial\aleph(\bar{w},\bar{s},\bar{v})$ is nonsingular [2]. To do so, it is not hard to see that we only need to show the nonsingularity of $\pi_{(w,s)}\partial\Xi(\bar{w},\bar{s},\bar{v})$. For any $V \in \pi_{(w,s)}\partial\Xi(\bar{w},\bar{s},\bar{v})$, it follows that there exists $W \in \partial\Pi_{\mathrm{epi}(\varphi)}(\Phi(\bar{w},\bar{s},\bar{v}))$ such that

$$V = I_{m+1} - W\left(I_{m+1} - \begin{bmatrix} I_m & 0 \\ 0 & 0 \end{bmatrix}\right),$$

where W is a convex combination of some finitely many matrices in $\partial_B\Pi_{\mathrm{epi}(\varphi)}(\Phi(\bar{w}, \bar{s},\bar{v}))$. Suppose $W_i \in \partial_B\Pi_{\mathrm{epi}(\varphi)}(\Phi(\bar{w},\bar{s},\bar{v}))$ and $\lambda_i \geq 0$, $i = 1,\ldots,\nu$, satisfying $\sum_{i=1}^{\nu}\lambda_i = 1$, such that $W = \sum_{i=1}^{\nu}\lambda_i W_i$, where each W_i is in the form of $W_i = \left[\begin{smallmatrix} U_i & \alpha_i \\ \alpha_i^T & \beta_i \end{smallmatrix}\right]$ with $U_i \in \mathbb{R}^{m\times m}$, $\alpha_i \in \mathbb{R}^m$, and $\beta_i \geq 0$. Thus,

$$W = \begin{bmatrix} \sum_{i=1}^{\nu}\lambda_i U_i & \sum_{i=1}^{\nu}\lambda_i\alpha_i \\ \sum_{i=1}^{\nu}\lambda_i\alpha_i^T & \sum_{i=1}^{\nu}\lambda_i\beta_i \end{bmatrix}.$$

To ease the notation, we write $W = \left[\begin{smallmatrix} U & \alpha \\ \alpha^T & \beta \end{smallmatrix}\right]$. Then, by [11, Proposition 3], there exists $\varrho_i \in (0,1)$, $i = 1,\ldots,\nu$, such that

$$0 \leq \beta_i \leq \varrho_i < 1 \qquad \forall\, i.$$

Hence,

$$0 \leq \beta < 1. \tag{35}$$

Then, we have

$$V = I_{m+1} - \begin{bmatrix} U & \alpha \\ \alpha^T & \beta \end{bmatrix} \left(I_{m+1} - \begin{bmatrix} I_m & 0 \\ 0 & 0 \end{bmatrix} \right)$$

$$= I_{m+1} - \begin{bmatrix} U & \alpha \\ \alpha^T & \beta \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

$$= I_{m+1} - \begin{bmatrix} 0 & \alpha \\ 0 & \beta \end{bmatrix} = \begin{bmatrix} I_m & -\alpha \\ 0 & 1-\beta \end{bmatrix}.$$

This together with (35) implies that $\det V = 1 - \beta > 0$ for any $V \in \pi_{(w,s)} \partial \Xi(\bar{w}, \bar{s}, \bar{v})$. So, $\Xi(w, s, v)$ is coherently oriented with respect to w and s at $(\bar{w}, \bar{s}, \bar{v})$ [17, 20]. Thereby, $\pi_{(w,s)} \partial \Xi(\bar{w}, \bar{s}, \bar{v})$ is nonsingular, and so is $\partial \aleph(\bar{w}, \bar{s}, \bar{v})$. Then, by [15, Theorem 6], $\aleph$ is a locally Lipschitz homeomorphism near $(\bar{w}, \bar{s}, \bar{v})$, and $\operatorname{sgn} \det V = \operatorname{ind}(\aleph, (\bar{w}, \bar{s}, \bar{v})) = \pm 1$ for any $V \in \partial_B \aleph(\bar{w}, \bar{s}, \bar{v})$. Further, noticing that $\aleph(\cdot)$ is coherently oriented at $(\bar{w}, \bar{s}, \bar{v})$ and is piecewise C^k on a neighborhood of $(\bar{w}, \bar{s}, \bar{v})$, then by [17, Theorem 5], it follows that $\aleph$ is a PC^k-homeomorphism near $(\bar{w}, \bar{s}, \bar{v})$. Thus, the desired results follow immediately. This completes the proof. □

Remark 7. The condition $\Pi_{\mathrm{epi}(\varphi)}(\Phi(\bar{w}, \bar{s}, \bar{v}))_w \in \mathrm{int}(\mathrm{dom}\varphi)$ in Proposition 12 holds automatically if φ is finite valued everywhere. The obtained results complement and enrich the framework of piecewise smooth functions [20, 17], and also enhance the recent results on the Moreau–Yosida regularization [11].

6. Conclusion. The Lagrangian dual is widely used for large-scale problems. A significant feature of the Lagrangian-dual function φ is the piecewise smoothness, which is studied in this paper and employed in the analysis of the Moreau–Yosida regularization of φ. We investigate the semismoothness of the gradient g of the Moreau–Yosida regularization of φ, which plays a key role in the superlinear or quadratic convergence analysis of generalized Newton methods for solving nonsmooth equations. As to problem (1) with the linear objective function, we show that the Lagrangian-dual function φ is piecewise C^2 and the gradient g is piecewise smooth and thereby semismooth if the inequality constraints in (1) either are all affine or all possess positive definite Hessian matrices. An example with an affine constraint and a strictly convex constraint is constructed. We find that the Lagrangian-dual function of this problem is *not* piecewise C^2, and that the gradient g of its Moreau–Yosida regularization *is* still semismooth. However, whether or not g is semismooth for general mixed affine and strictly convex constraints is still left unanswered. We also investigate problem (1) with a convex objective function. We show with an example that φ may not be piecewise C^2 for the problem with a general convex objective function. For problem (1) with an objective function which possesses a positive definite Hessian, φ and g can again be shown to be piecewise C^2 and semismooth, respectively. We have also provided an alternative way to study the semismoothness/piecewise smoothness of g under the structure of the epigraph of the Lagrangian dual function using the projection operator. For problem (1) with a linear objective, we have also established an interesting result characterizing the relations between the original problem and the Fenchel conjugate of the regularization of the Lagrangian dual problem. For future research, we will examine under which conditions the projection mapping over the epigraph of the Lagrangian-dual function φ is piecewise smooth or semismooth.

Acknowledgment. We would like to thank the two anonymous referees for their constructive suggestions, which helped improve the presentation of this paper.

REFERENCES

[1] X. Chen, Z. Nashed, and L. Qi, *Smoothing methods and semismooth methods for nondifferentiable operator equations*, SIAM J. Numer. Anal., 38 (2000), pp. 1200–1216.
[2] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley and Sons, New York, 1983.
[3] F. H. Clarke, Y. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Nonsmooth Analysis and Control Theory*, Springer-Verlag, New York, 1998.
[4] A. L. Dontchev, H.-D. Qi, and L. Qi, *Convergence of Newton method for convex best interpolation*, Numer. Math., 87 (2001), pp. 435–456.
[5] A. L. Dontchev, H.-D. Qi, and L. Qi, *Quadratic convergence of Newton method for convex interpolation and smoothing*, Constr. Approx., 19 (2003), pp. 1230–1143.
[6] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Springer, New York, 2003.
[7] M. Fukushima and L. Qi, *A globally and superlinearly convergent algorithm for nonsmooth convex minimization*, SIAM J. Optim., 6 (1996), pp. 1106–1120.
[8] J. B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms*, Springer-Verlag, Berlin, 1993.
[9] T. L. Magnanti, *Fenchel and Lagrange dual are equivalent*, Math. Program., 7 (1974), pp. 253–258.
[10] F. Meng and Y. Hao, *The property of piecewise smoothness of Moreau-Yosida approximation for a piecewise C^2 convex function*, Adv. Math. (China), 30 (2001), pp. 354–358.
[11] F. Meng, D. Sun, and G. Zhao, *Semismoothness of solutions to generalized equations and the Moreau-Yosida regularization*, Math. Program., 104 (2005), pp. 561–581.
[12] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optim., 15 (1977), pp. 959–972.
[13] R. Mifflin, L. Qi, and D. Sun, *Properties of the Moreau-Yosida regularization of a piecewise C^2 convex function*, Math. Program., 84 (1999), pp. 269–281.
[14] J. J. Moreau, *Proximite et dualite dans un espace hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.
[15] J.-S. Pang, D. Sun, and J. Sun, *Semismooth homeomorphisms and strong stability of semidefinite and Lorentz complementarity problems*, Math. Oper. Res., 28 (2003), pp. 39–63.
[16] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Program., 58 (1993), pp. 353–367.
[17] D. Ralph and S. Scholtes, *Sensitivity analysis of composite piecewise smooth equations*, Math. Program., 76 (1997), pp. 593–612.
[18] R. T. Rockafellar, *Convex Analysis*, Princeton, NJ, 1970.
[19] A. Ruszczyński and A. Shapiro, *Stochastic Programming, Handbook in Operations Research and Management Science*, A. Ruszczyński and A. Shapiro, eds., Elsevier Science, Amsterdam, 2003.
[20] S. Scholtes, *Introduction to Piecewise Smooth Equations*, Habilitation, University of Karlsruhe, Karlsruhe, Germany, 1994.
[21] D. Sun and J. Han, *On a conjecture in Moreau-Yosida regularization of a nonsmooth convex function*, Chinese Sci. Bull., 42 (1997), pp. 1140–1143.
[22] C. Y. Wang, X. Q. Yang, and X. M. Yang, *Nonlinear Lagrange duality theorems and penalty function methods in continuous optimization*, J. Global Optim., 27 (2003), pp. 473–484.
[23] K. Yosida, *Functional Analysis*, Springer-Verlag, Berlin, 1964.
[24] G. Zhao, *A Lagrangian dual method with self-concordant barrier for multi-stage stochastic convex nonlinear programming*, Math. Program., 102 (2005), pp. 1–24.

SIAM J. OPTIM.
Vol. 19, No. 1, pp. 62–76

LINEAR REGULARITY FOR A COLLECTION OF SUBSMOOTH SETS IN BANACH SPACES*

XI YIN ZHENG† AND KUNG FU NG‡

Abstract. Using variational analysis, we study the linear regularity for a collection of finitely many closed sets. In particular, we extend duality characterizations of the linear regularity for a collection of finitely many closed convex sets to the possibly nonconvex setting. Moreover, the sharpest linear regularity constant can also be dually represented under the subsmoothness assumption.

Key words. approximate projection theorem, linear regularity, normal cone, subsmoothness

AMS subject classifications. 90C31, 90C25, 49J52, 46B20

DOI. 10.1137/060659132

1. Introduction.

Linear regularity is a well-known notion in mathematical programming and approximation theory. In particular, it plays a key role in establishing a linear convergence rate of iterates generated by the cyclic projection algorithm for finding the projection from a point to the intersection of finitely many closed convex sets (see [3] and references therein).

In this paper, we study the linear regularity of a collection $\{A_1, \ldots, A_n\}$ of finitely many closed sets in a Banach space X. Here we say that the collection is locally linearly regular at $a \in \bigcap_{i=1}^n A_i$ (resp., linearly regular) if there exists $\tau \in (0, +\infty)$ such that

$$d\left(x, \bigcap_{i=1}^n A_i\right) \le \tau \sum_{i=1}^n d(x, A_i) \quad \text{for all } x \text{ close to } a \tag{1.1}$$

$$\left(\text{resp.,} \quad d\left(x, \bigcap_{i=1}^n A_i\right) \le \tau \sum_{i=1}^n d(x, A_i) \quad \text{for all } x \in X\right).$$

Linear regularity has been well studied by many authors in the case when each A_i is a closed convex set (see [2, 3, 4, 5, 18, 25] and references therein). In 1972, Jameson [12] established a dual characterization for the linear regularity of a collection of two closed convex cones. Pang [21] and Lewis and Pang [15] provided necessary conditions for the linear regularity of a collection of finitely many closed convex sets in terms of the normal cone. Afterwards, Bauschke, Borwein, and Li [4] established some sufficient conditions in the same line. Recently, it was proved (cf. [2, 20, 25]) that if $\{A_1, \ldots, A_n\}$ is a collection of closed convex sets in a Banach space X, then the following statements are equivalent:

*Received by the editors May 5, 2006; accepted for publication (in revised form) August 13, 2007; published electronically February 6, 2008.

http://www.siam.org/journals/siopt/19-1/65913.html

†Department of Mathematics, Yunnan University, Kunming 650091, P. R. China (xyzheng@ynu.edu.cn). The research of this author was supported by the National Natural Science Foundation of P. R. China (grant 10361008) and the Natural Science Foundation of Yunnan Province, China (grant 2003A0002M).

‡Department of Mathematics, The Chinese University of Hong Kong, Shatin, New Territory, Hong Kong (kfng@math.cuhk.edu.hk). The research of this author was supported by an Earmarked Grant from the Research Grant Council of Hong Kong.

(C1) $\{A_1, \ldots, A_n\}$ is linearly regular.
(C2) There exists $\tau \in (0, +\infty)$ such that

$$N\left(\bigcap_{i=1}^{n} A_i, x\right) \cap B_{X^*} \subset \tau \sum_{i=1}^{n} N(A_i, x) \cap B_{X^*} \quad \text{for all } x \in \bigcap_{i=1}^{n} A_i,$$

where X^* denotes the dual space of X and B_{X^*} denotes the closed unit ball of X^*.
(C3) There exists $\tau \in (0, +\infty)$ such that for any $x \in \bigcap_{i=1}^{n} A_i$,

$$\text{(SC)} \qquad N\left(\bigcap_{i=1}^{n} A_i, x\right) = \sum_{i=1}^{n} N(A_i, x)$$

and

$$\inf\left\{\sum_{i=1}^{n} \|x_i^*\| : \ x_i^* \in N(A_i, x) \text{ and } x^* = \sum_{i=1}^{n} x_i^*\right\} \le \tau \|x^*\| \quad \forall x^* \in N\left(\bigcap_{i=1}^{n} A_i, x\right).$$

In the terminology of Deutsch, Li, and Ward [11], (SC) means that the collection has the strong conical hull intersection property (strong CHIP) at x, which has been extensively studied in variational analysis (cf. [4, 5, 10, 11, 16]).

In this paper, we will study the nonconvex case. In view of the fact that a collection $\{A_1, \ldots, A_n\}$ of closed convex sets is linearly regular with a constant τ if and only if $\{A_1, \ldots, A_n\}$ is locally linearly regular at each $a \in \mathrm{bd}(\bigcap_{i=1}^{n} A_i)$ with the same constant, it is natural to adopt the local version when one considers a collection of closed sets. While the equivalences among (C1), (C2), and (C3) are no longer valid if one drops the convexity assumption of some A_i, a natural substitute of convexity in this respect is the subsmoothness, a notion recently introduced and studied by Aussel, Daniilidis, and Thibault [1], which is a generalization of the well-known notion of the prox-regularity (cf. [6, 7, 9, 23, 24] and references therein).

In section 2, we recall some notions in variational analysis and provide some properties of the subsmoothness. In section 3, as an application of the Ekeland variational principle, we provide a kind of approximate projection result for a closed set, which is very useful for our analysis. In section 4, in terms of the subsmoothness and the approximate projection result, we establish sufficient and/or necessary conditions for the local linear regularity of a collection of finitely many subsmooth sets, and extend the equivalences among (C1), (C2), and (C3) to the nonconvex case. Moreover, the constants τ satisfying (1.1) are represented quantitatively by duality formulas.

2. Subsmoothness of a closed set. First we provide some notions in variational analysis. For a closed subset A of a Banach space X and $a \in A$, let $T_c(A, a)$ and $T(A, a)$ denote respectively the Clarke tangent cone and the contingent cone of A at a, which are respectively defined by

$$T_c(A, a) = \liminf_{x \xrightarrow{A} a, t \to 0^+} \frac{A - x}{t} \quad \text{and} \quad T(A, a) = \limsup_{t \to 0^+} \frac{A - a}{t},$$

where $x \xrightarrow{A} a$ means that $x \to a$ with $x \in A$. Thus, $v \in T_c(A, a)$ if and only if, for each sequence $\{a_n\}$ in A converging to a and each sequence $\{t_n\}$ in $(0, \infty)$ decreasing to 0, there exists a sequence $\{v_n\}$ in X converging to v such that $a_n + t_n v_n \in A$ for all n, while $v \in T(A, a)$ if and only if there exist a sequence $\{v_n\}$ converging to v and a sequence $\{t_n\}$ in $(0, \infty)$ decreasing to 0 such that $a + t_n v_n \in A$ for all n.

We denote by $N_c(A,a)$ the Clarke normal cone of A at a, that is,

$$N_c(A,a) := \{x^* \in X^* : \ \langle x^*, h\rangle \le 0 \ \text{ for all } h \in T_c(A,a)\}.$$

For $\varepsilon \ge 0$ and $a \in A$, the nonempty set

$$\hat{N}_\varepsilon(A,a) := \left\{ x^* \in X^* : \ \limsup_{x \xrightarrow{A} a} \frac{\langle x^*, x-a\rangle}{\|x-a\|} \le \varepsilon \right\}$$

is called the set of Fréchet ε-normals of A at a. When $\varepsilon = 0$, $\hat{N}_\varepsilon(A,a)$ is a convex cone which is called the Fréchet normal cone of A at a and is denoted by $\hat{N}(A,a)$.

Let $N(A,a)$ denote the limiting normal cone of A at a, that is,

$$N(A,a) = \limsup_{x \xrightarrow{A} a, \varepsilon \to 0^+} \hat{N}_\varepsilon(A,x).$$

Thus, $x^* \in N(A,a)$ if and only if there exists a sequence $\{(x_n, \varepsilon_n, x_n^*)\}$ in $A \times R_+ \times X^*$ such that $(x_n, \varepsilon_n) \to (a,0)$, $x_n^* \overset{w^*}{\to} x^*$, and $x_n^* \in \hat{N}_{\varepsilon_n}(A, x_n)$ for each n. It is known that

$$\hat{N}(A,a) \subset N(A,a) \subset N_c(A,a)$$

(cf. [17] and [18]).

If A is convex, then $T_c(A,a) = T(A,a)$ and

$$N_c(A,a) = \hat{N}(A,a) = \{x^* \in X^* : \ \langle x^*, x\rangle \le \langle x^*, a\rangle \ \text{ for all } x \in A\}.$$

Recall that a Banach space X is called an Asplund space if every continuous convex function on X is Fréchet differentiable at each point of a dense subset of X. It is well known (cf. [22]) that X is an Asplund space if and only if every separable subspace of X has a separable dual space. In particular, every reflexive Banach space is an Asplund space. When X is an Asplund space, Mordukhovich and Shao [18] proved that

$$N_c(A,a) = \mathrm{cl}^*(\mathrm{co}(N(A,a))) \quad \text{and} \quad N(A,a) = \limsup_{x \xrightarrow{A} a} \hat{N}(A,x), \tag{2.1}$$

where $\mathrm{cl}^*(\cdot)$ denotes the closure with respect to the weak* topology w^*.

Recall that a closed set A in X is said to be prox-regular at $a \in A$ if there exist $\sigma, r > 0$ such that

$$\langle x^* - u^*, x - u\rangle \ge -\sigma \|x-u\|^2$$

whenever $x, u \in A \cap B(a,r)$, $x^* \in N_c(A,x) \cap B_{X^*}$, and $u^* \in N_c(A,u) \cap B_{X^*}$. Readers can find some interesting properties of the prox-regularity in [23] and [24].

As a generalization of the prox-regularity, Aussel, Daniilidis, and Thibault [1] introduced and studied the subsmoothness. A closed set A in X is said to be subsmooth at $a \in A$ if for any $\varepsilon > 0$ there exists $r > 0$ such that

$$\langle x^* - u^*, x - u\rangle \ge -\varepsilon \|x-u\| \tag{2.2}$$

whenever $x, u \in A \cap B(a,r)$, $x^* \in N_c(A,x) \cap B_{X^*}$, and $u^* \in N_c(A,u) \cap B_{X^*}$.

Taking $x^* = 0$, (2.2) reduces to $\langle u^*, x-u\rangle \le \varepsilon\|x-u\|$. On the other hand, noting that $\langle x^*-u^*, x-u\rangle \ge -2\varepsilon\|x-u\|$ if $\langle x^*, u-x\rangle \le \varepsilon\|x-u\|$ and $\langle u^*, x-u\rangle \le \varepsilon\|x-u\|$, it follows that A is subsmooth at $a \in A$ if and only if for any $\varepsilon > 0$ there exists $r > 0$ such that

$$\langle u^*, x-u\rangle \le \varepsilon\|x-u\| \quad \text{for all } x \in A \cap B(u,r) \tag{2.3}$$

whenever $u \in A \cap B(a,r)$ and $u^* \in N_c(A,u) \cap B_{X^*}$. Thus, for any $\varepsilon > 0$ there exists $r > 0$ such that

$$N_c(A,u) \subset \hat{N}_\varepsilon(A,u) \quad \text{for all } u \in A \cap B(a,r),$$

provided that A is subsmooth at a. Hence

$$\text{subsmoothness of } A \text{ at } a \Longrightarrow N_c(A,a) = \hat{N}(A,a). \tag{2.4}$$

Usually, A is said to be Clarke regular at a if $N_c(A,a) = \hat{N}(A,a)$.

It is known (cf. [17, Corollary 1.96]) that $\hat{N}(A,u) \cap B_{X^*} = \hat{\partial} d(\cdot, A)(u)$ for $u \in A$, where $\hat{\partial}$ denotes the Fréchet subdifferential. Hence $x^* \in \hat{N}(A,u) \cap B_{X^*}$ if and only if for any $\varepsilon > 0$ there exists $r > 0$ such that

$$\langle x^*, x-u\rangle \le d(x,A) + \varepsilon\|x-u\| \quad \text{for all } x \in B(u,r). \tag{2.5}$$

The following proposition shows that a strengthened condition similar to (2.5) provides a characterization of the subsmoothness.

PROPOSITION 2.1. *Let A be a closed subset of X. Then A is subsmooth at $a \in A$ if and only if for any $\varepsilon > 0$ there exists $r > 0$ such that*

$$\langle u^*, x-u\rangle \le d(x,A) + \varepsilon\|x-u\| \quad \text{for all } x \in B(a,r) \tag{2.6}$$

whenever $u \in A \cap B(a,r)$ and $u^* \in N_c(A,u) \cap B_{X^*}$.

Proof. Since $d(x,A) = 0$ for all $x \in A$, (2.6) implies (2.3). Hence, the sufficiency part holds. Conversely, suppose that A is subsmooth at $a \in A$. Let $\varepsilon > 0$ and take $r > 0$ such that

$$\langle u^*, z-u\rangle \le \frac{\varepsilon}{2}\|z-u\| \quad \text{for all } z \in A \cap B(a,2r) \tag{2.7}$$

whenever $u \in A \cap B(a,r)$ and $u^* \in N_c(A,u) \cap B_{X^*}$. Let $x \in B(a,r)$, $u \in A \cap B(a,r)$, and $u^* \in N_c(A,u) \cap B_{X^*}$. Then $d(x,A) \le \|x-a\| < r$. Thus, there exists a sequence $\{u_n\}$ in $A \cap B(x,r)$ such that $\|x-u_n\| \to d(x,A)$. Hence $\|u_n - a\| \le \|u_n - x\| + \|x-a\| < 2r$. It follows from (2.7) that

$$\begin{aligned}\langle u^*, x-u\rangle &= \langle u^*, x-u_n\rangle + \langle u^*, u_n-u\rangle \\ &\le \|x-u_n\| + \frac{\varepsilon}{2}\|u_n-u\| \\ &\le \|x-u_n\| + \frac{\varepsilon}{2}(\|u_n-x\| + \|x-u\|).\end{aligned}$$

Letting $n \to \infty$, one has

$$\langle u^*, x-u\rangle \le d(x,A) + \frac{\varepsilon}{2}(d(x,A) + \|x-u\|) \le d(x,A) + \varepsilon\|x-u\|.$$

This shows that the necessity part holds. The proof is completed. ∎

PROPOSITION 2.2. *Let X, Y be Banach spaces, Ω be a closed convex subset of Y, and $g : X \to Y$ be a continuously differentiable function. Let $a \in g^{-1}(\Omega)$ and suppose that $g'(a)$ is surjective, where $g'(a)$ denotes the derivative of g at a. Then there exists a neighborhood U of a such that $g^{-1}(\Omega)$ is subsmooth at each point of $g^{-1}(\Omega) \cap U$.*

Proof. Since $g'(a)$ is surjective, there exists $l > 0$ such that $2lB_Y \subset g'(a)(B_X)$. Since $x \mapsto g'(x)$ is continuous, it follows that there exists $r > 0$ such that

$$lB_Y \subset g'(x)(B_X) \quad \text{for all } x \in B(a, r). \tag{2.8}$$

We first establish the following inclusion:

$$N_c(g^{-1}(\Omega), u) \subset g'(u)^*(N(\Omega, g(u))) \quad \text{for all } u \in g^{-1}(\Omega) \cap B(a, r). \tag{2.9}$$

Suppose to the contrary that there exists $u \in g^{-1}(\Omega) \cap B(a, r)$ such that

$$x^* \in N_c(g^{-1}(\Omega), u) \setminus g'(u)^*(N(\Omega, g(u))). \tag{2.10}$$

Since the adjoint operator $g'(u)^*$ is weak*-weak* continuous and $g'(u)$ is surjective (by (2.8)), $g'(u)^*(N(\Omega, g(u))) \cap B_{X^*}$ is weakly* closed. This and the Krein–Smulian theorem imply that $g'(u)^*(N(\Omega, g(u)))$ is weakly* closed. By (2.10) and the separation theorem, there exists $h_0 \in X$ such that

$$\begin{aligned} \langle x^*, h_0\rangle &> \sup\{\langle g'(u)^*(y^*), h_0\rangle : \ y^* \in N(\Omega, g(u))\} \\ &= \sup\{\langle y^*, g'(u)(h_0)\rangle : \ y^* \in N(\Omega, g(u))\}. \end{aligned}$$

It follows from the convexity of Ω that $\langle x^*, h_0\rangle > 0$ and $g'(u)(h_0) \in T_c(\Omega, g(u))$. Take an arbitrary sequence $\{x_n\}$ in $g^{-1}(\Omega)$ converging to u, and an arbitrary sequence $\{t_n\}$ in $(0, +\infty)$ decreasing to 0. Then $g(x_n) \overset{\Omega}{\to} g(u)$, and hence there exists a sequence $y_n \to g'(u)(h_0)$ such that $g(x_n) + t_n y_n \in \Omega$ for all n. Since g is continuously differentiable, (2.8) and the Lyusternik–Graves theorem (cf. [20, Theorem 1.57]) imply that

$$d(x_n + t_n h_0, g^{-1}(g(x_n) + t_n y_n)) \le L\|g(x_n + t_n h_0) - g(x_n) - t_n y_n\| \tag{2.11}$$

for some $L \in (0, +\infty)$ and all n large enough. Noting that

$$g(x_n + t_n h_0) - g(x_n) = g'(u)(t_n h_0) + o(t_n),$$

it follows that for each n large enough there exists $\tilde{x}_n \in X$ such that

$$\tilde{x}_n \in g^{-1}(g(x_n) + t_n y_n) \subset g^{-1}(\Omega)$$

and

$$\|x_n + t_n h_0 - \tilde{x}_n\| \le 2L(t_n\|g'(u)(h_0) - y_n\| + \|o(t_n)\|).$$

This and $y_n \to g'(u)(h_0)$ imply that $h_n := \frac{\tilde{x}_n - x_n}{t_n} \to h_0$ and $x_n + t_n h_n = \tilde{x}_n \in g^{-1}(\Omega)$. This shows that $h_0 \in T_c(g^{-1}(\Omega), u)$, which is not possible because $x^* \in N_c(g^{-1}(\Omega), u)$ and $\langle x^*, h_0\rangle > 0$. This shows that (2.9) holds. Let $z \in g^{-1}(\Omega) \cap B(a, \frac{r}{2})$ and $\varepsilon > 0$. Then there exists $\delta \in (0, \frac{r}{2})$ such that

$$\|g'(u_1) - g'(u_2)\| < \frac{l\varepsilon}{2} \quad \text{for any } u_1, u_2 \in B(z, 2\delta). \tag{2.12}$$

Let $u \in g^{-1}(\Omega) \cap B(z,\delta)$ and $u^* \in N_c(g^{-1}(\Omega),u) \cap B_{X^*}$. Then, by (2.9), there exists $y^* \in N(\Omega, g(u))$ such that $u^* = (g'(u))^*(y^*)$. Take $y \in B_Y$ such that $\|y^*\| \le 2\langle y^*, y\rangle$. By (2.8), there exists $v \in B_X$ such that $ly = g'(u)(v)$. Hence,

$$l\|y^*\| \le 2\langle y^*, g'(u)(v)\rangle = 2\langle u^*, v\rangle \le 2.$$

By the convexity of Ω, one has

$$\langle y^*, g(x) - g(u)\rangle \le 0 \quad \text{for all } x \in g^{-1}(\Omega).$$

Noting that

$$\langle u^*, x-u\rangle = \langle (g'(u))^*(y^*), x-u\rangle = \langle y^*, g'(u)(x-u)\rangle,$$

it follows that for any $x \in g^{-1}(\Omega) \cap B(u,\delta)$,

$$\begin{aligned}\langle u^*, x-u\rangle &\le \langle y^*, g'(u)(x-u) - (g(x)-g(u))\rangle \\ &\le \frac{2}{l}\|g(x)-g(u)-g'(u)(x-u)\| \\ &\le \frac{2}{l}\|g'(u+\theta(x-u))-g'(u)\|\|x-u\|,\end{aligned}$$

where $\theta \in (0,\ 1)$. Since $\|u+\theta(x-u)-z\| \le \|u-z\| + \theta\|x-u\| < 2\delta$, it follows from (2.12) that

$$\langle u^*, x-u\rangle \le \varepsilon\|x-u\| \quad \text{for any } x \in g^{-1}(\Omega) \cap B(z,\delta).$$

Therefore, $g^{-1}(\Omega)$ is subsmooth at z. The proof is complete. □

We don't know whether Proposition 2.2 holds if the continuous differentiability of g on X is weakened to the strict differentiability of g at a.

The following Proposition 2.3 demonstrates an interesting fact that, in an Asplund space, the subsmoothness on an open subset of A can be described in terms of the Fréchet normal cone (rather than the Clarke normal cone). To do this, we need the following lemma.

LEMMA 2.1. *Let A be a closed subset of X and $a \in A$. Suppose that for any $\varepsilon > 0$ there exists $r > 0$ such that*

$$\langle u^*, x-a\rangle \le \varepsilon\|x-a\| \ \text{ for all } x \in A \cap B(a,r), \ \text{ for all } u^* \in \hat{N}(A,a) \cap B_{X^*}. \tag{2.13}$$

Then $\hat{N}(A,a)$ is weak closed.*

Proof. Let ε be an arbitrary number in $(0,\ +\infty)$ and take $r > 0$ such that (2.13) holds. Since $\hat{N}(A,a)$ is convex, by the Krein–Smulian theorem it suffices to show that $\hat{N}(A,a) \cap B_{X^*}$ is weakly* closed. Let $\{u_j^*\}$ be a net in $\hat{N}(A,a) \cap B_{X^*}$ convergent to $x^* \in X^*$ with respect to the weak* topology. Then, $x^* \in B_{X^*}$ (because B_{X^*} is weakly* closed) and $\langle u_j^*, x\rangle \to \langle x^*, x\rangle$ for all $x \in X$. It follows from (2.13) that $\langle x^*, x-a\rangle \le \varepsilon\|x-a\|$ for all $x \in A \cap B(a,r)$. This and the arbitrariness of ε imply that $x^* \in \hat{N}(A,a)$. Hence $x^* \in \hat{N}(A,a) \cap B_{X^*}$. This shows that $\hat{N}(A,a) \cap B_{X^*}$ is weakly* closed. The proof is complete. □

PROPOSITION 2.3. *Let A be a closed subset of X, and U be an open subset of X. Suppose that X is an Asplund space. Then A is subsmooth at each point of $A \cap U$ if and only if for any $z \in A \cap U$ and $\varepsilon > 0$ there exists $r > 0$ such that* (2.3) *holds whenever $u \in \mathrm{bd}(A) \cap B(z,r)$ and $u^* \in \hat{N}(A,u) \cap B_{X^*}$.*

Proof. Since $\hat{N}(A,x) \subset N_c(A,x)$ for all $x \in A$, the necessity part is clear. For the sufficiency part, we need only show that $N_c(A,z) = \hat{N}(A,z)$ for any $z \in A \cap U$. Let $z \in A \cap U$ and $z^* \in N(A,z)$. Since X is an Asplund space, there exists a sequence $\{(u_n, u_n^*)\}$ in $X \times X^*$ such that

$$u_n \xrightarrow{A} z, \quad u_n^* \xrightarrow{w^*} z^*, \quad \text{and} \quad u_n^* \in \hat{N}(A,u_n).$$

Hence, there exists $M \in (0,\ +\infty)$ such that each $\|u_n^*\| < M$. For any $\varepsilon > 0$, take $r > 0$ such that (2.3) holds for any $u \in B(z,r) \cap \mathrm{bd}(A)$ and $u^* \in \hat{N}(A,u) \cap B_{X^*}$. Without loss of generality, we assume that $u_n \in B(z,r)$ for all n. It follows from (2.3) that

$$\left\langle \frac{u_n^*}{M}, x - u_n \right\rangle \le \varepsilon \|x - u_n\| \quad \text{for all } x \in B(u_n,r) \cap A.$$

Letting $n \to \infty$, one has

$$\left\langle \frac{z^*}{M}, x - z \right\rangle \le \varepsilon \|x - z\| \quad \text{for all } x \in B(z,r) \cap A.$$

This implies that $\limsup_{x \xrightarrow{A} z} \frac{\langle z^*, x-z\rangle}{\|x-z\|} \le 0$, and so $z^* \in \hat{N}(A,z)$. Hence, $N(A,z) \subset \hat{N}(A,z)$. Since the converse inclusion automatically holds, $N(A,z) = \hat{N}(A,z)$. Since X is an Asplund space, it follows from (2.1) that $N_c(A,z) = \mathrm{cl}^*(\mathrm{co}(\hat{N}(A,z)))$. Noting that $\hat{N}(A,z)$ is a convex cone, this means that $N_c(A,z) = \mathrm{cl}^*(\hat{N}(A,z))$. It follows from Lemma 2.1 that $N_c(A,z) = \hat{N}(A,z)$. The proof is complete. □

A natural question is: Can the open set U in Proposition 2.3 be replaced by a singleton $\{a\}$ with $a \in A$? That is, is A subsmooth at a given point a of A if for any $\varepsilon > 0$ there exists $r > 0$ such that (2.3) holds whenever $u \in \mathrm{bd}(A) \cap B(a,r)$ and $u^* \in \hat{N}(A,u) \cap B_{X^*}$?

While we don't have the answer to this question, we can prove the following result similar to the proof of Proposition 2.3.

PROPOSITION 2.4. *Let X be an Asplund space, A be a closed subset of X, and $a \in A$. Then the following two statements are equivalent:*

(i) *For any $\varepsilon > 0$ there exists $r > 0$ such that* (2.3) *holds whenever $u \in A \cap B(a,r)$ and $u^* \in N(A,u) \cap B_{X^*}$.*

(ii) *For any $\varepsilon > 0$ there exists $r > 0$ such that* (2.3) *holds whenever $u \in A \cap B(a,r)$ and $u^* \in \hat{N}(A,u) \cap B_{X^*}$.*

Remark 2.1. Let X, A, and a be as in Proposition 2.4. Then A is Clarke regular at a when (ii) of Proposition 2.4 holds. Indeed, by Proposition 2.4, $N(A,a) = \hat{N}(A,a)$. It follows from (2.1) that

$$N_c(A,a) = \mathrm{cl}^*(\mathrm{co}(\hat{N}(A,a))) = \mathrm{cl}^*(\hat{N}(A,a)).$$

This and Lemma 2.1 show that A is Clarke regular at a.

3. Approximate projection theorem in Banach spaces. Using the Bronstead–Rockafellar theorem, it was proved in [19] that if A is a closed convex nonempty subset of a Banach space X and $x \in X \setminus A$, then for any $\gamma \in (0,\ 1)$ there exist $a \in \mathrm{bd}(A)$ and $a^* \in N(A,a)$ with $\|a^*\| = 1$ such that

$$\gamma \|x - a\| < \min\{d(x,A),\ \langle a^*, x - a\rangle\}. \tag{3.1}$$

By virtue of the well-known Ekeland variational principle, we provide below a nonconvex generalization of the above projection result, which will play a key role in the proofs of our main results in section 4.

THEOREM 3.1. *Let X be a Banach space (resp., Asplund space) and A be a closed nonempty subset of X. Let $\gamma \in (0,\ 1)$. Then for any $x \notin A$ there exist $a \in \mathrm{bd}(A)$ and $a^* \in N_c(A,a)$ (resp., $a^* \in \hat{N}(A,a)$) with $\|a^*\| = 1$ such that (3.1) holds.*

Proof. First suppose that X is an Asplund space. Let $x \in X\backslash A$. Then $d(x,A) > 0$. Let $\varepsilon \in (0,\ +\infty)$ be such that

$$\varepsilon < \frac{(1-\gamma^{\frac{1}{2}})d(x,A)}{4+(2+2\gamma^{\frac{1}{2}})d(x,A)}, \tag{3.2}$$

and take $z_0 \in A$ such that $\|z_0 - z\| < d(x,A) + \varepsilon$. Let $\phi(z) := \|z-x\| + \delta_A(z)$ for all $z \in X$. Then ϕ is a proper lower semicontinuous function and $\phi(z_0) < \inf_{z\in X}\phi(z) + \varepsilon$. By the Ekeland variational principle, there exists $\bar{z} \in A$ such that $\phi(\bar{z}) \le \phi(z_0)$ and $\phi(\bar{z}) \le \phi(z) + \varepsilon\|z-\bar{z}\|$ for all $z \in X$. Hence

$$\|\bar{z} - x\| < d(x,A) + \varepsilon, \tag{3.3}$$

and the continuous convex function $f(z) := \|z-x\| + \varepsilon\|z-\bar{z}\|$ attains its global minimum over A at $\bar{z}$. Noting that $\varepsilon < d(x,A)$, it follows from [17, Theorem 2.33] that there exist $z_1, a \in B(\bar{z},\varepsilon)$ such that

$$z_1 \neq x, \quad a \in A, \quad \text{and} \quad 0 \in \partial f(z_1) + \hat{N}(A,a) + \varepsilon B_{X^*},$$

where ∂ denotes the subdifferential in the sense of convex analysis. Hence there exist $z_1^*, z_2^* \in X^*$ such that

$$\|z_1^*\| = 1, \quad \|z_2^*\| < 2\varepsilon, \quad \langle z_1^*, z_1 - x\rangle = \|z_1 - x\|, \quad \text{and} \quad -z_1^* + z_2^* \in \hat{N}(A,a).$$

It follows from (3.3) that $\|x-a\| < d(x,A) + 2\varepsilon$. This and (3.2) imply that

$$\gamma^{\frac{1}{2}}\|x-a\| < d(x,A). \tag{3.4}$$

Let $a^* := \frac{-z_1^*+z_2^*}{\|-z_1^*+z_2^*\|}$. Then, $a^* \in \hat{N}(A,a)$ and so $a \in \mathrm{bd}(A)$. Note that

$$\begin{aligned}
\langle a^*, x-a\rangle &= \frac{\langle z_1^*, z_1 - x\rangle + \langle z_1^*, a - z_1\rangle + \langle z_2^*, x-a\rangle}{\|-z_1^*+z_2^*\|} \\
&\ge \frac{\|z_1-x\| - 2\varepsilon - 2\varepsilon\|x-a\|}{1+2\varepsilon} \\
&\ge \frac{(1-2\varepsilon)\|x-a\| - 4\varepsilon}{1+2\varepsilon} \\
&\ge \frac{(1-2\varepsilon)d(x,A) - 4\varepsilon}{1+2\varepsilon}.
\end{aligned}$$

This and (3.2) imply that $\gamma^{\frac{1}{2}}d(x,A) < \langle a^*, x-a\rangle$. It follows from (3.4) and $\gamma \in (0,\ 1)$ that (3.1) holds. When X is a general Banach space, the conclusion can be similarly proved (with [8, Corollary, p. 52] replacing [17, Theorem 2.33]). The proof is complete. □

Remark 3.1. When X is a general Banach space, the Clarke normal cone in Theorem 3.1 cannot be replaced by the Fréchet normal cone. Indeed, take X to be

a non-Asplund space (e.g., let $X := l_1$). Then, by [17, Corollary 2.21], there exists a proper closed nonempty subset A of X such that $\hat{N}(A,a) = \{0\}$ for any $a \in \mathrm{bd}(A)$. It follows that for any $x \in X \setminus A$ there does not exist $a \in \mathrm{bd}(A)$ and $a^* \in \hat{N}(A,a)$ with $\|a^*\| = 1$ such that (3.1) holds.

Remark 3.2. In general, in Theorem 3.1 one cannot take $\gamma = 1$ even when A is a closed convex set. For example, let X be a nonreflexive Banach space. Then, by the James theorem there exists $x^* \in X^*$ with $\|x^*\| = 1$ such that

$$\langle x^*, z\rangle < 1 \quad \text{for all } z \in B_X. \tag{3.5}$$

Let $A := \{x \in X : \langle x^*, x\rangle \le 1\}$. We claim that $\langle a^*, x-a\rangle < \|x-a\|$ for any $x \in X\setminus A$, any $a \in A$, and any $a^* \in N(A,a)$ with $\|a^*\| = 1$. Indeed, suppose to the contrary that there exist $x_0 \in X \setminus A$, $a_0 \in A$, and $a_0^* \in N(A,a_0)$ with $\|a_0^*\| = 1$ such that $\langle a_0^*, x_0 - a_0\rangle = \|x_0 - a_0\|$. Then $a_0 \in \mathrm{bd}(A)$; that is, $\langle x^*, a_0\rangle = 1$. By the definition of A, it is clear that $N(A,a_0) = R_+x^*$ and so $a_0^* = x^*$. Hence $\langle x^*, \frac{x_0-a_0}{\|x_0-a_0\|}\rangle = 1$, contradicting (3.5).

4. Main results. In this section, we establish some relationships concerning the local linear regularity of a collection of closed sets in a Banach space. The following proposition, which can be proved by Theorem 3.1 and convex analysis techniques, provides a relationship between the local linear regularity and the linear regularity for a collection of finitely many closed convex sets.

PROPOSITION 4.1. *Let X be a Banach space and $C_1, \ldots, C_n$ be closed convex subsets of X such that $\bigcap_{i=1}^n C_i \neq \emptyset$. Then $\{C_1, \ldots, C_n\}$ is linearly regular if and only if there exists $\tau \in (0, +\infty)$ such that $\{C_1, \ldots, C_n\}$ is locally linearly regular at each point of $\mathrm{bd}(\bigcap_{i=1}^n C_i)$ with the same constant τ.*

In view of Proposition 4.1, we see that it is pertinent to study the local linear regularity for a collection of nonconvex closed sets. For convenience, let us fix some notation. Let $\{A_1, \ldots, A_n\}$ be a collection of closed sets in a Banach space X with intersection A containing a:

$$a \in A := \bigcap_{i=1}^n A_i.$$

The modulus of the linear regularity of the collection $\{A_1, \ldots, A_n\}$ at a is denoted by $\eta(A;a)$ and defined by

$$\eta(A;a) := \inf\{\tau > 0 : \ (1.1) \text{ holds}\}.$$

Thus, $\eta(A;a) < +\infty$ if and only if $\{A_1, \ldots, A_n\}$ is locally linearly regular at a.

We will provide necessary and/or sufficient conditions for the local linear regularity and establish formulas for the modulus $\eta(A;a)$. Let $\tau, \delta \in (0, +\infty)$. For convenience of presenting our results, we list the following inclusions:

$$\hat{N}(A,u) \cap B_{X^*} \subset \tau \sum_{i=1}^n N_c(A_i,u) \cap B_{X^*} \quad \text{for all } u \in A \cap B(a,\delta), \tag{4.1}$$

$$N_c(A,u) \cap B_{X^*} \subset \tau \sum_{i=1}^n N_c(A_i,u) \cap B_{X^*} \quad \text{for all } u \in A \cap B(a,\delta). \tag{4.2}$$

In terms of these two inclusions, we define the following quantities:

$$\beta_c^f(\delta) := \inf\{\tau > 0 : \ (4.1) \text{ holds}\} \quad \text{and} \quad \beta_c(\delta) := \inf\{\tau > 0 : \ (4.2) \text{ holds}\}.$$

THEOREM 4.1. *Let X be a Banach space. Then*

$$\lim_{\delta\to 0^+} \beta_c^f(\delta) \le \eta(A;a). \tag{4.3}$$

If each A_i is subsmooth at a, then

$$\lim_{\delta\to 0^+} \beta_c(\delta) \ge \eta(A;a). \tag{4.4}$$

Consequently, if each A_i is subsmooth at a and if A is Clarke regular at all its points close to a, then

$$\eta(A;a) = \lim_{\delta\to 0^+} \beta_c^f(\delta). \tag{4.5}$$

Proof. If $\eta(A;a) = +\infty$, then (4.3) holds trivially. Next assume that $\eta(A;a) < \infty$. Let $\tau \in (\eta(A;a),\ +\infty)$. Then there exists $\delta_0 > 0$ such that

$$d(x,A) \le \tau \sum_{i=1}^n d(x,A_i) \quad \text{for all } x \in B(a,\delta_0). \tag{4.6}$$

Let $u \in A \cap B(a, \frac{\delta_0}{2})$ and $x^* \in \hat{N}(A,u) \cap B_{X^*}$. Let k be an arbitrary natural number. Then there exists $r \in (0,\ \frac{\delta_0}{2})$ such that (2.5) holds with $\varepsilon = \frac{1}{k}$. Noting that $B(u,r) \subset B(a,\delta_0)$, it follows from (4.6) that

$$\langle x^*, x-u\rangle \le \tau \sum_{i=1}^n d(x,A_i) + \frac{1}{k}\|x-u\| \quad \text{for all } x \in B(u,r).$$

This and [8, Proposition 2.3.3] imply that there exist $u_{k,i}^* \in X^*$ and $x_k^* \in \frac{1}{k}B_{X^*}$ $(i = 1,\dots,n)$ such that

$$u_{k,i}^* \in \partial_c d(\cdot,A_i)(u) \subset N_c(A_i,u) \cap B_{X^*} \quad \text{and} \quad x^* = \tau \sum_{i=1}^n u_{k,i}^* - x_k^*,$$

where ∂_c denotes the Clarke subdifferential. Since B_{X^*} is weakly* compact and $N_c(A,x)$ is weakly* closed, without loss of generality (consider generalized subsequences if necessary) we assume that $u_{k,i}^* \xrightarrow{w^*} u_i^* \in N_c(A_i,u) \cap B_{X^*}$ as $k \to \infty$. Hence

$$x^* = \tau \sum_{i=1}^n u_i^* \in \tau \sum_{i=1}^n N_c(A_i,u) \cap B_{X^*}.$$

This shows that (4.1) holds for any $\delta \in (0,\ \frac{\delta_0}{2})$, and so $\lim_{\delta\to 0^+} \beta_c^f(\delta) \le \tau$. Letting $\tau \to \eta(A;a)$, it follows that (4.3) holds.

Now suppose that each A_i is subsmooth at a. Since (4.4) holds trivially if $\lim_{\delta\to 0^+} \beta_c(\delta) = +\infty$, we assume that $\lim_{\delta\to 0^+} \beta_c(\delta) < +\infty$. Let τ be an arbitrary number in $(\lim_{\delta\to 0^+} \beta_c(\delta),\ +\infty)$. Then there exists $\delta > 0$ such that (4.2) holds. Consider any $\varepsilon \in (0,\ \frac{1}{2})$ with $\tau\varepsilon < 1$. By Proposition 2.1 there exists $\delta_1 \in (0,\ \delta)$ such that

$$\langle x_i^*, x-a_i\rangle \le d(x,A_i) + \frac{\varepsilon}{n}\|x-a_i\| \quad \text{for all } x \in B(a,\delta_1) \tag{4.7}$$

whenever $a_i \in A_i \cap B(a, \delta_1)$, $x_i^* \in N_c(A_i, a_i) \cap B_{X^*}$, and $i = 1, \ldots, n$. Let $r := \frac{\delta_1}{2}$ and $x \in B(a, r) \setminus A$. Then $d(x, A) \le \|x - a\| < r$. Let $\gamma \in (\max\{\frac{d(x,A)}{r}, \tau\varepsilon\}, 1)$. By Theorem 3.1, there exist $z \in \mathrm{bd}(A)$ and $x^* \in N_c(A, z)$ with $\|x^*\| = 1$ such that

$$\langle x^*, x - z\rangle \ge \gamma\|x - z\| \tag{4.8}$$

and $\gamma\|x - z\| \le d(x, A)$. Thus, $\|x - z\| \le \frac{d(x,A)}{\gamma} < r$. Hence

$$\|z - a\| \le \|z - x\| + \|x - a\| < 2r = \delta_1 < \delta.$$

It follows from (4.2) that there exists $x_i^* \in N_c(A_i, z) \cap B_{X^*}$ such that $x^* = \tau \sum_{i=1}^n x_i^*$. By (4.7), one has

$$\langle x^*, x - z\rangle = \tau \sum_{i=1}^n \langle x_i^*, x - z\rangle \le \tau \sum_{i=1}^n \left(d(x, A_i) + \frac{\varepsilon}{n}\|x - z\| \right).$$

This and (4.8) imply that $(\gamma - \tau\varepsilon)\|x - z\| \le \tau \sum_{i=1}^n d(x, A_i)$ and hence

$$d(x, A) \le \frac{\tau}{\gamma - \tau\varepsilon} \sum_{i=1}^n d(x, A_i)$$

(because $z \in A$). Therefore, $\eta(A; a) \le \frac{\tau}{\gamma - \tau\varepsilon}$. It follows that (4.4) holds by letting $\gamma \to 1^-$, $\varepsilon \to 0^+$, and $\tau \to \lim_{\delta \to 0^+} \beta_c(\delta)$. The last assertion of the theorem is obvious and this completes the proof. □

The following example shows that if the subsmoothness assumption is dropped in Theorem 4.1, then inequality (4.4) is not necessarily true even when X is finite dimensional.

Example 4.1. Let $X = R^2$, $A_1 = \{(s, t) \in R^2 : st \le 0\}$, and

$$A_2 = \{(s, t) \in R_+^2 : (s - 1)^2 + t^2 \le 1 \text{ and } s^2 + (t - 1)^2 \le 1\}.$$

By the definition of the Clarke tangent cone, it is easy to verify that $T_c(A_1, (0, 0)) = \{(0, 0)\}$. This means that $N_c(A_1, (0, 0)) = X^*$. Hence

$$N_c(A_1 \cap A_2, (0, 0)) \cap B_{X^*} \subset N_c(A_1, (0, 0)) \cap B_{X^*} + N_c(A_2, (0, 0)) \cap B_{X^*}.$$

On the other hand, for every natural number k, let $x_k = (\frac{1}{k}, (\frac{2}{k} - \frac{1}{k^2})^{\frac{1}{2}})$. Noting that $A_1 \cap A_2 = \{(0, 0)\}$, it is easy to verify that

$$d(x_k, A_1 \cap A_2) = \left(\frac{2}{k}\right)^{\frac{1}{2}}, \quad d(x_k, A_1) = \frac{1}{k} \quad \text{and} \quad d(x_k, A_2) = 0.$$

Hence, $\frac{d(x_k, A_1 \cap A_2)}{d(x_k, A_1) + d(x_k, A_2)} = (2k)^{\frac{1}{2}} \to +\infty$. This shows that $\{A_1, A_2\}$ is not locally linearly regular at $(0, 0)$.

The following proposition shows that the last assertion of Theorem 4.1 remains true even if the Clarke regularity assumption on the intersection A is dropped, provided that X is an Asplund space, but we don't know if this is also so for a general Banach space.

PROPOSITION 4.2. *Let X be an Asplund space and each A_i be subsmooth at a. Then* (4.5) *holds.*

Applying the Asplund space version of Theorem 3.1, by virtue of Theorem 4.1 together with the obvious modification of the proof of inequality (4.4), one can obtain the proof of Proposition 4.2.

The following result (together with Theorem 4.1) implies that, under the local linear regularity assumption, the subsmoothness and Clarke regularity of each A_i imply the subsmoothness and Clarke regularity of the intersection A.

PROPOSITION 4.3. *Let X be an Asplund space and $\lim_{\delta\to 0^+}\beta_c^f(\delta) < +\infty$. Suppose that each A_i is Clarke regular at all points of A close to a. Then A is Clarke regular at all points of A close to a. If, in addition, each A_i is subsmooth at a, then A is subsmooth at a.*

Proof. Take $\delta > 0$ such that each A_i is Clarke regular at each point of $A\cap B(a,\delta)$. Then

$$N_c(A_i,u) = \hat{N}(A_i,u) \quad \text{for all } u \in A\cap B(a,\delta). \tag{4.9}$$

Considering smaller δ if necessary, one can find $\tau \in (\lim_{\delta\to 0^+}\beta_c^f(\delta),\ +\infty)$ such that (4.1) holds. We claim that

$$N_c(A,u) = \hat{N}(A,u) \quad \text{for all } u \in A\cap B(a,\delta). \tag{4.10}$$

Let $u \in A\cap B(a,\delta)$ and $u^* \in N(A,u)$. By (2.1), there exists a sequence $\{(u_k,u_k^*)\}$ in $(A\cap B(a,\delta))\times X^*$ such that $u_k \to u$ and $u_k^* \overset{w^*}{\to} u^*$ with $u_k^* \in \hat{N}(A,u_k)$ for all k. It follows that $\{u_k^*\}$ is bounded. Without loss of generality, we assume that each $u_k^* \in B_{X^*}$ (if necessary, replace (u_k^*,u^*) by $(\gamma u_k^*,\gamma u^*)$ with small a constant $\gamma > 0$). By (4.1), there exist $u_k^*(i) \in N_c(A_i,u_k)\cap B_{X^*}$ such that $u_k^* = \tau\sum_{i=1}^n u_k^*(i)$. We can also assume that

$$u_k^*(i) \overset{w^*}{\to} u^*(i) \in N(A_i,u)\cap B_{X^*}, \quad i = 1,\dots,n \tag{4.11}$$

(passing to subsequences if necessary). Hence, $u^* = \tau\sum_{i=1}^n u^*(i)$. On the other hand, by the definition of the Fréchet normal cone it is easy to verify that

$$\sum_{i=1}^n \hat{N}(A_i,u) \subset \hat{N}(A,u) \quad \text{for all } u \in A. \tag{4.12}$$

It follows from (4.9) and (4.11) that $u^* \in \hat{N}(A,u)$. Therefore, $N(A,u) \subset \hat{N}(A,u)$, and so $N(A,u) = \hat{N}(A,u)$. Since $\hat{N}(A,u)$ is a convex cone, this and (2.1) imply that $N_c(A,u) = \mathrm{cl}^*(\hat{N}(A,u))$. Thus, to prove (4.10), it suffices to show that $\hat{N}(A,u)\cap B_{X^*}$ is weakly* closed (by the Krein–Smulian theorem). Let $\{x_\alpha^*\}$ be a net in $\hat{N}(A,u)\cap B_{X^*}$ convergent to $x^* \in X^*$ with respect to the weak* topology. Then $x^* \in B_{X^*}$ (because B_{X^*} is weakly* compact). By (4.1), there exist $x_\alpha^*(i) \in N_c(A_i,u)\cap B_{X^*}$ $(i = 1,\dots,n)$ such that $x_\alpha^* = \tau\sum_{i=1}^n x_\alpha^*(i)$. Since every Clarke normal cone $N_c(A_i,u)$ is weakly* closed, we can assume that $x_\alpha^*(i) \overset{w^*}{\to} x^*(i) \in N_c(A_i,u)\cap B_{X^*}$ (passing to a subnet if necessary). Hence $x^* = \tau\sum_{i=1}^n x^*(i)$. It follows from (4.9) and (4.12) that $x^* \in \hat{N}(A,u)\cap B_{X^*}$. This shows that $\hat{N}(A,u)\cap B_{X^*}$ is weakly* closed.

Next suppose that each A_i is subsmooth at a. Then, for any $\varepsilon > 0$ there exists $r \in (0,\ \delta)$ such that

$$\langle u^*, x-u\rangle \le \frac{\varepsilon\|x-u\|}{n\tau} \quad \text{for all } x \in A_i\cap B(u,r) \tag{4.13}$$

whenever $u \in A_i \cap B(a,r)$ and $u^* \in N_c(A_i,u) \cap B_{X^*}$. Let $u \in A \cap B(a,r)$ and $u^* \in N_c(A,u) \cap B_{X^*}$. Then, by $0 < r < \delta$, (4.1), and (4.10), there exists $u_i^* \in N_c(A_i,u) \cap B_{X^*}$ such that $u^* = \tau \sum_{i=1}^n u_i^*$. It follows from (4.13) that

$$\langle u^*, x-u\rangle = \tau \sum_{i=1}^n \langle u_i^*, x-u\rangle \le \varepsilon \|x-u\| \text{ for all } x \in A \cap B(u,r).$$

This shows that A is subsmooth at a. The proof is complete. □

We don't know whether Proposition 4.3 can be extended to the Banach space setting. The difficulty lies in the fact that the two equalities in (2.1) (which are essential in our present proof of the proposition) are no longer valid for a general Banach space.

In view of Proposition 4.1, the following corollary can be regarded as a nonconvex extension of the equivalences among (C1), (C2), and (C3) mentioned in section 1 (note that a closed convex set is Clarke regular and subsmooth at all its points).

COROLLARY 4.1. *Let $a \in A$ and consider the following statements:*

(i) $\{A_1,\dots,A_n\}$ *is locally linearly regular at a.*

(ii) *There exist $\tau,\delta \in (0,\ +\infty)$ such that* (4.1) *holds.*

(iii) *There exist $\tau,\delta \in (0,\ +\infty)$ such that* (4.2) *holds.*

(iv) *There exist $\tau,\delta \in (0,\ +\infty)$ such that for any $u \in A \cap B(a,\delta)$,*

$$(*)\quad \inf\left\{\sum_{i=1}^n \|x_i^*\| : \sum_{i=1}^n x_i^* = x^* \text{ and } x_i^* \in N(A_i,u)\right\} \le \tau\|x^*\| \quad \text{for all } x^* \in N(A,u).$$

Then, the following statements hold:

(1) (i)⇒(ii) *always holds.*

(2) (iii)⇒(i) *holds if each A_i is subsmooth at a.*

(3) (i)–(iv) *are equivalent if each A_i is subsmooth and A is Clarke regular at all points of A close to a.*

(4) (i)⇔(ii) *holds if each A_i is subsmooth at a and X is an Asplund space.*

Proof. By Theorem 4.1 and Proposition 4.2, we need only show that (iii) and (iv) are equivalent under the assumption in (3). By the subsmoothness and regularity assumptions, there exists $\delta_0 \in (0,\ +\infty)$ such that

$$(4.14)\ \hat{N}(A,u) = N_c(A,u) \text{ and } N(A_i,u) = N_c(A_i,u) \quad \text{for all } u \in A \cap B(a,\delta_0).$$

Suppose that there exist $\tau > 0$ and $\delta \in (0,\ \delta_0)$ such that (4.2) holds. Let $u \in A \cap B(a,\delta)$ and $x^* \in N(A,u) \setminus \{0\}$. Then, by (4.2) and (4.14), there exists $x_i^* \in N(A_i,u) \cap B_{X^*}$ such that $\frac{x^*}{\|x^*\|} = \tau \sum_{i=1}^n x_i^*$. Letting $z_i^* = \tau\|x^*\|x_i^*$, it follows that $z_i^* \in N(A_i,u)$, $x^* = \sum_{i=1}^n z_i^*$, and $\sum_{i=1}^n \|z_i^*\| \le n\tau\|x^*\|$. Hence,

$$\inf\left\{\sum_{i=1}^n \|x_i^*\| : \sum_{i=1}^n x_i^* = x^* \text{ and } x_i^* \in N(A_i,u)\right\} \le n\tau\|x^*\|.$$

This shows that (iii)⇒(iv) holds.

Conversely, suppose that there exist $\tau > 0$ and $\delta \in (0,\ \delta_0)$ such that $(*)$ holds for any $u \in A \cap B(a,\delta)$. Let $u \in A \cap B(a,\delta)$ and $x^* \in N_c(A,u) \cap B_{X^*}$. By $(*)$ and (4.14), for every natural number k there exist $x_i^*(k) \in N_c(A_i,u)$ such that

$$(4.15)\qquad x^* = \sum_{i=1}^n x_i^*(k) \quad \text{and} \quad \sum_{i=1}^n \|x_i^*(k)\| \le \left(\tau + \frac{1}{k}\right)\|x^*\| \le \tau + \frac{1}{k}.$$

Since $(\tau+1)B_{X^*}$ is weakly* compact and each $N(A_i,u)$ is weakly* closed, we can assume that $x_i^*(k) \overset{w^*}{\to} x_i^* \in N(A_i,u)$ as $k \to \infty$ (passing to a generalized subsequence if necessary). It follows from (4.15) that $x^* = \sum_{i=1}^n x_i^*$ and $\sum_{i=1}^n \|x_i^*\| \le \tau$ (because the dual norm of X^* is lower semicontinuous with respect to the weak* topology). Therefore, $x_i^* \in \tau(N_c(A_i,u) \cap B_{X^*})$ and $x^* \in \tau \sum_{i=1}^n N_c(A_i,u) \cap B_{X^*}$. Hence, (4.2) holds. This shows that (iv)⇒(iii) holds. The proof is complete. □

Under the assumption in (3) of Corollary 4.1, (4.2) implies that

$$N(A,u) = \sum_{i=1}^n N(A_i,u) \tag{SC}$$

for all $u \in A$ close to a. In the case when each A_i is convex, (SC) and $(*)$ mean that $\{A_1,\dots,A_n\}$ has strong conical hull intersection property (strong CHIP) at u and $\{N(A_1,u),\dots,N(A_n,u)\}$ has property (G), respectively. These two properties have been well studied in convex analysis (see [2, 4, 5, 11] and references therein).

The following corollary is immediate from Propositions 2.2 and 4.3, Theorem 4.1, and Corollary 4.1.

COROLLARY 4.2. *Let X,Y be Banach spaces and $f_i : X \to Y$ be a continuously differentiable function ($i = 1,\dots,n$). Let C_i be a closed convex subset of Y and $A_i := f_i^{-1}(C_i)$ ($i = 1,\dots,n$). Let $a \in A$, and suppose that each $f_i'(a)$ is surjective. Then the following statements hold:*

(1) (i)⇒(ii) *holds.*

(2) (iii) *implies any one of* (i)–(iv).

(3) *If X is an Asplund space, then* (i)—(iv) *are equivalent.*

Where (i)–(iv) *are as in Corollary* 4.1.

Recently, Kruger [13, 14] studied a different kind of regularity of $\{A_1,\dots,A_n\}$ at a defined by $0 < \lim_{\rho\to 0^+} \frac{1}{\rho} \sup\{r \ge 0 : \bigcap_{i=1}^n (A_i - a_i) \cap (a + \rho B_X) \ne \emptyset, \text{ for all } a_i \in rB_X\}$.

Acknowledgment. The authors wish to thank the referees for careful reading of the paper and many valuable comments, which helped to improve our presentation.

REFERENCES

[1] D. Aussel, A. Daniilidis, and L. Thibault, *Subsmooth sets: Functional characterizations and related concepts*, Trans. Amer. Math. Soc., 357 (2005), pp. 1275–1301.

[2] A. Bakan, F. Deutsch, and W. Li, *Strong CHIP, normality, and linear regularity of convex sets*, Trans. Amer. Math. Soc., 357 (2005), pp. 3831–3863.

[3] H. H. Bauschke and J. M. Borwein, *On projection algorithms for solving convex feasibility problems*, SIAM Rev., 38 (1996), pp. 367–426.

[4] H. H. Bauschke, J. M. Borwein, and W. Li, *Strong conical hull intersection property, bounded linear regularity, Jameson's property (G), and error bounds in convex optimization*, Math. Program., 86 (1999), pp. 135–160.

[5] H. H. Bauschke, J. M. Borwein, and P. Tseng, *Bounded linear regularity, strong CHIP, and CHIP are distinct properties*, J. Convex Anal., 7 (2000), pp. 395–413.

[6] F. Bernard and L. Thilbaut, *Uniform prox-regularity of functions and epigraphs in Hilbert spaces*, Nonlinear Anal., 60 (2005), pp. 187–207.

[7] F. Bernard and L. Thilbaut, *Prox-regularity of functions and sets in Banach spaces*, Set-Valued Anal., 12 (2004), pp. 25–47.

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

[9] F. H. Clarke, R. Stern, and P. Wolenski, *Proximal smoothness and the lower-C^2 property*, J. Convex Anal., 2 (1995), pp. 117–144.

[10] F. Deutsch, W. Li, and J. Swetits, *Fenchel duality and the strong conical hull intersection property*, J. Optim. Theory Appl., 102 (1999), pp. 681–695.

[11] F. Deutsch, W. Li, and J. D. Ward, *Best approximation from the intersection of a closed convex set and a polyhedron in Hilbert space, weak Slater conditions, and the strong conical hull intersection property*, SIAM J. Optim., 10 (1999), pp. 252–268.
[12] G. Jameson, *The duality of pair of wedges*, Proc. Lond. Math. Soc., 24 (1972), pp. 531–547.
[13] A. Y. Kruger, *Stationary and regularity of set systems*, Pac. J. Optim., 1 (2005), pp. 101–126.
[14] A. Y. Kruger, *About regularity of collection of sets*, Set-Valued Anal., 14 (2006), pp. 187–206.
[15] A. S. Lewis and J. S. Pang, *Error bounds for convex inequality systems*, in Generalized Convexity, Generalized Monotonicity: Recent Results, Proceedings of the Fifth Symposium on Generalized Convexity, Luminy, 1996, J.-P. Crouzeix, J.-E. Martinez-Legaz, and M. Volle, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 75–10.
[16] C. Li and K. F. Ng, *Constraint qualification, the strong CHIP, and best approximation with convex constraints in Banach spaces*, SIAM J. Optim., 14 (2003), pp. 584–607.
[17] B. S. Mordukhovich, *Variational Analysis and Generalized Differentiation* I/II, Springer-Verlag, Berlin, Heidelberg, 2006.
[18] B. S. Mordukhovich and Y. Shao, *Nonsmooth sequential analysis in Asplund spaces*, Trans. Amer. Math. Soc., 348 (1996), pp. 1235–1280.
[19] K. F. Ng and W. H. Yang, *Error bounds for abstract linear inequality systems*, SIAM J. Optim., 13 (2002), pp. 24–43.
[20] K. F. Ng and W. H. Yang, *Regularities and relations to error bounds*, Math. Program., 99 (2004), pp. 521–538.
[21] J. S. Pang, *Error bounds in mathematical programming*, Math. Program., 79 (1997), pp. 299–332.
[22] R. R. Phelps, *Convex Functions, Monotone Operators and Differentiability*, Lecture Notes in Math. 1364, Springer, New York, 1989.
[23] R. Poliquin, R. T. Rockafellar, and L. Thibault, *Local differentiability of distance functions*, Trans. Amer. Math. Soc., 352 (2000), pp. 5231–5249.
[24] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[25] X. Y. Zheng and K. F. Ng, *Metric regularity and constraint qualifications for convex inequalities on Banach spaces*, SIAM J. Optim., 14 (2004), pp. 757–772.

SIAM J. OPTIM.
Vol. 19, No. 1, pp. 77–113

RELAXATION-BASED BOUNDS FOR SEMI-INFINITE PROGRAMS*

ALEXANDER MITSOS†, PANAYIOTIS LEMONIDIS†, CHA KUN LEE†, AND PAUL I. BARTON†‡

Abstract. Finite formulations are presented for the calculation of lower and upper bounds on the optimal solution value of semi-infinite programs (SIPs) involving smooth, potentially nonconvex objective function and constraints. The lower bounding problem is obtained by a formulation that combines the first- and second-order KKT necessary conditions of the lower-level problem with a discretization of the parameter set. The resulting mathematical program with equilibrium constraints (MPEC) is a relaxation of the original SIP and furnishes valid lower bounds. If the parameter set is subdivided, the optimal solution value of the lower bounding problem converges to the optimal solution value of the SIP. The upper bounding problem is based on convex and linear relaxations of the lower-level problem resulting in a restriction of the SIP. If the parameter set is subdivided, the constructed relaxations converge to the original lower-level program. The existence of SIP Slater points ensures convergence of the upper bounding problems to the optimal solution value of the SIP. Several alternatives for the upper bounding problem are presented and discussed. Numerical results are presented for a number of test problems from the literature.

Key words. SIP, MPEC, KKT, nonconvex, global optimization, convex relaxation, linearization

AMS subject classifications. 90C34, 65K05, 90C26, 90C33

DOI. 10.1137/060674685

1. Introduction.

Semi-infinite programs (SIPs) are optimization problems that involve a finite number of decision variables subject to an infinite number of constraints. We consider SIPs of the form

$$\begin{aligned} f^* = \min_{\mathbf{x}} \; & f(\mathbf{x}) \\ \text{s.t. } & g(\mathbf{x},\mathbf{p}) \leq 0 \quad \forall \mathbf{p} \in P \subset \mathbb{R}^{n_p}, \\ & \mathbf{x} \in X \subset \mathbb{R}^{n_x}, \end{aligned} \tag{1.1}$$

without any convexity/concavity assumptions and with $|P| \leq +\infty$. Regular inequality and equality constraints would not change anything significant in the proposed bounding methods and are only omitted for simplicity.

The infinite number of constraints of similar functional form that arise in SIP typically originate from design problems that impose either a constraint for any given point in time or for every point in a geometric region. SIPs are encountered in diverse scientific and engineering applications such as Chebyshev approximation, including the design of digital and FIR filters [24, 34], optimal control systems [38], neural networks [33], kinetic model reduction [11, 36], robust optimization [9], the design of water and air pollution control models [23], and the design of adaptive array processors [28].

Traditional algorithms for nonlinear SIPs can be categorized as either discretization or local-reduction methods [41]. In discretization approaches [14, 25, 37, 43, 49,

*Received by the editors November 10, 2006; accepted for publication (in revised form) August 20, 2007; published electronically February 6, 2008. This work is based upon work supported by the National Science Foundation under grant CTS-0521962.

http://www.siam.org/journals/siopt/19-1/67468.html

†Department of Chemical Engineering, Massachusetts Institute of Technology, 66-464, 77 Massachusetts Avenue, Cambridge, MA 02139 (amitsos@alum.mit.edu, plemonid@alum.mit.edu, chakun@alum.mit.edu, pib@mit.edu).

‡Corresponding author.

58] a sequence of subproblems is solved in which the objective function of (1.1) is minimized subject to a subset of the constraints indexed by the finite set $P_D \subset P$, which is generated from a uniform or adaptive grid on P. Under the assumptions that the resulting finite nonlinear programs (NLPs) are solved to global optimality, the discretization is exhaustive, and the cardinality of the set P_D tends to infinity, it can be shown that the accumulation points of discretization algorithms are global minima of the original SIP [14, 58]. In local reduction methods, a local solution of the SIP is also found by solving a sequence of finite NLPs. However, in contrast to discretization methods, the finite subset of constraints is not generated from a grid on the parameter set P, but rather from finding implicitly all the local maxima $\bar{\mathbf{p}}(\mathbf{x})$ of $g(\mathbf{x},\cdot)$ on P for each $\mathbf{x} \in X$. Under relatively strong assumptions on the problem structure, the set of local maxima is finite, and the SIP can, at least conceptually, be reduced to an equivalent finite problem. The accumulation points of local-reduction algorithms are local minima of the original SIP. Based on this reduction principle, a number of exact penalty and Lagrangian approaches have been suggested [16, 17, 39, 50]. Discretization and local-reduction-based methods generate approximations of local and global minima of nonconvex SIPs. On finite termination, these approximations are not guaranteed to be feasible points of the original SIP. For more thorough reviews of applications and algorithmic contributions in semi-infinite programming, the reader is referred to [23, 26, 40, 41, 55, 57].

An alternative to considering a finite subset of the constraints is to reformulate (1.1) as the following nonsmooth problem:

$$\begin{aligned} &\min_{\mathbf{x}\in X} f(\mathbf{x}) \\ &\text{s.t.} \quad g^U(\mathbf{x}) \equiv \max_{\mathbf{p}\in P} g(\mathbf{x},\mathbf{p}) \leq 0. \end{aligned} \tag{1.2}$$

The superscript U denotes the (exact) upper bound of $g(\mathbf{x},\cdot)$ on P. To determine feasibility of a candidate point $\bar{\mathbf{x}} \in X$, the *lower-level* or *inner problem* must be solved to global optimality,

$$g^U(\bar{\mathbf{x}}) = \max_{\mathbf{p}\in P} g(\bar{\mathbf{x}},\mathbf{p}). \tag{1.3}$$

Obviously, if $g^U(\bar{\mathbf{x}}) \leq 0$, then $\bar{\mathbf{x}}$ is feasible in (1.1); otherwise it is not. Recently, a method to generate guaranteed feasible points for SIP was proposed by Bhattacharjee and coworkers [12, 13]. Assuming that the host set X is compact and P is an interval, the functions f and $g(\cdot,\mathbf{p})$ are continuously differentiable on an open set containing X for all $\mathbf{p} \in P$, $g(\mathbf{x},\cdot)$ is continuous on P for all $\mathbf{x} \in X$, and there exist SIP Slater points arbitrarily close to all global minima, the method generates guaranteed feasible points and an ε-optimal estimate of the global solution of the SIP on finite termination. To generate these feasible points, interval arithmetic [32, 42] is used to construct an interval extension of the constraining function $g(\mathbf{x},\cdot)$ with respect to the parameter set P, i.e., an interval-valued function $G : X \to \mathbb{IR} : \mathbf{x} \mapsto [G^L(\mathbf{x}), G^U(\mathbf{x})]$ satisfying in particular

$$g^U(\mathbf{x}) \leq G^U(\mathbf{x}) \quad \forall \mathbf{x} \in X,$$

where $\mathbb{IR}$ is the set of all intervals in $\mathbb{R}$. For each $\mathbf{x} \in X$, $G^U(\mathbf{x})$ is a valid upper bound for $g^U(\mathbf{x})$ and thus a relaxation of the inner problem (1.3); see also section 3. Therefore, the following interval constrained reformulation (ICR) is a valid restriction

of (1.1):

$$\begin{aligned} &\min_{\mathbf{x}\in X} f(\mathbf{x}) \\ &\text{s.t.} \quad G^U(\mathbf{x}) \leq 0. \end{aligned}$$

Under subdivision of the parameter set P, it can be shown [12, 13] that $G^U(\mathbf{x})$ converges to $g^U(\mathbf{x})$, and therefore the ICR provides feasible points with objective function value arbitrarily close to the global solution value of the SIP.

In a work concurrent with this paper, Floudas and Stein [19] propose an alternative relaxation of (1.3) and thus a valid restriction of (1.1) furnishing feasible points. Their method relies on constructing a concave relaxation of the constraining function g on the parameter set P using the αBB method [4], replacing the resulting convex lower-level problem with its necessary and sufficient KKT conditions and solving the resulting mathematical program with equilibrium constraints (MPEC) with NCP functions.

In this paper we propose finite formulations that generate lower and upper bounds on the optimal objective value of (1.1). Generating upper and lower bounds is not only a necessary subproblem in algorithms such as the one by Bhattacharjee et al. [13] for the global solution of SIPs, but also a useful problem in its own right. Generating a feasible point and an associated upper bound to the optimal objective value of SIPs is, for instance, used in kinetic model reduction [36]. A lower bound can be used for a conservative estimate of the distance of the obtained upper bound from the optimal solution value.

For the lower bounding problem we use a combination of the first- and second-order KKT necessary conditions in conjunction with a discretization of the parameter set. For the upper bounding problem we extend the idea from [12, 13] of relaxing the lower-level problem (1.3) in order to restrict the outer problem and thus find guaranteed feasible points. For relaxation of the lower-level problem we employ convex and linear relaxation techniques. In section 2 we present definitions and assumptions needed in the subsequent sections. In section 3 we provide a brief introduction to interval methods and convex relaxation techniques and discuss restrictions and relaxations of SIPs. In section 4 we combine and extend ideas from the literature to create a KKT-based lower bounding problem that is formulated and solved as a mixed-integer nonlinear program (MINLP). In section 5 we introduce relaxation-based upper bounding problems that are formulated using either αBB or McCormick's techniques and solved using either a KKT-based (MPEC) or a linearization approach. In section 6 we provide numerical results for the proposed bounding formulations applied to literature examples and comment on their relative performance. Finally, in section 7 we provide conclusions and suggestions for future work in both semi-infinite and generalized semi-infinite programs (GSIP). GSIPs differ from SIPs in that the host set for $\mathbf{p}$ in the lower-level problem depends on $\mathbf{x}$; see, e.g., [47].

2. Definitions and assumptions. In this section we present the assumptions required for construction of the bounding problems proposed in what follows. Note that for the sake of simplicity we do not present the weakest assumptions possible.

ASSUMPTION 1 (host sets). *The host sets $X \subset \mathbb{R}^{n_x}$, $P \subset \mathbb{R}^{n_p}$ are Cartesian products of (compact) intervals; i.e., for all variables and parameters explicit bounds are known ($X = [\mathbf{x}^L, \mathbf{x}^U]$ and $P = [\mathbf{p}^L, \mathbf{p}^U]$).*

The set of vertices of P is denoted P_e. Based on Assumption 1,

$$P_e = \left\{ \mathbf{p} \in P : p_j \in \{p_j^L, p_j^U\},\ \forall j = 1, \ldots, n_p \right\},$$

and the cardinality of P_e is given by $|P_e| = 2^{n_p}$.

ASSUMPTION 2 (basic properties of functions). *The functions $f : X \to \mathbb{R}$ and $g : X \times P \to \mathbb{R}$ are twice continuously differentiable on some open set containing X and $X \times P$, respectively. Moreover, the constraint g is a factorable composite function* [30] *of univariate functions with known convex underestimating and overestimating functions.*

By $g_{p_j} : X \times P \to \mathbb{R}$ we denote the partial derivative of g with respect to p_j. Similarly by $g_{x_j} : X \times P \to \mathbb{R}$ we denote the partial derivative of g with respect to x_j. For nonsmooth relaxations we will make use of subgradients, as follows:

DEFINITION 2.1 (subgradient). *Let $Z \subset \mathbb{R}^{n_z}$ be a nonempty convex set and $h : Z \to \mathbb{R}$ be concave. A vector $\mathbf{d} \in Z$ is called a* subgradient *of h at $\bar{\mathbf{z}} \in Z$ if*

$$h(\mathbf{z}) \leq h(\bar{\mathbf{z}}) + \mathbf{d}^T(\mathbf{z} - \bar{\mathbf{z}}) \qquad \forall \mathbf{z} \in Z.$$

The definition for convex functions is analogous, with the direction of the inequality reversed. Existence of subgradients on the interior of Z is guaranteed, and for differentiable functions the unique subgradient is the gradient [10].

Noting that under Assumptions 1 and 2 the SIP (1.1) is equivalent to the nonsmooth problem (1.2), the following definition is motivated.

DEFINITION 2.2 (lower-level program). *For a fixed $\mathbf{x} \in X$ we denote*

$$\begin{aligned} &\max_{\mathbf{p}} g(\mathbf{x}, \mathbf{p}) \\ &s.t.\ \mathbf{p}^L \leq \mathbf{p} \leq \mathbf{p}^U \end{aligned} \tag{2.1}$$

the inner program *or* lower-level program *and $g^U(\mathbf{x})$ its optimal objective value.*

DEFINITION 2.3 (relaxation of functions). *Given a convex set $C \subset \mathbb{R}^{n_z}$ and a function $h : C \to \mathbb{R}$, a convex function $h^u : C \to \mathbb{R}$ is a* convex relaxation *(or* convex underestimator*) of h on C if*

$$h^u(\mathbf{z}) \leq h(\mathbf{z}) \quad \forall \mathbf{z} \in C,$$

and a concave function $h^o : C \to \mathbb{R}$ is a concave relaxation *(or* concave overestimator*) of h on C if*

$$h^o(\mathbf{z}) \geq h(\mathbf{z}) \quad \forall \mathbf{z} \in C.$$

The convex envelope $\bar{h}^u : C \to \mathbb{R}$ *of h on C is a convex relaxation of h on C such that for any convex relaxation h^u of h on C*

$$h^u(\mathbf{z}) \leq \bar{h}^u(\mathbf{z}) \quad \forall \mathbf{z} \in C.$$

Similarly, the concave envelope $\bar{h}^o : C \to \mathbb{R}$ *of h on C is a concave relaxation of h on C such that for any concave relaxation h^o of h on C*

$$\bar{h}^o(\mathbf{z}) \leq h^o(\mathbf{z}) \quad \forall \mathbf{z} \in C.$$

DEFINITION 2.4 (relaxation of programs). *Let $Z^D, Z^E \subset \mathbb{R}^{n_z}$, and consider the optimization problems*

$$\inf_{\mathbf{z} \in Z^D} f^D(\mathbf{z}) \quad \text{and} \quad \inf_{\mathbf{z} \in Z^E} f^E(\mathbf{z}).$$

If $Z^D \subset Z^E$ *and* $f^E(\mathbf{z}) \leq f^D(\mathbf{z})\ \forall \mathbf{z} \in Z^D$*, the optimization problem* $\inf_{\mathbf{z}\in Z^E} f^E(\mathbf{z})$ *is said to be a* relaxation *of* $\inf_{\mathbf{z}\in Z^D} f^D(\mathbf{z})$*. Similarly, the optimization problem* $\inf_{\mathbf{z}\in Z^D} f^D(\mathbf{z})$ *is said to be a* restriction *of* $\inf_{\mathbf{z}\in Z^E} f^E(\mathbf{z})$ [22].

A direct consequence of relaxations/restrictions is that for the programs in the above definition, $\inf_{\mathbf{z}\in Z^E} f^E(\mathbf{z}) \leq \inf_{\mathbf{z}\in Z^D} f^D(\mathbf{x})$. For maximization problems the above inequalities are reversed.

DEFINITION 2.5 (convex program). *The minimization problem* $\inf_{\mathbf{z}\in Z} f(\mathbf{z})$ *is called convex if* $Z \subset \mathbb{R}^{n_z}$ *is convex and* f *is convex on* Z*. Similarly the maximization problem* $\sup_{\mathbf{z}\in Z} f(\mathbf{z})$ *is called convex if* $Z \subset \mathbb{R}^{n_z}$ *is convex and* f *is concave on* Z*.*

DEFINITION 2.6 (convex relaxation of programs). *Let* $Z^D, Z^E \subset \mathbb{R}^{n_z}$*. The optimization problem* $\inf_{\mathbf{z}\in Z^E} f^E(\mathbf{z})$ *is a* convex relaxation *of* $\inf_{\mathbf{z}\in Z^D} f^D(\mathbf{z})$ *if it is a convex program and a relaxation of* $\inf_{\mathbf{z}\in Z^D} f^D(\mathbf{z})$*.*

DEFINITION 2.7 (diameter of a set). *Let* $Z \subset \mathbb{R}^{n_z}$*. The diameter of* Z*, denoted* $w(Z)$*, is the maximal distance between two points in* Z*:*

$$w(Z) = \sup_{\mathbf{z}_1,\mathbf{z}_2\in Z} ||\mathbf{z}_1 - \mathbf{z}_2||.$$

DEFINITION 2.8 (subdivision). *A* subdivision *of the set* P *is a finite collection of subsets* $P^i \subset P$ *with index set* I *such that*

$$P = \bigcup_{\tau\in I} P_\tau \qquad \textit{and} \qquad \operatorname{int}(P_{\tau_1}) \cap \operatorname{int}(P_{\tau_2}) = \emptyset \quad \forall \tau_1, \tau_2 \in I : \tau_1 \neq \tau_2.$$

A subdivision of P *with index set* I_2 *is a* refinement *of the subdivision with index set* I_1 *if for all* $\tau_2 \in I_2$ *there exists* $\tau_1 \in I_1$ *such that* $P_{\tau_2} \subset P_{\tau_1}$ *and for some* $\tau_2 \in I_2$ *there exists* $\tau_1 \in I_1$ *such that* $P_{\tau_2} \subset P_{\tau_1}$ *and* $P_{\tau_2} \neq P_{\tau_1}$*. A sequence of refined subdivisions with index sets* $I_1, I_2, \ldots, I_k$ *is called* exhaustive *if for* $k \to \infty$ *for all* $\tau_k \in I_k$ *the diameter of the set corresponding to* τ_k *vanishes,* $w(P_{\tau_k}) \to 0$*.*

Note that unlike partitions in a branch-and-bound procedure, in a subdivision no subsets P_τ of the host set P are fathomed.

3. Background. In both the lower and upper bounding problems, techniques from interval arithmetic and from convex, concave, and linear relaxations of mathematical programs are employed. Therefore, to aid in the understanding of the bounding procedures, we provide some brief background on these concepts.

3.1. Interval extensions. Consider a continuous function $h : Z \to \mathbb{R}$, where Z is an n_z-dimensional interval defined as

$$Z = [z_1^L, z_1^U] \times \cdots \times [z_{n_z}^L, z_{n_z}^U] = [\mathbf{z}^L, \mathbf{z}^U].$$

The image of Z under h is denoted by the scalar interval $\bar{h}(Z) = [h^L(Z), h^U(Z)]$. Consider also an interval-valued function $H : Z \to \mathbb{IR} : Y \mapsto [H^L(Y), H^U(Y)]$. H is an *inclusion function* for h over Z if the following relation holds:

$$[h^L(Y), h^U(Y)] = \bar{h}(Y) \subset H(Y) = [H^L(Y), H^U(Y)] \quad \forall Y \in \mathbb{IR}^{n_z} : Y \subset Z.$$

The natural interval extension is an example of such an inclusion function. It is derived by replacing each variable z_i with the corresponding interval $[z_i^L, z_i^U]$, decomposing the resulting expression into compositions of elementary operations (multiplication, addition, etc.) and intrinsic functions (exponential, exponentiation, monomial, logarithmic, etc.), and evaluating them using the rules of interval arithmetic [32, 42].

The tightness of inclusion functions can be quantified using the Hausdorff metric $q(\bar{h}(Z), H(Z))$, which is defined as

$$q(\bar{h}(Z), H(Z)) = \max(|h^L(Z) - H^L(Z)|, |h^U(Z) - H^U(Z)|).$$

An inclusion function is convergent with a convergence order β if constants $\gamma \geq 0$ and $\delta \geq 0$ exist such that

$$\begin{aligned} q(\bar{h}(Y), H(Y)) &\leq \gamma w(Y)^\beta, \\ w(H(Y)) &\leq \delta w(Y)^\beta, \end{aligned}$$

for each $Y \in \mathbb{IR}^{n_z} : Y \subset Z$. Note that in general the constants γ, δ depend on the function h and the host set Z.

If subdivision of Z is employed and convergent inclusion functions are taken for each of the resulting subintervals, a tighter estimate of the image of Z under h is obtained. Assuming that the subdivision of Z is given by a collection of nondegenerate intervals Z^k for some finite m, i.e., $Z = \bigcup_{k=1}^m Z^k$, the range of the inclusion is defined as

$$H_m(Z) = \left[\min_k H^L(Z^k), \max_k H^U(Z^k)\right] = [H_m^L(Z), H_m^U(Z)].$$

It can be shown that the following relationship holds:

$$\bar{h}(Z) \subset H_m(Z) \subset H(Z) \Leftrightarrow [h^L(Z), h^U(Z)] \subset [H_m^L(Z), H_m^U(Z)] \subset [H^L(Z), H^U(Z)].$$

Finally, if the subdivision of Z is exhaustive, then the bounds from the inclusion functions converge to the bounds of the true range of the function

$$\lim_{m\to\infty} H_m^L(Z) = h^L(Z) \quad \text{and} \quad \lim_{m\to\infty} H_m^U(Z) = h^U(Z).$$

Natural interval extensions have a first-order convergence rate, while Taylor models (standard or optimally centered forms) have a second-order convergence rate [6] but are typically more expensive to evaluate. For a thorough discussion, the reader is referred to the literature, e.g., [6, 32]. Furthermore, it should be noted that interval methods can be automated; see, e.g., [53, 54].

3.2. Convex relaxation. Most deterministic global optimization algorithms, such as spatial branch-and-bound and outer approximation, rely on the construction of convex relaxations. Given a box-constrained NLP

$$\begin{aligned} &\max_{\mathbf{z}} h(\mathbf{z}) \\ &\text{s.t. } \mathbf{z} \in [\mathbf{z}^L, \mathbf{z}^U] \equiv Z \subset \mathbb{R}^{n_z} \end{aligned} \tag{3.1}$$

with a nonconcave objective function $h(\mathbf{z})$, the goal is to construct a convex maximization problem, i.e., a program with convex constraints and a concave objective function, whose optimal objective value overestimates the optimal solution value of (3.1). Convex and concave envelopes or tight relaxations are known for a variety of simple nonlinear terms [46, 3, 52], and this allows the construction of convex and concave relaxations for a quite general class of functions through several methods [30, 4, 46, 21]. All the methods proposed in the literature essentially rely on a few key ideas and elements. McCormick's results [30] allow the construction of convex

and concave relaxations of functions defined by recursive compositions of elementary operations and intrinsic functions. Floudas and coworkers [2, 3, 5] have proposed convex relaxations for arbitrary twice continuously differentiable functions by the addition of a simple, sufficiently negative function that is known to be convex; concave relaxations are handled similarly. Both approaches can also introduce auxiliary variables and constraints. Smith and Pantelides [46] formalized the use of auxiliary variables, while Gatzke, Tolsma, and Barton [21] demonstrated how these methods can be combined and automated. Tawarmalani and Sahinidis [52, 51] proposed to further relax the convex relaxations via linearization to take advantage of the scalability and reliability of linear programming (LP) solvers. Wang and Chang [56] proposed piecewise-affine relaxations based on linearization. While many combinations of the above ideas are conceivable, we consider three extreme cases of convex relaxation that are of particular interest for the relaxation of the lower-level program.

3.2.1. Nonsmooth concave overestimation. The first alternative we consider is to construct a concave relaxation of the objective function in (3.1) by successively applying McCormick's composition theorem, without the introduction of auxiliary variables and constraints. McCormick [30] presents convex and concave relaxations of a function

$$h^t(\mathbf{z}) = T(t(\mathbf{z})) + U(u(\mathbf{z}))V(v(\mathbf{z})),$$

where $T, U, V : \mathbb{R} \to \mathbb{R}$ are continuous and $t, u, v : Z \to \mathbb{R}$ are continuous on Z. Assuming that convex and concave relaxations are known for all functions (t, u, v and T, U, V), and bounds are known for the ranges of the inner functions (t, u, v), McCormick's composition result provides convex and concave relaxations for h^t on Z.

Consider, for instance, $T(t(\mathbf{z}))$ and let t^u and t^o be convex and concave relaxations of t on Z. Let further $t(\mathbf{z}) \in [t^L, t^U]$ for all $\mathbf{z} \in Z$, and T^u, T^o be convex and concave relaxations of T on $[t^L, t^U]$. Let, finally, $w_{min} \in \arg\min_{w \in [t^L, t^U]} T^u(w)$ and $w_{max} \in \arg\max_{w \in [t^L, t^U]} T^o(w)$. Then

$$T^u\left(\text{mid}\{t^u(\mathbf{z}), t^o(\mathbf{z}), w_{min}\}\right),$$
$$T^o\left(\text{mid}\{t^u(\mathbf{z}), t^o(\mathbf{z}), w_{max}\}\right)$$

are respectively a convex and concave relaxation of the composite function $T(t)$ on Z. For the product of two functions see Appendix B.

Recursive application of McCormick's result allows the derivation of convex and concave relaxations for complicated expressions termed *factorable* expressions. Assuming that the objective function h in (3.1) is factorable, we denote $h^{o,mc} : Z \to \mathbb{R}$ the concave relaxation constructed by the recursive application of the composition theorem. Since $h^{o,mc}(\mathbf{z}) \geq h(\mathbf{z})$ for all $\mathbf{z} \in Z$, the optimal objective value of

$$\max_{\mathbf{z} \in Z} h^{o,mc}(\mathbf{z}) \tag{3.2}$$

overestimates the optimal objective value of (3.1). While convex, (3.2) is not necessarily smooth, and therefore standard optimization techniques relying on the satisfaction of KKT conditions are not applicable in general. Since it is box-constrained, the linearization (using subgradients) at an arbitrary interior point $\bar{\mathbf{z}} \in \text{int}(Z)$ results in a linear program which is a further relaxation (and trivially smooth),

$$\max_{\mathbf{z} \in Z} h^{o,mc,lin}(\mathbf{z}),$$

where $h^{o,mc,lin}(\mathbf{z}) = h(\bar{\mathbf{z}}) + \mathbf{d}^{\mathrm{T}}(\mathbf{z} - \bar{\mathbf{z}})$ for some subgradient $\mathbf{d}$. While the existence of subgradients is guaranteed, obtaining them within the McCormick relaxation requires some development. The reader is referred to [8] for this development. Note also that a subgradient is guaranteed to exist even at boundary points under mild assumptions, e.g., differentiability of the univariate intrinsic functions on their domains.

3.2.2. Smooth concave overestimation without auxiliary variables. The second alternative we consider is based on the ideas of αBB relaxation by Adjiman and coworkers [4, 3, 1] and γBB relaxation by Akrotirianakis and Floudas [5]. To avoid the introduction of auxiliary variables and constraints, which add complications, we deviate from the framework presented in these references. Instead of splitting the nonlinear objective h into the sum of concave terms, special nonconcave terms, and general nonconcave terms, we apply the relaxation to the original function. Note also that we consider the simplest variant of uniform diagonal shift of the Hessian matrix.

Since univariate quadratic terms are convex,

$$h^{o,\alpha}(\mathbf{z}) = h(\mathbf{z}) + \alpha \sum_{i=1}^{n_z} (z_i - z_i^L)(z_i^U - z_i)$$

is concave for sufficiently large values of α. Moreover, for any $\mathbf{z} \in Z$,

$$h^{o,\alpha}(\mathbf{z}) \geq h(\mathbf{z}) \qquad \forall \alpha \geq 0.$$

The smallest possible value for α is obtained by finding the largest eigenvalue of the Hessian matrix on Z, i.e., by the global solution of a nonconvex optimization problem. Instead, Adjiman et al. [3] have proposed efficient methods for overestimating α. One such method is the application of Gerschgorin's theorem and estimating

$$\frac{1}{2} \max_{\mathbf{z} \in Z} \max_i \max \left\{ 0, H_{ii}(\mathbf{z}) + \sum_{j \neq i} |H_{ij}(\mathbf{z})| \right\}$$

using interval arithmetic on the Hessian matrix. Note that $H_{ij} = \frac{\partial^2 h}{\partial z_i \partial z_j}$.

Since $h^{o,\alpha}(\mathbf{z}) \geq h(\mathbf{z})$ for all $\mathbf{z} \in Z$ and all $\alpha \geq 0$, the optimal objective value of

$$\max_{\mathbf{z} \in Z} h^{o,\alpha}(\mathbf{z}) \tag{3.3}$$

overestimates the optimal objective value of (3.1). The formulated relaxation (3.3) is a box-constrained maximization problem with a smooth concave objective function. The polyhedral feasible set along with the concavity of the objective function make the first-order KKT conditions necessary and sufficient for a global maximum. Standard, gradient-based optimization algorithms can reliably solve (3.3). Finally, since (3.3) is box-constrained, the linearization at an arbitrary point $\bar{\mathbf{z}} \in Z$ results in a linear program which is a further relaxation.

The application of γBB relaxation [5] is analogous. In this method, relaxation is achieved by the addition of exponential terms,

$$h^{o,\gamma}(\mathbf{z}) = h(\mathbf{z}) + \sum_{i=1}^{n_z} \left(1 - e^{\gamma_i (z_i - z_i^L)}\right) \left(1 - e^{\gamma_i (z_i^U - z_i)}\right).$$

3.2.3. Smooth concave overestimation with auxiliary variables. The third alternative we consider is the introduction of auxiliary variables $\mathbf{w}$ and constraints as described in [21]. First, a factorable representation of the nonconcave function h is developed, introducing a new variable w_i for each distinct factor. Subsequently, the bounds for the auxiliary variables $\mathbf{w}$ are propagated via natural interval extensions from the bounds on $\mathbf{z}$ and the auxiliary variables already introduced. At the next step an equivalent equality constrained program is generated by introducing the definition of each factor as an equality constraint and replacing each occurrence of a nonconvex function with the relevant factor. Then, each nonlinear equality constraint is rewritten as a pair of inequalities. Finally, the inequalities are relaxed by relaxing each nonlinear expression; if the (smooth) convex and concave envelopes (or tight relaxations) of a nonlinear expression are known, these are introduced; otherwise convex and concave relaxations are computed by the αBB or γBB method. Nonsmoothness in an envelope can be represented by multiple smooth convex inequalities (e.g., the bilinear case). Sums of linear terms are also replaced by new variables w_i along with a linear equality constraint.

In the special case that the objective function contains additive univariate convex terms, these terms can be directly overestimated by the secant without auxiliary variables. Similarly, additive concave terms in the objective are left unchanged. The resulting program is

$$
\begin{aligned}
&\max_{\mathbf{z},\mathbf{w}} h^{o,ex}(\mathbf{z},\mathbf{w}) \\
&\text{s.t. } t_i^u(\mathbf{z},\mathbf{w}) - w_j \le 0, \quad i \in I_j^u, \quad j = 1,\dots,n_w, \\
&\qquad w_j - t_i^o(\mathbf{z},\mathbf{w}) \le 0, \quad i \in I_j^o, \quad j = 1,\dots,n_w, \\
&\qquad\qquad \mathbf{t}^l(\mathbf{z},\mathbf{w}) = \mathbf{0}, \\
&\qquad\qquad\qquad \mathbf{z} \in Z, \\
&\qquad\qquad\qquad \mathbf{w} \in [\mathbf{w}^L, \mathbf{w}^U] \subset \mathbb{R}^{n_w},
\end{aligned}
\tag{3.4}
$$

where $\mathbf{t}^l$, $\mathbf{t}^u$, and $\mathbf{t}^o$ denote affine, convex, and concave functions, respectively, and the objective function $h^{o,ex}$ is concave. The (possibly empty) finite index sets I_j^u and I_j^o represent the multiple smooth convex inequalities. By construction, the optimal solution value of (3.4) overestimates the optimal solution value of (3.1). It is a convex program with linear equality constraints and differentiable convex inequality constraints. Due to convexity, the KKT conditions are sufficient for a global minimum, and we employ this for the upper bounding procedure. The number of auxiliary variables and constraints introduced depends on the problem size and on the problem structure. Since it is bounded by a multiple of the number of factors in the McCormick factorization, it is typically a small multiple of the number of variables.

The existence of a Slater point provides a constraint qualification [10, p. 325], and in this case the first-order KKT conditions are also necessary for a local and global minimum. While typically the existence of a Slater point is expected, to the best of our knowledge it has not been proved in general for this type of convex relaxations.

Note that since the procedure described here is analogous to the procedure used when constructing natural interval extensions, which in turn are used to calculate bounds for the auxiliary variables, the relaxation provided by (3.4) is expected to be at least as tight as the natural interval extension of h over Z. Moreover, by the introduction of auxiliary variables the relaxations can furnish tighter relaxations than the ones furnished by McCormick's composition theorem without auxiliary variables

[51, p. 128].

A further relaxation of (3.4) can be performed via linearization of the objective function and the constraints [51]. A weaker linear relaxation can be obtained by removing all nonlinear constraints. Finally, an even weaker linear relaxation is generated by removing all constraints but the variable bounds to obtain a box-constrained program.

3.3. Restrictions/relaxations of SIP. Recall that SIPs can be interpreted as the nonsmooth program (1.2). A restriction of the lower-level program (1.3), pointwise in $\mathbf{x}$, leads to an underestimation of the optimal solution value $g^U(\mathbf{x})$. Therefore, the constraint $g^U(\mathbf{x}) \leq 0$ is relaxed. Similarly, a relaxation of the lower-level program (1.3), pointwise in $\mathbf{x}$, leads to an overestimation of the optimal solution value $g^U(\mathbf{x})$, and the constraint $g^U(\mathbf{x}) \leq 0$ is restricted. In other words, a restriction of the lower-level program results in a relaxation of the SIP, and similarly a relaxation of the lower-level program results in a restriction of the SIP.

To be of practical significance, the restrictions and relaxations of the lower-level program have to be valid for the entire set X. Discretization methods replace the host set P with a finite set $P_D \subset P$, uniformly in $\mathbf{x}$. This is a restriction of the lower-level program, since its feasible set is replaced by a subset and results in a relaxation of the SIP. If the resulting NLP is solved to global optimality, a lower bound to the SIP is obtained. On the other hand, the ICR proposal of Bhattacharjee and coworkers [12, 13] does not alter the feasible set of the lower-level program, but overestimates the objective function of the lower-level program pointwise in $\mathbf{x}$. This overestimation leads to a relaxation of the lower-level program and a restriction of the SIP. Any feasible point of the resulting NLP gives an upper bound to the SIP. Similarly, the work by Floudas and Stein [19] and our proposed upper bounds are based on a convex relaxation of the lower-level program resulting in a restriction of the SIP.

4. KKT-based lower bound. In single-level optimization lower bounds are typically obtained by the solution of a convex relaxation. As discussed in the introduction, a well established relaxation of SIPs is obtained by replacing the infinite set P with a finite subset P_D. Stein and Still [48] solve GSIP with a convex lower-level program satisfying a constraint qualification via an equivalent representation as a MPEC and remark that under nonconvexity this approach would give a lower bound. Here we combine these two ideas. We also propose a simple method for calculating bounds on the KKT multipliers which, depending on the solution method employed for the lower bounding problem, are either helpful or required [31]. Finally, we propose a simple method of employing (partially) the second-order KKT necessary conditions. The formulated lower bounding problem is computationally more expensive than either of the two known ideas. However, the lower bound furnished is at least as tight as either of the existing ones, since the solution obtained simultaneously satisfies both conditions. Moreover, numerical examples in section 6 show that the lower bound proposed may even be tighter than either of the two existing ideas.

A valid relaxation of the constraint "$\mathbf{p}$ is a global maximum of the lower-level program" is the constraint "$\mathbf{p}$ is a local maximum of the lower-level program." By Assumption 2, for each $\mathbf{x} \in X$ the function $g(\mathbf{x}, \cdot)$ is differentiable on some open set containing P, and therefore by linearity of the constraints in the lower-level program (2.1) the KKT conditions are necessary for a local maximum. The solution value of the following MPEC, for which the parameters $\mathbf{p}$ and the KKT multipliers $\boldsymbol{\mu}$ have been added to the set of variables, provides a valid lower bound for (1.1):

$$
\begin{aligned}
f^{LBD} = \min_{\mathbf{x},\mathbf{p},\boldsymbol{\mu}} \; & f(\mathbf{x}) \\
\text{s.t.} \; & -g_{p_j}(\mathbf{x},\mathbf{p}) + \mu_j - \mu_{n_p+j} = 0, \quad j = 1,\dots,n_p, \\
& \mu_j(p_j - p_j^U) = 0, \quad j = 1,\dots,n_p, \\
& \mu_{n_p+j}(-p_j + p_j^L) = 0, \quad j = 1,\dots,n_p, \\
& g(\mathbf{x},\mathbf{p}) \le 0, \\
& g(\mathbf{x},\hat{\mathbf{p}}) \le 0 \quad \forall \hat{\mathbf{p}} \in P_D, \\
& 0 \le \mu_j \le \mu_j^{max}, \quad j = 1,\dots,2n_p, \\
& \mathbf{x} \in X, \qquad \mathbf{p} \in P.
\end{aligned} \tag{4.1}
$$

Recall that g_{p_j} denotes the partial derivative of g with respect to p_j. To obtain a valid lower bound for (1.1) the above MPEC must be solved to global optimality. Another alternative is to further relax (4.1).

Recall that replacing the infinite set P with the finite subset P_D is a restriction of the lower-level program, and therefore a relaxation of the SIP. On the other hand, the addition of the parameters and KKT multipliers to the variable list is a relaxation of the optimality constraint in the lower-level program. The necessity of the KKT conditions ensures that the introduced constraints are feasible. The promise of this relaxation is that when P is subdivided and the subdivisions are successively refined in an exhaustive manner, for some τ, the only KKT points in P_τ will be global maxima of the lower-level program for the obtained $\bar{\mathbf{x}}$; see also the subsection devoted to convergence.

The multiplier bounds $\boldsymbol{\mu}^{max}$ need not be exact. Finite bounds are needed for some of the solution methods employed for (4.1) as well as the reformulation via integer programming used in this paper, and tighter bounds will typically accelerate the solution of (4.1). However, if the bounds on the KKT multipliers are underestimated, the lower bounding problem may be invalid (by excluding some feasible $\mathbf{x}$), irrespective of the algorithm used to solve (4.1).

4.1. Bounds on the KKT multipliers. In the following we discuss how to obtain the bounds $\boldsymbol{\mu}^{max}$ for the KKT multipliers $\boldsymbol{\mu}$ in the lower bounding problem (4.1).

PROPOSITION 4.1 (multiplier bounds). *Valid upper bounds for the KKT multipliers $\boldsymbol{\mu}$ in* (4.1) *are given by*

$$
\mu_j^{max} = \max_{\mathbf{x}\in X, \mathbf{p}\in P, p_j = p_j^U} g_{p_j}(\mathbf{x},\mathbf{p}), \quad j = 1,\dots,n_p, \tag{4.2}
$$

$$
\mu_{n_p+j}^{max} = -\min_{\mathbf{x}\in X, \mathbf{p}\in P, p_j = p_j^L} g_{p_j}(\mathbf{x},\mathbf{p}), \quad j = 1,\dots,n_p. \tag{4.3}
$$

Proof. Consider an arbitrary but fixed index $j \in \{1,2,\dots,n_p\}$. The KKT multipliers $\mu_j \ge 0$ and $\mu_{n_p+j} \ge 0$ must satisfy the stationarity and complementary slackness conditions

$$
-g_{p_j}(\mathbf{x},\mathbf{p}) + \mu_j - \mu_{n_p+j} = 0, \tag{4.4}
$$

$$
\mu_j(p_j - p_j^U) = 0, \tag{4.5}
$$

$$
\mu_{n_p+j}(-p_j + p_j^L) = 0. \tag{4.6}
$$

Note first that for $p_j \in (p_j^L, p_j^U)$ we obtain $\mu_j = 0$ from (4.5) and $\mu_{n_p+j} = 0$ from (4.6). In the following we therefore consider only $p_j \in \{p_j^L, p_j^U\}$.

For $p_j = p_j^U$ we obtain $\mu_{n_p+j} = 0$ from (4.6), and since any μ_j satisfies (4.5), μ_j is calculated by (4.4) as

$$\mu_j = g_{p_j}(\mathbf{x}, \mathbf{p}).$$

Therefore, the largest value that μ_j can take is given by the maximum of the partial derivative g_{p_j} over X and P, constrained by $p_j = p_j^U$,

$$\mu_j^{max} = \max_{\mathbf{x}\in X, \mathbf{p}\in P, p_j = p_j^U} g_{p_j}(\mathbf{x}, \mathbf{p}).$$

Similarly for $p_j = p_j^L$ we obtain $\mu_j = 0$ from (4.5), and since any μ_{n_p+j} satisfies (4.6), μ_{n_p+j} is calculated by (4.4) as

$$\mu_{n_p+j} = -g_{p_j}(\mathbf{x}, \mathbf{p}).$$

Therefore the largest value that μ_{n_p+j} can take is given by the maximum of the negative partial derivative $-g_{p_j}$ over X and P such that $p_j = p_j^L$,

$$\mu_{n_p+j}^{max} = \max_{\mathbf{x}\in X, \mathbf{p}\in P, p_j = p_j^L} (-g_{p_j}(\mathbf{x}, \mathbf{p})) = -\min_{\mathbf{x}\in X, \mathbf{p}\in P, p_j = p_j^L} g_{p_j}(\mathbf{x}, \mathbf{p}). \qquad \square$$

Note that in the optimization problems (4.2), (4.3), p_j is fixed and is used only as a dummy variable. In general, solving the above optimization programs is expensive, and we therefore propose to overestimate the bounds by interval extensions over $X \times P$ (with p_j fixed). When nonpositive values are obtained for the bounds, i.e., when the partial derivative g_{p_j} is nonpositive for $p_j = p_j^U$ or nonnegative for $p_j = p_j^L$, the corresponding variable and the complementary slackness conditions can be eliminated. In a branch-and-bound procedure such as the one proposed by Bhattacharjee et al. [13], where the host set of X is branched and the host set P subdivided, this elimination criterion is likely to be valid in some subsets of X and P.

4.2. Implementation with integer variables. Due to the complementary slackness conditions, (4.1) violates the Mangasarian–Fromowitz constraint qualification (MFCQ) [48] and is difficult to solve [7]. Stein and Still [48] solve their MPEC by regularizing the complementary slackness conditions via NCP functions and solving a sequence of regularized NLPs. Each of these NLPs is an approximation to (4.1), but it can be easily verified that it provides a valid lower bound. In the limit (4.1) is solved. General-purpose global NLP solvers such as BARON [44] obtain lower bounds on (4.1) by constructing convex relaxations of the equality constraints as pairs of inequalities. To ensure convergence, bounds on the KKT multipliers are needed. Fortuny-Amat and McCarl [20] reformulate the complementary slackness conditions of the inner program in a class of bilevel programs using integer variables and the big-M formulation to obtain an MINLP. To do so bounds are needed for the multipliers, as well as for the constraints. In the special case of SIP, since the lower-level problem contains only box constraints, bounds on the constraints are readily available, and the big-M reformulation is particularly simple. Moreover, due to the simple structure of the lower-level program the KKT multipliers can be eliminated. Let y_j be an integer variable used to indicate whether $p_j = p_j^U$; i.e., let $y_j = 0$ imply $p_j < p_j^U$ and $y_j = 1$ imply $p_j = p_j^U$. Similarly, let y_{n_p+j} be an integer variable used to indicate whether $p_j = p_j^L$; i.e., let $y_{n_p+j} = 0$ imply $p_j > p_j^L$ and $y_{n_p+j} = 1$ imply $p_j = p_j^L$. We claim

that MPEC (4.1) is equivalent to the following MINLP:

$$
\begin{aligned}
f^{LBD} = \min_{\mathbf{x},\mathbf{p},\mathbf{y}} & f(\mathbf{x}) \\
\text{s.t. } & g_{p_j}(\mathbf{x},\mathbf{p}) \le y_j \mu_j^{max}, && j=1,\dots,n_p, \\
& p_j^U - p_j \le \left(p_j^U - p_j^L\right)(1-y_j), && j=1,\dots,n_p, \\
& -g_{p_j}(\mathbf{x},\mathbf{p}) \le y_{n_p+j}\mu_{n_p+j}^{max}, && j=1,\dots,n_p, \\
& p_j - p_j^L \le \left(p_j^U - p_j^L\right)\left(1-y_{n_p+j}\right), && j=1,\dots,n_p, \\
& y_j + y_{n_p+j} \le 1, && j=1,\dots,n_p, \\
& g(\mathbf{x},\mathbf{p}) \le 0 && \\
& g(\mathbf{x},\hat{\mathbf{p}}) \le 0 && \forall \hat{\mathbf{p}} \in P_D, \\
& \mathbf{y} \in \{0,1\}^{2n_p}, && \\
& \mathbf{x} \in X, \qquad \mathbf{p} \in P, &&
\end{aligned}
\tag{4.7}
$$

where $\boldsymbol{\mu}^{max}$ are calculated by (4.2), (4.3). The reason we employ this reformulation is because the use of binary variables allows more flexibility, such as the introduction of second-order KKT conditions.

Note that if (4.2) furnishes $\mu_j^{max} \le 0$, then y_j is eliminated, and the two inequalities containing y_j are replaced by $g_{p_j}(\mathbf{x},\mathbf{p}) \le 0$. Similarly, if (4.3) gives $\mu_{n_p+j}^{max} \le 0$, then y_{n_p+j} is eliminated, and the two inequalities containing y_{n_p+j} are replaced by $g_{p_j}(\mathbf{x},\mathbf{p}) \ge 0$. It is very simple to verify the validity of these eliminations.

PROPOSITION 4.2 (validity of MPEC formulation). *If values for $\boldsymbol{\mu}^{max}$ are calculated by* (4.2), (4.3)*, then MPEC* (4.1) *is equivalent to* (4.7).

Proof. The two problems have the same objective function, which depends only on the original variables $\mathbf{x}$, and therefore it suffices to show that the projection of the feasible sets on X is the same. Note also that both formulations contain the constraints from the discretization of P, namely $g(\mathbf{x},\hat{\mathbf{p}}) \le 0$, for all $\hat{\mathbf{p}} \in P_D$, and we can ignore these.

1. $\bar{\mathbf{x}}$ *is feasible in* (4.1) $\Rightarrow$ $\bar{\mathbf{x}}$ *is feasible in* (4.7).
   Let $\bar{\mathbf{x}}$ be feasible in (4.1). Therefore, there exist $\bar{\boldsymbol{\mu}}$, $\bar{\mathbf{p}}$ such that $(\bar{\mathbf{x}},\bar{\mathbf{p}},\bar{\boldsymbol{\mu}})$ satisfy the constraints of (4.1). We will show that there also exists $\bar{\mathbf{y}}$ such that $(\bar{\mathbf{x}},\bar{\mathbf{p}},\bar{\mathbf{y}})$ is feasible in (4.7). For an arbitrary but fixed $j \in \{1,2,\dots,n_p\}$ consider the following cases:
   (a) $\bar{p}_j \in (p_j^L, p_j^U)$. From the complementary slackness conditions we obtain $\bar{\mu}_j = \bar{\mu}_{n_p+j} = 0$, and therefore from the stationarity condition $g_{p_j}(\bar{\mathbf{x}},\bar{\mathbf{p}}) = 0$. Pick now $\bar{y}_j = \bar{y}_{n_p+j} = 0$. The first five constraints of (4.7) become
   $$
   \begin{aligned}
   & g_{p_j}(\bar{\mathbf{x}},\bar{\mathbf{p}}) \le 0, \\
   & p_j^U - \bar{p}_j \le \left(p_j^U - p_j^L\right), \\
   & g_{p_j}(\bar{\mathbf{x}},\bar{\mathbf{p}}) \ge 0, \\
   & \bar{p}_j - p_j^L \le \left(p_j^U - p_j^L\right), \\
   & 0 + 0 \le 1,
   \end{aligned}
   $$
   and are clearly satisfied.
   (b) $\bar{p}_j = p_j^U$. From the complementary slackness condition we obtain $\bar{\mu}_{n_p+j} = 0$, and therefore from the stationarity condition $g_{p_j}(\bar{\mathbf{x}},\bar{\mathbf{p}}) \ge 0$. Pick now

$\bar{y}_j = 1$, $\bar{y}_{n_p+j} = 0$. The first five constraints of (4.7) become

$$\begin{aligned} g_{p_j}(\bar{\mathbf{x}}, \bar{\mathbf{p}}) &\leq \mu_j^{max}, \\ 0 &\leq 0, \\ g_{p_j}(\bar{\mathbf{x}}, \bar{\mathbf{p}}) &\geq 0, \\ p_j^U - p_j^L &\leq \left(p_j^U - p_j^L\right), \\ 1 + 0 &\leq 1, \end{aligned}$$

and are satisfied; recall also the calculation of μ_j^{max} from (4.2).

(c) $\bar{p}_j = p_j^L$. This case is analogous to the previous case.

Since we also have $g(\bar{\mathbf{x}}, \bar{\mathbf{p}}) \leq 0$, the desired result is shown.

2. $\bar{\mathbf{x}}$ *is feasible in* (4.7) $\Rightarrow$ $\bar{\mathbf{x}}$ *is feasible in* (4.1) . Let $\bar{\mathbf{x}}$ be feasible in (4.7). Therefore, there exist $\bar{\mathbf{y}}$, $\bar{\mathbf{p}}$ such that $(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\mathbf{y}})$ satisfy the constraints of (4.7). We will show that there also exists $\bar{\boldsymbol{\mu}}$, such that $(\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\boldsymbol{\mu}})$ is feasible in (4.1). For an arbitrary but fixed $j \in \{1, 2, \ldots, n_p\}$ consider the following cases:

(a) $\bar{p}_j \in (p_j^L, p_j^U)$. The constraint $p_j^U - \bar{p}_j \leq \left(p_j^U - p_j^L\right)(1 - \bar{y}_j)$ directly gives $\bar{y}_j = 0$, and therefore we obtain $g_{p_j}(\bar{\mathbf{x}}, \bar{\mathbf{p}}) \leq 0$. Similarly we obtain $\bar{y}_{n_p+j} = 0$ and $g_{p_j}(\bar{\mathbf{x}}, \bar{\mathbf{p}}) \geq 0$. We therefore have $g_{p_j}(\bar{\mathbf{x}}, \bar{\mathbf{p}}) = 0$. Pick $\bar{\mu}_j = \bar{\mu}_{n_p+j} = 0$ and note that the first three constraints of (4.1) are satisfied.

(b) $\bar{p}_j = p_j^U$. Similarly to the previous case we obtain $\bar{y}_{n_p+j} = 0$ and $g_{p_j}(\bar{\mathbf{x}}, \bar{\mathbf{p}}) \geq 0$. Pick $\bar{\mu}_j = g_{p_j}(\bar{\mathbf{x}}, \bar{\mathbf{p}})$ and $\bar{\mu}_{n_p+j} = 0$ and note that the first three constraints of (4.1) are satisfied.

(c) $\bar{p}_j = p_j^L$. This case is analogous to the previous case.

Since we also have $g(\bar{\mathbf{x}}, \bar{\mathbf{p}}) \leq 0$, the desired result is shown. ∎

4.3. Second-order conditions. Typically, some infeasible points $\bar{\mathbf{x}}$ have a lower objective value than the optimal objective value $f(\bar{\mathbf{x}}) < f^*$, for otherwise (1.1) is essentially unconstrained. Moreover, while for these points $\max_{\mathbf{p}\in P} g(\bar{\mathbf{x}}, \mathbf{p}) > 0$, it is possible that $\min_{\mathbf{p}\in P} g(\bar{\mathbf{x}}, \mathbf{p}) \leq 0$. Therefore, since the parameters are added to the variable list, the solution of (4.1) tends to furnish such points $\bar{\mathbf{x}}$ along with points $\bar{\mathbf{p}}$ which are stationary points (e.g., unconstrained minima) of the lower-level program, because for the given $\bar{\mathbf{x}}$ these are the least restrictive for the constraint $g(\bar{\mathbf{x}}, \mathbf{p}) \leq 0$. The second-order KKT conditions can partially alleviate this phenomenon. The second-order constraint qualification is trivially satisfied for the box-constrained lower-level problem. Instead of the full second-order conditions we propose a simple necessary check, by requiring that the second derivative with respect to each parameter be nonpositive unless this parameter is at one of its bounds. With the use of the binary variables described above, this condition can be formulated as

$$g_{p_j p_j}(\mathbf{x}, \mathbf{p}) \leq (y_j + y_{n_p+j})\, g_{p_j p_j}^{max}, \qquad j = 1, \ldots, n_p,$$

where $g_{p_j p_j}^{max}$ is an upper bound for the second derivative $g_{p_j p_j}$ on $X \times P$. Similarly to the bounds on the KKT multipliers, it suffices to take a bound satisfying

$$g_{p_j p_j}^{max} \geq \max_{\mathbf{x}\in X, \mathbf{p}\in P, p_j \in \{p_j^L, p_j^U\}} g_{p_j p_j}(\mathbf{x}, \mathbf{p}).$$

This bound can be calculated by interval extensions on $X \times P$. Note that a negative bound is acceptable here.

4.4. Validity and convergence of lower bounding problem. In discretization approaches, tightening of the lower bounds can be achieved by gradually increasing the cardinality of the finite set of parameters considered, i.e., increasing the number of constraints. Methods for efficient convergence of the proposed lower bounding problem to the optimal solution value of the SIP are outside the scope of this paper, and here we only briefly discuss basic convergence properties. We will show that a subdivision of P leads to convergence of the lower bound.

Suppose that the host set P is subdivided into a collection of subsets $P_\tau \subset P$ with index set I, satisfying $P_\tau = [\mathbf{p}_\tau^L, \mathbf{p}_\tau^U]$. We now require matrices of variables $\mathbf{P}$ and $\mathbf{Y}$ whose columns are indexed by the set I, i.e., for $\tau \in I$ the vectors $\mathbf{p}_\tau \in P_\tau$ and $\mathbf{y}_\tau \in \{0,1\}^{2n_p}$ represent columns of $\mathbf{P}$ and $\mathbf{Y}$, respectively. The lower bounding problem (4.7) with the inclusion of the second-order conditions now becomes

$$
\begin{aligned}
f^{I,LBD} = \min_{\mathbf{x},\mathbf{P},\mathbf{Y}} \; & f(\mathbf{x}) \\
\text{s.t. } & g_{p_{j,\tau}}(\mathbf{x}, \mathbf{p}_\tau) \le y_{j,\tau}\mu_{j,\tau}^{max}, && j = 1,\dots,n_p, \quad \tau \in I, \\
& p_{j,\tau}^U - p_{j,\tau} \le \left(p_{j,\tau}^U - p_{j,\tau}^L\right)(1 - y_{j,\tau}), && j = 1,\dots,n_p, \quad \tau \in I, \\
& -g_{p_j}(\mathbf{x}, \mathbf{p}_\tau) \le y_{n_p+j,\tau}\mu_{n_p+j,\tau}^{max}, && j = 1,\dots,n_p, \quad \tau \in I, \\
& p_{j,\tau} - p_{j,\tau}^L \le \left(p_{j,\tau}^U - p_{j,\tau}^L\right)\left(1 - y_{n_p+j,\tau}\right), && j = 1,\dots,n_p, \quad \tau \in I, \\
& y_{j,\tau} + y_{n_p+j,\tau} \le 1, && j = 1,\dots,n_p, \quad \tau \in I, \\
& g_{p_j p_j}(\mathbf{x}, \mathbf{p}_\tau) \le (y_{j,\tau} + y_{n_p+j,\tau})\, g_{p_j p_j,\tau}^{max}, && j = 1,\dots,n_p, \quad \tau \in I, \\
& g(\mathbf{x}, \mathbf{p}_\tau) \le 0, && \tau \in I, \\
& g(\mathbf{x}, \hat{\mathbf{p}}) \le 0 && \forall \hat{\mathbf{p}} \in P_D, \\
& \mathbf{Y} \in \{0,1\}^{2n_p \times |I|}, \\
& \mathbf{p}_\tau \in P_\tau, && \tau \in I, \\
& \mathbf{x} \in X.
\end{aligned}
\tag{4.8}
$$

Note that, to ensure the existence of a KKT point in each subset P_τ, it is necessary to use the interior bounds in the KKT conditions ($p_{j,\tau}^U$ as opposed to p_j^U).

PROPOSITION 4.3 (validity of the lower bounding problem). *The optimal solution value $f^{I,LBD}$ obtained by* (4.8) *is a valid lower bound to the optimal objective value of SIP* (1.1)*, or* (4.8) *is a relaxation of* (1.1)*.*

Proof. Since the objective function in (4.8) is the same as in (1.1), it suffices to show that if a point $\bar{\mathbf{x}}$ is feasible in (1.1), then it is feasible in (4.8). This is done by showing that one can pick points $\bar{\mathbf{P}}$, $\bar{\mathbf{Y}}$ satisfying the constraints.

Let $\bar{\mathbf{x}}$ be feasible in (1.1); i.e., $g(\bar{\mathbf{x}}, \mathbf{p}) \le 0$ for all $\mathbf{p} \in P$. Since $P_D \subset P$, we obtain directly $g(\bar{\mathbf{x}}, \hat{\mathbf{p}}) \le 0$ for all $\hat{\mathbf{p}} \in P_D$. For each $\tau \in I$, pick $\bar{\mathbf{p}}_\tau \in \arg\max_{\mathbf{p}_\tau \in P_\tau} g(\bar{\mathbf{x}}, \mathbf{p}_\tau)$. Since $P_\tau \subset P$, we directly obtain $\bar{\mathbf{p}}_\tau \in P$ and therefore $g(\bar{\mathbf{x}}, \bar{\mathbf{p}}_\tau) \le 0$. Since each program $\max_{\mathbf{p}_\tau \in P_\tau} g(\bar{\mathbf{x}}, \mathbf{p}_\tau)$ is box-constrained, it satisfies the linear constraint qualification, and the first and second-order KKT conditions are necessary for a maximum. As a consequence, in analogy to the proof of Proposition 4.2, one can pick $\bar{\mathbf{y}}_\tau$ that satisfy the first six constraints of (4.8). ∎

PROPOSITION 4.4 (convergence of lower bounding problem). *Consider a sequence of successively refined subdivisions with index set I_k. Consider the relaxation of* (4.8) *(with $P_D = \emptyset$ and without the KKT necessary conditions),*

$$
\begin{aligned}
f^{LBD,k} = \min_{\mathbf{x},\mathbf{P}}\ & f(\mathbf{x}) \\
s.t.\ & g(\mathbf{x},\mathbf{p}_\tau) \le 0, \quad \tau \in I_k, \\
& \mathbf{p}_\tau \in P_\tau, \quad \tau \in I_k, \\
& \mathbf{x} \in X.
\end{aligned} \tag{4.9}
$$

If the subdivision is exhaustive, then for $k \to \infty$, $f^{LBD,k} \to f^*$.

Note that since the subdivision is refined, essentially an implicit discretization is defined, and convergence results for discretizations are well known, e.g., [41]. Therefore the proof of Proposition 4.4 is based on the proof of Theorem 2.8 of [41].

Proof. Since the subdivision is successively refined, the sequence of lower bounds obtained is nondecreasing. Moreover, by Proposition 4.3 we have $f^{LBD,k} \le f^*$. Therefore, $f^{LBD,k}$ converges for $k \to \infty$. Denote its limit f^{lim} and note that $f^{lim} \le f^*$.

Since X is compact, a subsequence $\mathbf{x}^{k_i}$ which converges to $\mathbf{x}^{lim,i}$ with $f(\mathbf{x}^{lim,i}) = f^{lim}$ exists. Note now that the solution of (4.9) furnishes $\mathbf{p}_\tau^{k_i}$ for each $\mathbf{x}^{k_i}$ satisfying $g(\mathbf{x}^{k_i}, \mathbf{p}_\tau^{k_i}) \le 0$ for all $\tau \in I_k$ Since the subdivision is exhaustive, for k_i large enough and some τ the points $\mathbf{p}_\tau^{k_i}$ are arbitrarily close to points in $\arg\max_{\mathbf{p}\in P} g(\mathbf{x}^{lim,i}, \mathbf{p})$. The continuity of g therefore implies $g^U(\mathbf{x}^{lim,i}) \le 0$, or $\mathbf{x}^{lim,i}$ is feasible for the original SIP. Therefore, $f^{lim} \ge f^*$. ∎

5. Upper bounding problems. The basic principle of our upper bounding proposals is the same as that of Bhattacharjee and coworkers [12, 13], namely that a relaxation of the lower-level program leads to a restriction of the semi-infinite constraint and thus a restriction of the SIP. Here, instead of using an inclusion function based on interval analysis, we employ convex relaxations and/or linear relaxations. First, we extend the convex relaxations to the parametric case. Subsequently, we describe the alternatives and then discuss some basic convergence properties.

5.1. Parametric concave relaxations. We are interested in constructing a concave relaxation of the lower-level program $\max_{\mathbf{p}\in P} g(\mathbf{x},\mathbf{p})$ for each $\mathbf{x} \in X$ and hence a parametric concave relaxation. Procedures for parametric relaxations are not available in the literature in the extent needed for our upper bounding proposals. Relaxations have been used in parametric optimization, but the focus has been on LP-relaxations of parametric mixed integer linear programs; see, e.g., Ohtake and Nishida [35]. In nonlinear programming, convex relaxations have been applied to the right-hand side case, e.g., [18], in which the dependence on the parameter is very simple. Relaxations have also been constructed for optimization with dynamic systems embedded, e.g., [45], where the role of the integration variable is similar to a parameter, but the focus is different than what is needed here.

5.1.1. Nonsmooth concave relaxation. Here we extend the procedure described in section 3.2.1 for concave relaxation and linearization, using the composition results of McCormick, to the parametric case. In particular we need to describe how concave relaxations on P can be calculated for joint terms in $\mathbf{x}$ and $\mathbf{p}$. By Assumption 2, the constraint g can be decomposed in such a way that all joint terms are of the form

$$v_j(\mathbf{x},\mathbf{p}) = t(\mathbf{x}) + u(\mathbf{x})v_i(\mathbf{x},\mathbf{p}),$$

where, for v_i, convex and concave relaxations ($v_i^u(\mathbf{x},\cdot)$, $v_i^o(\mathbf{x},\cdot)$) as well as bounds ($v_i^L(\mathbf{x})$, $v_i^U(\mathbf{x})$) are known on P for each $\mathbf{x} \in X$. The convex and concave relaxations

of $v_j(\mathbf{x},\cdot)$ on P are given by

$$v_j^u(\mathbf{x},\mathbf{p}) = \min\{t(\mathbf{x}) + u(\mathbf{x})v_i^u(\mathbf{x},\mathbf{p}), t(\mathbf{x}) + u(\mathbf{x})v_i^o(\mathbf{x},\mathbf{p})\},$$
$$v_j^o(\mathbf{x},\mathbf{p}) = \max\{t(\mathbf{x}) + u(\mathbf{x})v_i^u(\mathbf{x},\mathbf{p}), t(\mathbf{x}) + u(\mathbf{x})v_i^o(\mathbf{x},\mathbf{p})\},$$

and the bounds of $v_j(\mathbf{x},\cdot)$ on P are given by

$$v_j^L(\mathbf{x},\mathbf{p}) = \min\{t(\mathbf{x}) + u(\mathbf{x})v_i^L(\mathbf{x},\mathbf{p}), t(\mathbf{x}) + u(\mathbf{x})v_i^U(\mathbf{x},\mathbf{p})\},$$
$$v_j^U(\mathbf{x},\mathbf{p}) = \max\{t(\mathbf{x}) + u(\mathbf{x})v_i^L(\mathbf{x},\mathbf{p}), t(\mathbf{x}) + u(\mathbf{x})v_i^U(\mathbf{x},\mathbf{p})\}.$$

By recursive application any factorable function can be handled, generally with a nesting of min and max statements. Note at this point the similarity to the ICR proposal by Bhattacharjee, Green, and Barton [12]. Note also that propagating subgradients from v_i to v_j is straightforward and constructing linearizations poses no significant challenge over the NLP case considered in [8].

Example 5.1. Consider a simple SIP for which $x \in [-1,1]$, $p \in [-1,1]$, $f = -x$, and $g = e^{x\,p^2} - 1$. The feasible set is easily calculated to be $x \le 0$, and therefore the optimal solution value is 0 at $x = 0$. A factorable representation of $g(x,p)$ is given by

$$v_1 = p^2,$$
$$v_2 = x\,v_1,$$
$$v_3 = e^{v_2} - 1.$$

Since p^2 is convex, its convex envelope on P is given by $v_1^u = p^2$, and its concave envelope on P is given by $v_1^o = 1$; its bounds are given by $v_1^L = 0$ and $v_1^U = 1$. For the term v_2 we obtain convex and concave envelopes as $v_2^u = \min\{xv_1^u, xv_1^o\}$ and $v_2^o = \max\{xv_1^u, xv_1^o\}$. Its bounds are given by $v_2^L = \min\{x \cdot 0, x \cdot 1\} = \min\{0, x\}$ and $v_2^U = \max\{0, x\}$. For the term v_3 we need to invoke McCormick's composition theorem. The exponential function e^z is convex, and therefore its concave envelope is given by the secant $T^o = e^{z^L} + \frac{z - z^L}{z^U - z^L}(e^{z^U} - e^{z^L})$, and the convex envelope is given by the function itself $T^u = e^z$. Moreover, the exponential function is monotone increasing, and therefore $\arg\min_{z\in[z^L,z^U]} e^z = \{z^L\}$ and $\arg\max_{z\in[z^L,z^U]} e^z = \{z^U\}$. Therefore, $\mathrm{mid}\{v_2^u, v_2^o, w_{min}\} = v_2^u$ and $\mathrm{mid}\{v_2^u, v_2^o, w_{max}\} = v_2^o$. By McCormick's composition theorem,

$$e^{v_2^u} - 1 = e^{\min\{xp^2, x\}} - 1$$

is a convex underestimator of $g(\mathbf{x},\cdot)$ on P and

$$e^{v_2^L} + \frac{v_2^o - v_2^L}{v_2^U - v_2^L}\left(e^{v_2^U} - e^{v_2^L}\right) - 1$$
$$= e^{\min\{0,x\}} + \frac{\max\{xp^2, x\} - \min\{0,x\}}{\max\{0,x\} - \min\{0,x\}}\left(e^{\max\{0,x\}} - e^{\min\{0,x\}}\right) - 1$$

is a concave overestimator of $g(\mathbf{x},\cdot)$ on P.

5.1.2. Smooth concave relaxations without auxiliary variables. We now consider the extension of the smooth overestimation of the lower-level program via the addition of known concave terms; compare section 3.2.2. Without loss of generality we consider the αBB relaxations, which is also handled in [19]. Use of the γBB

relaxations is analogous. The αBB overestimation $g^{o,\alpha}(\mathbf{x},\cdot)$ of $g(\mathbf{x},\cdot)$ on P is given by

$$g^{o,\alpha}(\mathbf{x},\mathbf{p}) = g(\mathbf{x},\mathbf{p}) + \alpha \sum_{i=1}^{n_p} (p_i - p_i^L)(p_i^U - p_i).$$

For sufficiently large values of α, the overestimating function $g^{o,\alpha}(\mathbf{x},\cdot)$ is partially concave on P for each $\mathbf{x} \in X$.

In principle α can be taken as a function of $\mathbf{x}$, but since no closed form for the calculation of α exists in general and introducing an $\mathbf{x}$-dependence may lead to nonsmooth constraints with respect to $\mathbf{x}$, we refrain from doing this. Instead we obtain α via interval extensions of the eigenvalue estimates of the Hessian matrix on $X \times P$. This overestimation guarantees the desired concavity with the drawback that the relaxations are weaker than necessary. The relaxed lower-level program

$$\max_{\mathbf{p}\in P} g^{o,\alpha}(\mathbf{x},\mathbf{p})$$

is a box-constrained maximization program with a smooth concave objective function. Note that for convergence of $g^{o,\alpha}$ to g it is sufficient to subdivide P, without partitioning X. Within a branch-and-bound procedure, it is advisable, though, to recalculate α for each node, since this accelerates convergence.

Example 5.2. Recall Example 5.1. The second derivative of g with respect to p is given by $(2x + 4x^2p^2)e^{xp^2}$. Calculating the natural interval extension of the second derivative on $X \times P$ gives $6e$. A concave overestimator of $g(x,\cdot)$ on P is therefore given by

$$e^{x\,p^2} - 1 + 3e(p+1)(1-p) = e^{x\,p^2} - 1 + 3e(1-p^2).$$

5.1.3. Smooth concave relaxation with auxiliary variables. We now consider the alternative of introducing auxiliary variables and constraints; compare also section 3.2.3. This is similar to the nonsmooth concave relaxation. By Assumption 2, g can be decomposed in such a way that the joint terms are of the form

$$t(\mathbf{x}) + u(\mathbf{x})w_i,$$

where w_i is a previously introduced variable (or a parameter p_i). If such terms directly appear as a summand in g, no relaxation is needed. Otherwise a new variable w_j is introduced along with a linear equality constraint

$$w_j = t(\mathbf{x}) + u(\mathbf{x})w_i.$$

The bounds on this new auxiliary variable can in principle be calculated as functions of $\mathbf{x}$. Since our final goal is a smooth program with respect to $\mathbf{x}$ we calculate the bounds on this auxiliary variable by taking natural interval extensions with respect to both $\mathbf{p}$ and $\mathbf{x}$. In general this is a further relaxation.

For instance, the term x_ip_j would be replaced by a new variable $w_k \in [w_k^L, w_k^U]$ and a constraint $w_k = x_ip_j$, where the variable bounds are given by $w_k^L = \min\{x_i^Lp_j^L, x_i^Lp_j^U, x_i^Up_j^L, x_i^Up_j^U\}$ and $w_k^U = \max\{x_i^Lp_j^L, x_i^Lp_j^U, x_i^Up_j^L, x_i^Up_j^U\}$.

To obtain a compact presentation we augment the parameters $\mathbf{p}$ with the auxiliary variables and denote these $\tilde{\mathbf{p}} \in \tilde{P} \subset \mathbb{R}^{n_{\tilde{p}}}$. Also, we lump the box and auxiliary constraints into inequality constraints formed from the functions $\mathbf{u} : X \times \mathbb{R}^{n_{\tilde{p}}} \to \mathbb{R}^{n_u}$.

The presence of linear equalities is omitted for simplicity. The relaxed lower-level program is then given by

$$\begin{aligned} &\max_{\tilde{\mathbf{p}}\in\mathbb{R}^{n_{\tilde{p}}}} g^{o,ex}(\mathbf{x},\tilde{\mathbf{p}}) \\ &\text{s.t. } \mathbf{u}(\mathbf{x},\tilde{\mathbf{p}}) \leq \mathbf{0}. \end{aligned} \tag{5.1}$$

The resulting restriction of (1.1),

$$\begin{aligned} f_{gsip}^{UBD,ex} = &\min_{\mathbf{x}\in X} f(\mathbf{x}) \\ &\text{s.t. } g^{o,ex}(\mathbf{x},\tilde{\mathbf{p}}) \leq 0 \quad \forall \tilde{\mathbf{p}} \in \mathbb{R}^{n_{\tilde{p}}} : \mathbf{u}(\mathbf{x},\tilde{\mathbf{p}}) \leq \mathbf{0}, \end{aligned}$$

is a GSIP. By construction, for all $\mathbf{x} \in X$ there exists a $\tilde{\mathbf{p}}$ such that $\mathbf{u}(\mathbf{x},\tilde{\mathbf{p}}) \leq \mathbf{0}$ and the GSIP can be reformulated to a bilevel program [47].

Example 5.3. *Recall Example* 5.1 *and the factorable presentation of* $g = e^{x\,p^2} - 1$. *The nonlinear term* p^2 *is replaced by an auxiliary variable* $\tilde{p}_2 \in [0,1]$ *along with two convex inequality constraints*

$$\begin{aligned} \tilde{p}_1^2 - \tilde{p}_2 &\leq 0, \\ \tilde{p}_2 - 1 &\leq 0. \end{aligned}$$

The joint term $x\tilde{p}_2$ *is replaced by a new auxiliary variable* $\tilde{p}_3 \in [0,1]$ *(bounds incidentally exact) along with a linear (in* $\tilde{\mathbf{p}}$*) equality constraint*

$$\tilde{p}_3 - x\tilde{p}_2 = 0.$$

The exponential term $e^{\tilde{p}_3}$ *is convex and can be overestimated by the secant*

$$e^{\tilde{p}_3^L} + \frac{\tilde{p}_3 - \tilde{p}_3^L}{\tilde{p}_3^U - \tilde{p}_3^L}\left(e^{\tilde{p}_3^U} - e^{\tilde{p}_3^L}\right) = e^0 + \frac{\tilde{p}_3 - 0}{1-0}\left(e^1 - e^0\right) = 1 + \tilde{p}_3(e-1).$$

Since the exponential term directly appears in g*, no additional auxiliary variable is introduced. The concave relaxation of the lower-level program is given by*

$$\begin{aligned} &\max_{\tilde{\mathbf{p}}} \tilde{p}_3(e-1) \\ &s.t.\ \tilde{p}_1^2 - \tilde{p}_2 \leq 0, \\ &\qquad \tilde{p}_2 - 1 \leq 0, \\ &\qquad \tilde{p}_3 - x\tilde{p}_2 = 0, \\ &\qquad \tilde{\mathbf{p}} \in [-1,1] \times [0,1] \times [0,1]. \end{aligned}$$

5.2. KKT-based upper bound. In the following we describe how to obtain an upper bound from the solution of an MPEC. The first step in obtaining the upper bound is to construct a relaxation of (2.1), i.e., a maximization problem with constraints that are partially convex on $\mathbf{p} \in P$ for each $\mathbf{x} \in X$ and an objective function that is partially concave on $\mathbf{p} \in P$ for each $\mathbf{x} \in X$ and overestimates $g(\mathbf{x},\cdot)$ on P. This relaxation of the lower-level program results in a restriction of the SIP.

The next step is to replace the resulting SIP with an MPEC similar to the one described for the lower bounding problem. A basic requirement for this transformation is differentiability of the relaxed lower-level program, and therefore only the smooth relaxations described in section 5.1 are applicable. Moreover, for the MPEC to be a

valid restriction, the KKT conditions need to be sufficient only for a global maximum. This is ensured by the (partial) convexity of the programs. Note that necessity of the KKT conditions is not required. If the relaxed lower-level program attains its maximum only at points that are not KKT points, the MPEC will be infeasible. A local solution of the formulated MPEC is a valid upper bound to the original SIP. Therefore, unlike the lower bounding problem, obtaining rigorous bounds on the KKT multipliers is not necessary (it is still helpful). If the bounds on the KKT multipliers are underestimated, the upper bounding problem is further restricted and therefore remains valid, but may be rendered infeasible.

PROPOSITION 5.1 (validity of KKT-based upper bounding problem). *Consider a (generalized) SIP whose lower-level program is a relaxation of the lower-level program* (1.1) *for each* $\mathbf{x} \in X$ *and as such has a nonempty feasible set for each* $\mathbf{x} \in X$. *Suppose that for each* $\mathbf{x} \in X$ *the first-order KKT conditions are sufficient for a global maximum of the lower-level program. Construct an MPEC with* $\mathbf{x}$, $\mathbf{p}$, *and the KKT multipliers of the relaxed lower-level program* $\boldsymbol{\mu}$ *as variables, the same objective function as* (1.1), *the constraint*

$$g^o(\mathbf{x}, \mathbf{p}) \leq 0,$$

along with the constraints of the lower-level program and its first-order KKT conditions, and a finite bound on the KKT multipliers

$$\mu_j \leq \mu_j^{max}.$$

If $\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\boldsymbol{\mu}}$ *is a feasible point of this MPEC, then* $\bar{\mathbf{x}}$ *is a feasible point of* (1.1).

Proof. Any feasible point $\bar{\mathbf{x}}, \bar{\mathbf{p}}, \bar{\boldsymbol{\mu}}$ of the constructed MPEC is also feasible in the MPEC without the bounds on the KKT multipliers. Since the first-order KKT conditions are sufficient for a global maximum of the lower-level program, $\bar{\mathbf{p}}$ is a global maximum of the lower-level program for $\mathbf{x} = \bar{\mathbf{x}}$. The constraint $g^o(\bar{\mathbf{x}}, \bar{\mathbf{p}}) \leq 0$ therefore ensures that the optimal solution value of the relaxed lower-level program satisfies $g^{o,U}(\bar{\mathbf{x}}) \leq 0$. Therefore, because g^o is a concave overestimator, also the optimal solution value of the original lower-level program satisfies $g^U(\bar{\mathbf{x}}) \leq 0$, or $\mathbf{x}$ is feasible in (1.1). Finally, adding an upper bound on the on the KKT multipliers is a third restriction. ∎

At this point a comparison with the interval inclusion approach by Bhattacharjee and coworkers [12, 13] is warranted. The MPEC problems typically have many more additional variables and constraints than the ICR and therefore are significantly harder to solve. Moreover, the stationarity and complementary slackness constraints are equality constraints, and state-of-the-art finitely terminating algorithms guarantee the feasibility of nonlinear equality constrained problems only within a tolerance. In some cases (see below) it can be shown that despite this approximation the generated points are guaranteed feasible in (1.1). On the other hand, typically, convex relaxations are tighter than interval extensions. As a consequence the proposed upper bounds will typically be tighter than those furnished by the ICR.

5.2.1. Smooth concave overestimation without auxiliary variables. We now consider a KKT-based upper bound based on the αBB method; such a bound has also been proposed in Floudas and Stein [19], and here we present some additional concepts, namely how to obtain bounds on the KKT multipliers and how to estimate the maximal constraint violation. As described in section 5.1.2, the smooth relaxation of the lower-level program via αBB results in a box-constrained maximization program

with a smooth concave objective function. Therefore, the first-order KKT conditions are necessary and sufficient for a global maximum and

$$
\begin{aligned}
f^{UBD,\alpha} = \min_{\mathbf{x},\mathbf{p},\boldsymbol{\mu}} \; & f(\mathbf{x}) \\
\text{s.t.} \; & -g^{o,\alpha}_{p_j}(\mathbf{x},\mathbf{p}) + \mu_j - \mu_{n_p+j} = 0, \quad j = 1,\dots,n_p, \\
& \mu_j(p_j - p_j^U) = 0, \quad j = 1,\dots,n_p, \\
& \mu_{n_p+j}(-p_j + p_j^L) = 0, \quad j = 1,\dots,n_p, \\
& g^{o,\alpha}(\mathbf{x},\mathbf{p}) \le 0, \\
& 0 \le \mu_j \le \mu_j^{max}, \quad j = 1,\dots,2n_p, \\
& \mathbf{x} \in X, \quad \mathbf{p} \in P,
\end{aligned}
\tag{5.2}
$$

is equivalent to the restricted SIP for sufficiently large $\boldsymbol{\mu}^{max}$. Note that the number of variables in (5.2) is equal to the original number of variables n_x plus up to three times the number of parameters ($3n_p$). In addition to the box constraints there are up to $3n_p$ equality constraints and one (most likely nonconvex) inequality constraint. Similar to the lower bounding problem, a reformulation to an MINLP is possible by introducing binary variables and eliminating the KKT multipliers.

The calculation of bounds on the KKT multipliers is equivalent to (4.2), (4.3), replacing g with $g^{o,\alpha}$. We first note that

$$
g^{o,\alpha}_{p_j}(\mathbf{x},\mathbf{p}) = g_{p_j}(\mathbf{x},\mathbf{p}) + \alpha(p_j^U - p_j) - \alpha(p_j - p_j^L).
$$

Therefore

$$
\begin{aligned}
\max_{\mathbf{x}\in X,\mathbf{p}\in P,p_j=p_j^U} g^{o,\alpha}_{p_j}(\mathbf{x},\mathbf{p}) &= -\alpha(p_j^U - p_j^L) + \max_{\mathbf{x}\in X,\mathbf{p}\in P,p_j=p_j^U} g_{p_j}(\mathbf{x},\mathbf{p}), \\
\max_{\mathbf{x}\in X,\mathbf{p}\in P,p_j=p_j^L} -g^{o,\alpha}_{p_j}(\mathbf{x},\mathbf{p}) &= \max_{\mathbf{x}\in X,\mathbf{p}\in P,p_j=p_j^L} -\left(g_{p_j}(\mathbf{x},\mathbf{p}) - \alpha(p_j^U - p_j^L)\right) \\
&= \alpha(p_j^U - p_j^L) - \min_{\mathbf{x}\in X,\mathbf{p}\in P,p_j=p_j^L} g_{p_j}(\mathbf{x},\mathbf{p}).
\end{aligned}
$$

Similar to the lower bounding problem, whenever a bound is nonpositive (function monotone) the corresponding variable and complementary slackness conditions are eliminated. For the γBB relaxation [5] the derivatives of the underestimating terms with respect to p_j are variable-dependent, but, evaluated at the variable bounds, they are given by $\gamma(e^{\gamma(p_j^U - p_j^L)} - 1)$ for $p_j = p_j^L$ and $\gamma(1 - e^{\gamma(p_j^U - p_j^L)})$ for $p_j = p_j^U$. Therefore, the calculation of bounds on the KKT multipliers is analogous.

As stated above, typical finitely terminating NLP solvers only approximate equality constraints. We will show that the feasibility of the points furnished can be easily verified, or the extent of constraint violation estimated.

PROPOSITION 5.2 (maximal constraint violation). *Consider $\bar{\mathbf{x}} \in X$, $\bar{\mathbf{p}} \in P$, and $\bar{\boldsymbol{\mu}} \ge \mathbf{0}$ such that $(\bar{\mathbf{x}},\bar{\mathbf{p}},\bar{\boldsymbol{\mu}})$ is an approximate feasible point of (5.2) in the sense of ε_{tol}-violation of the equality constraints*

$$
|-g^{o,\alpha}_{p_j}(\bar{\mathbf{x}},\bar{\mathbf{p}}) + \bar{\mu}_j - \bar{\mu}_{n_p+j}| \le \varepsilon_{tol}, \quad j = 1,\dots,n_p, \tag{5.3}
$$

$$
|\bar{\mu}_j(\bar{p}_j - p_j^U)| \le \varepsilon_{tol}, \quad j = 1,\dots,n_p, \tag{5.4}
$$

$$
|\bar{\mu}_{n_p+j}(-\bar{p}_j + p_j^L)| \le \varepsilon_{tol}, \quad j = 1,\dots,n_p, \tag{5.5}
$$

$$
g^{o,\alpha}(\bar{\mathbf{x}},\bar{\mathbf{p}}) \le 0.
$$

Then $\bar{\mathbf{x}}$ *is an approximately feasible point of* (1.1) *in the sense*

$$g^U(\bar{\mathbf{x}}) \equiv \max_{\mathbf{p}\in P} g(\bar{\mathbf{x}},\mathbf{p}) \leq \varepsilon_{tol} \sum_{j=1}^{n_p} (1 + p_j^U - p_j^L).$$

Proof. Since $g^{o,\alpha}$ is a concave overestimator of g for each $\bar{\mathbf{x}}$ we have

$$g^U(\bar{\mathbf{x}}) \leq g^{o,\alpha,U}(\bar{\mathbf{x}}) \equiv \max_{\mathbf{p}\in P} g^{o,\alpha}(\bar{\mathbf{x}},\mathbf{p}). \tag{5.6}$$

Since $g^{o,\alpha}$ is partially concave on the convex set P we obtain [10, p. 675]

$$g^{o,\alpha}(\bar{\mathbf{x}},\mathbf{p}) \leq g^{o,\alpha}(\bar{\mathbf{x}},\bar{\mathbf{p}}) + \sum_{j=1}^{n_p} g^{o,\alpha}_{p_j}(\bar{\mathbf{x}},\bar{\mathbf{p}})(p_j - \bar{p}_j) \qquad \forall \mathbf{p} \in P,$$

and therefore also

$$g^{o,\alpha,U}(\bar{\mathbf{x}}) \leq \max_{\mathbf{p}\in P} \left(g^{o,\alpha}(\bar{\mathbf{x}},\bar{\mathbf{p}}) + \sum_{j=1}^{n_p} g^{o,\alpha}_{p_j}(\bar{\mathbf{x}},\bar{\mathbf{p}})(p_j - \bar{p}_j) \right).$$

The maximum of the above sum over P is attained at a vertex of P, and therefore

$$g^{o,\alpha,U}(\bar{\mathbf{x}}) \leq g^{o,\alpha}(\bar{\mathbf{x}},\bar{\mathbf{p}}) + \sum_{j=1}^{n_p} \max\left\{ g^{o,\alpha}_{p_j}(\bar{\mathbf{x}},\bar{\mathbf{p}})(p_j^U - \bar{p}_j), g^{o,\alpha}_{p_j}(\bar{\mathbf{x}},\bar{\mathbf{p}})(p_j^L - \bar{p}_j) \right\}. \tag{5.7}$$

Since $\bar{\boldsymbol{\mu}} \geq \mathbf{0}$ from (5.3) we obtain

$$\begin{aligned} |g^{o,\alpha}_{p_j}(\bar{\mathbf{x}},\bar{\mathbf{p}})| &\leq \bar{\mu}_j + \varepsilon_{tol}, \\ |g^{o,\alpha}_{p_j}(\bar{\mathbf{x}},\bar{\mathbf{p}})| &\leq \bar{\mu}_{n_p+j} + \varepsilon_{tol}, \end{aligned}$$

and therefore by (5.4), (5.5)

$$\begin{aligned} |g^{o,\alpha}_{p_j}(\bar{\mathbf{x}},\bar{\mathbf{p}})(p_j^U - \bar{p}_j)| &\leq \varepsilon_{tol}(1 + p_j^U - \bar{p}_j) \leq \varepsilon_{tol}(1 + p_j^U - p_j^L), \\ |g^{o,\alpha}_{p_j}(\bar{\mathbf{x}},\bar{\mathbf{p}})(p_j^L - \bar{p}_j)| &\leq \varepsilon_{tol}(1 + \bar{p}_j - p_j^L) \leq \varepsilon_{tol}(1 + p_j^U - p_j^L). \end{aligned}$$

Combining the above inequalities with (5.6) and (5.7), we obtain the desired result. ∎

A practical way of ensuring that the points furnished by (5.2) are indeed feasible is to replace the constraint $g^{o,\alpha}(\mathbf{x},\mathbf{p}) \leq 0$ with $g^{o,\alpha}(\mathbf{x},\mathbf{p}) \leq -\varepsilon_{tol} \sum_{j=1}^{n_p}(1 + p_j^U - p_j^L)$. Note that this is a further restriction, and therefore again a valid upper bound is obtained.

5.2.2. Concave overestimation with auxiliary variables. We now consider the alternative of introducing auxiliary variables and constraints, (cf. section 5.1.3), which resulted in a lower-level program of the type (5.1). By convexity, the KKT

conditions are sufficient for a global maximum of the lower-level program and

$$
\begin{aligned}
f^{UBD,ex} = \min_{\mathbf{x},\tilde{\mathbf{p}},\boldsymbol{\mu}} \; & f(\mathbf{x}) \\
\text{s.t.} \; & -\boldsymbol{\nabla}_{\tilde{\mathbf{p}}} g^{o,ex}(\mathbf{x},\tilde{\mathbf{p}}) + \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{\nabla}_{\tilde{\mathbf{p}}} \mathbf{u}(\mathbf{x},\tilde{\mathbf{p}}) = \mathbf{0}, \\
& \mu_j u_j(\mathbf{x},\tilde{\mathbf{p}}) = 0, \quad j = 1,\ldots,n_u, \\
& g^{o,ex}(\mathbf{x},\tilde{\mathbf{p}}) \leq 0, \\
& \mathbf{u}(\mathbf{x},\tilde{\mathbf{p}}) \leq \mathbf{0}, \\
& \mu_j^{min} \leq \mu_j \leq \mu_j^{max}, \quad j = 1,\ldots,n_u, \\
& \mathbf{x} \in X, \quad \tilde{\mathbf{p}} \in \mathbb{R}^{n_{\tilde{p}}},
\end{aligned}
\tag{5.8}
$$

provides a valid upper bound of (1.1), where for inequality constraints $\mu_j^{min} = 0$. Recall that the number of variables in the lower-level problem $n_{\tilde{p}}$ and the number of KKT multipliers n_u depend on the number and type of factors in the factorable representation. Here a reformulation as an MINLP is possible by the introduction of binary variables, but elimination of the KKT multipliers does not seem possible in general.

At this point a discussion of potential disadvantages of (5.8) is warranted. The KKT conditions may not be necessary for problems of the type (5.1) since it has not been shown that the formulated constraints satisfy some constraint qualification (see also section 3.2.3). Moreover, obtaining valid upper bounds on the KKT multipliers is not always possible, so replacing the restricted SIP with an MPEC may be a further restriction and render the upper bounding program infeasible. To ensure convergence of the upper bounding problem this issue has to be addressed. Finally, through the introduction of extra variables, providing tight bounds on the maximal constraint violation seems intractable.

Example 5.4. Recall Example 5.1 and the relaxed lower-level program constructed in Example 5.3. The following MPEC is obtained as the upper bounding problem:

$$
\begin{aligned}
\min_{x,\tilde{\mathbf{p}},\boldsymbol{\mu}} \; & -x \\
s.t. \quad & 2\mu_1\tilde{p}_1 + \mu_4 - \mu_5 = 0, \\
& -\mu_1 + \mu_2 - x\mu_3 + \mu_6 - \mu_7 = 0, \\
& -(e-1) + \mu_3 + \mu_8 - \mu_9 = 0, \\
& \mu_1(\tilde{p}_1^2 - \tilde{p}_2) = 0, \\
& \mu_2(\tilde{p}_2 - 1) = 0, \\
& \mu_3(\tilde{p}_3 - x\tilde{p}_2) = 0, \\
& \mu_4(\tilde{p}_1 - 1) = 0, \\
& \mu_5(-1 - \tilde{p}_1) = 0, \\
& \mu_6(\tilde{p}_2 - 1) = 0, \\
& \mu_7(-\tilde{p}_2) = 0, \\
& \mu_8(\tilde{p}_3 - 1) = 0, \\
& \mu_9(-\tilde{p}_2) = 0, \\
& \tilde{p}_3(e-1) \leq 0, \\
& \tilde{p}_1^2 - \tilde{p}_2 \leq 0, \\
& \tilde{p}_2 - 1 \leq 0, \\
& \tilde{p}_3 - x\tilde{p}_2 = 0,
\end{aligned}
$$

$$\begin{aligned}
&\mu_3 \in [\mu_3^{min}, \mu_3^{max}],\\
&\mu_j \in [0, \mu_j^{max}], \quad j = 1, \ldots, 9 : j \neq 3,\\
&x \in [-1, 1],\\
&\tilde{\mathbf{p}} \in [-1, 1] \times [0, 1] \times [0, 1].
\end{aligned}$$

5.3. Linearization-based upper bound. Similar to the MPEC-based upper bounds, the first step in the linearization-based upper bounds is to construct a convex relaxation of the lower-level program and thus a restriction of (1.1):

$$\begin{aligned}
&\min_{\mathbf{x}\in X} f(\mathbf{x})\\
&\text{s.t. } g^o(\mathbf{x}, \mathbf{p}) \leq 0 \quad \forall \mathbf{p} \in P,
\end{aligned}$$

where $g^o : X \times P \to \mathbb{R}$ is partially concave on P for each $\mathbf{x} \in X$ and $g^o(\mathbf{x}, \mathbf{p})$ overestimates $g(\mathbf{x}, \mathbf{p})$. Note that for the approach involving auxiliary variables a somewhat different treatment is needed, and this is described in section 5.3.3.

The second step further restricts the generated SIP by linearizing at an arbitrary interior point $\bar{\mathbf{p}} \in \text{int}(P)$, pointwise in X, and creating the following SIP:

$$\begin{aligned}
&\min_{\mathbf{x}\in X} f(\mathbf{x})\\
&\text{s.t. } g^{o,lin}(\mathbf{x}, \mathbf{p}) \leq 0 \quad \forall \mathbf{p} \in P,
\end{aligned} \tag{5.9}$$

where $g^{o,lin}(\mathbf{x}, \mathbf{p}) \equiv g^o(\mathbf{x}, \bar{\mathbf{p}}) + \sum_{j=1}^{n_p} g^o_{p_j}(\mathbf{x}, \bar{\mathbf{p}})(p_j - \bar{p}_j)$. Here g_{p_j} denotes a partial derivative or, with an abuse of notation, subgradient. An equivalent nonsmooth reformulation of (5.9) is the following problem:

$$\begin{aligned}
&\min_{\mathbf{x}\in X} f(\mathbf{x})\\
&\text{s.t. } \max_{\mathbf{p}\in P} g^{o,lin}(\mathbf{x}, \mathbf{p}) \leq 0.
\end{aligned} \tag{5.10}$$

Since $g^{o,lin}$ is affine in $\mathbf{p}$, the maximum of $g^{o,lin}(\mathbf{x}, \cdot)$ on P will be attained at one of the vertices P_e of P for each $\mathbf{x}$ in X. Therefore an equivalent finite representation of (5.10) is

$$\begin{aligned}
&\min_{\mathbf{x}\in X} f(\mathbf{x})\\
&\text{s.t. } g^{o,lin}(\mathbf{x}, \mathbf{p}) \leq 0 \quad \forall \mathbf{p} \in P_e.
\end{aligned} \tag{5.11}$$

While for any $\bar{\mathbf{p}} \in P$ the formulated finite NLP (5.11) is a valid restriction of (1.1), the choice of $\bar{\mathbf{p}}$ greatly affects the strength of the generated upper bounds. Compared to the MPEC-based upper bound, this linearization approach presents the inherent advantage that it avoids the use of equality constraints (complementarity and stationarity conditions), and any feasible point of (5.11) is guaranteed feasible for (1.1). On the other hand, the MPEC approach introduces a polynomial (in the number of inner variables or in the number of inner variables and nonconvex terms) number of constraints, whereas the linearization approach introduces a potentially exponential number of constraints. Moreover, the linearization approach produces bounds that are at best as tight as the MPEC-based ones, assuming that both problems are solved to global optimality.

If either of the two following relationships holds for variable p_j,

$$\max_{\mathbf{x}\in X} g^o_{p_j}(\mathbf{x},\bar{\mathbf{p}}) \le 0,$$
$$\min_{\mathbf{x}\in X} g^o_{p_j}(\mathbf{x},\bar{\mathbf{p}}) \ge 0,$$

the number of constraints can be reduced. The following procedure describes how to obtain the (sufficient) subset of extreme points P_{e^*} that needs to be considered in problem (5.11):

- Initialize $P_{e^*} := P_e$.
- **FOR** $j = 1,\dots,n_p$ **DO**
  - **IF** $\max_{\mathbf{x}\in X} g^o_{p_j}(\mathbf{x},\bar{\mathbf{p}}) \le 0$ **THEN** $P_{e^*} := \{\mathbf{p}\in P_{e^*} : p_j = p_j^L\}$
    ELSE IF $\min_{\mathbf{x}\in X} g^o_{p_j}(\mathbf{x},\bar{\mathbf{p}}) \ge 0$ **THEN** $P_{e^*} := \{\mathbf{p}\in P_{e^*} : p_j = p_j^U\}$.

  END

Evaluating the above optimization programs is expensive, and we propose to estimate them using interval extensions.

5.3.1. Smooth concave overestimation without auxiliary variables. Recall that the concave relaxation of g on P using αBB techniques has the form

$$g^{o,\alpha}(\mathbf{x},\mathbf{p}) = g(\mathbf{x},\mathbf{p}) + \alpha\sum_{j=1}^{n_p}(p_j - p_j^L)(p_j^U - p_j),$$

and the linearized approximation of the αBB concave relaxation around a point $\bar{\mathbf{p}}\in P$ is

$$g^{o,\alpha,lin}(\mathbf{x},\mathbf{p}) = g^{o,\alpha}(\mathbf{x},\bar{\mathbf{p}}) + \sum_{j=1}^{n_p} g_{p_j}(\mathbf{x},\bar{\mathbf{p}})(p_j-\bar{p}_j) + \alpha\sum_{j=1}^{n_p}(-2\bar{p}_j + p_j^L + p_j^U)(p_j-\bar{p}_j).$$

Therefore, the αBB-based linearized upper bounding problem is of the form

$$\min_{\mathbf{x}\in X} f(\mathbf{x})$$
$$\text{s.t. } g^{o,\alpha,lin}(\mathbf{x},\mathbf{p}) \le 0 \quad \forall \mathbf{p}\in P_{e^*},$$

where P_{e^*} is calculated by the following procedure:

- Initialize $P_{e^*} := P_e$.
- **FOR** $j = 1,\dots,n_p$ **DO**
  - **IF** $\max_{\mathbf{x}\in X} g_{p_j}(\mathbf{x},\bar{\mathbf{p}}) \le \alpha(2\bar{p}_j - p_j^L - p_j^U)$ **THEN** $P_{e^*} := \{\mathbf{p}\in P_{e^*} : p_j = p_j^L\}$
    ELSE IF $\min_{\mathbf{x}\in X} g_{p_j}(\mathbf{x},\bar{\mathbf{p}}) \ge \alpha(2\bar{p}_j - p_j^L - p_j^U)$ **THEN** $P_{e^*} := \{\mathbf{p}\in P_{e^*} : p_j = p_j^U\}$.

  END

Again we propose to estimate the above optimization problems by interval extensions.

5.3.2. Nonsmooth concave overestimation without auxiliary variables. Similar to the aforementioned technique, the goal of this method is to introduce a concave overestimator of the constraint g with respect to the inner variables $\mathbf{p}$ using the McCormick technique, and then to linearize the resulting expression around an arbitrary point $\bar{\mathbf{p}}\in P$.

As described in section 5.1.1 and demonstrated in Example 5.1, the value of $\mathbf{x}$ can influence the functional form of the convex and concave overestimators. Therefore,

the linearized constraint $g^{o,mc,lin}(\cdot,\mathbf{p}) \leq 0$ evaluated at the vertices of the parameter set $\mathbf{p} \in P_e$ is typically nonsmooth with respect to $\mathbf{x}$. This is very similar to the ICR by Bhattacharjee, Green, and Barton [12], and we propose to work around the nonsmoothness in the same way, namely to use both constraints for each min / max statement resulting from a joint term in $\mathbf{x}$ and $\mathbf{p}$. This is a further restriction and as such provides a valid upper bound. For nested min / max statements the number of constraints becomes exponential in the number min / max statements. Alternatives such as the MINLP reformulation in [12] are also possible.

Similar to the linearization of the smooth concave overestimation, the set of vertices at which the constraints are evaluated can be reduced if the subgradients $g^{o,mc}(\mathbf{x},\cdot)$ at $\bar{\mathbf{p}}$ have positive or negative elements for all $\mathbf{x} \in X$.

Example 5.5. Recall Example 5.1. Linearizing the McCormick relaxation at $\bar{p} = 0$ gives

$$\begin{aligned} &e^{\min\{0,x\}} + \frac{\max\{x,0\} - \min\{0,x\}}{\max\{0,x\} - \min\{0,x\}} \left(e^{\max\{0,x\}} - e^{\min\{0,x\}}\right) - 1 \\ &= e^{\max\{0,x\}} - 1 = \max\{e^0, e^x\} - 1 = \max\{1, e^x\} - 1 = \max\{0, e^x - 1\}. \end{aligned}$$

Introducing the constraint $\max\{0, e^x - 1\} \leq 0$ would give a nonsmooth NLP, and we instead use two constraints corresponding to $x \geq 0$ and $x \leq 0$. In principle these constraints are evaluated at the vertices of P, but since the subgradient at $\bar{p} = 0$ is given by 0, the constraints are introduced only for one parameter value. The resulting single-level program is

$$\begin{aligned} \min_{x \in [-1,1]} \; & -x \\ \text{s.t. } & 0 \leq 0, \\ & e^x - 1 \leq 0. \end{aligned}$$

Incidentally, the above NLP is convex and has the same feasible set as the original program and therefore the same optimal solution point and value.

5.3.3. Smooth concave overestimation with auxiliary variables. A method to create a valid upper bound for (1.1) based on smooth concave overestimation of $g(\mathbf{x},\cdot)$ using auxiliary variables was presented in section 5.2.2. Recall that the following GSIP is a restriction of (1.1):

$$\begin{aligned} &\min_{\mathbf{x} \in X} f(\mathbf{x}) \\ &\text{s.t. } g^{o,ex}(\mathbf{x},\tilde{\mathbf{p}}) \leq 0 \quad \forall \tilde{\mathbf{p}} \in \mathbb{R}^{n_{\tilde{p}}} : \mathbf{u}(\mathbf{x},\tilde{\mathbf{p}}) \leq \mathbf{0}, \end{aligned} \tag{5.12}$$

where the parameters $\tilde{\mathbf{p}}$ contain the original parameters $\mathbf{p}$ and auxiliary parameters representing expressions of the variables and parameters. Bounds on the auxiliary parameters are propagated through interval extensions. The linearization approaches require that the set of parameter vertices be easily calculated, which is not the case here. Therefore, a further restriction of (5.12) is obtained by dropping the lower-level constraints with the exception of the bound constraints

$$\begin{aligned} &\min_{\mathbf{x} \in X} f(\mathbf{x}) \\ &\text{s.t. } g^{o,ex}(\mathbf{x},\tilde{\mathbf{p}}) \leq 0 \quad \forall \tilde{\mathbf{p}} \in \tilde{P}, \end{aligned} \tag{5.13}$$

and therefore further relaxing the lower-level program and thus further restricting (5.12). Taking into consideration that $g^{o,ex}(\mathbf{x},\cdot)$ is partially concave on $\tilde{P}$ for each $\mathbf{x} \in X$ and similar to the linearization approaches already presented, the following linearization of (5.13) around an arbitrary point $\bar{\mathbf{p}} \in \tilde{P}$ furnishes an upper bound for (1.1):

$$\begin{aligned} &\min_{\mathbf{x}\in X} f(\mathbf{x}) \\ &\text{s.t. } g^{o,ex}(\mathbf{x},\bar{\mathbf{p}}) + \sum_{j=1}^{n_{\tilde{p}}} g^{o,ex}_{p_j}(\mathbf{x},\bar{\mathbf{p}})(\tilde{p}_j - \bar{p}_j) \le 0 \quad \forall \tilde{\mathbf{p}} \in \tilde{P}_e, \end{aligned}$$

where $\tilde{P}_e$ denotes the set of vertices of $\tilde{P}$. Recall that the set of vertices considered can be reduced if the functions can be demonstrated to be monotone with respect to some parameters.

5.4. Relaxation over X and P. The upper bounding methodologies that have been presented so far rely on creating a function g^o that is partially concave with respect to the parameters $\mathbf{p}$ pointwise for each $\mathbf{x} \in X$. Another way of creating a valid overestimator of g is to construct a jointly concave function $g^{o,j}$ on $X \times P$, i.e., with respect to both the variables $\mathbf{x}$ and the parameters $\mathbf{p}$, using either McCormick or αBB concave relaxation methods, that satisfies

$$g^{o,j}(\mathbf{x},\mathbf{p}) \ge g(\mathbf{x},\mathbf{p}) \quad \forall (\mathbf{x},\mathbf{p}) \in X \times P.$$

Then, the following SIP is a restriction of (1.1):

$$\begin{aligned} &\min_{\mathbf{x}\in X} f(\mathbf{x}) \\ &\text{s.t. } g^{o,j}(\mathbf{x},\mathbf{p}) \le 0 \quad \forall \mathbf{p} \in P. \end{aligned} \tag{5.14}$$

Note that for convergence both host sets (X and P) need to be refined.

5.4.1. Linearization. Similar to the linearization approaches that have been presented so far, and since $g^{o,j}$ is concave on $X \times P$, we can linearize (5.14) around an arbitrary interior point $(\bar{\mathbf{x}},\bar{\mathbf{p}}) \in \operatorname{int}(X \times P)$ to obtain the following restriction of (1.1):

$$\begin{aligned} &\min_{\mathbf{x}\in X} f(\mathbf{x}) \\ &\text{s.t. } g^{o,j}(\bar{\mathbf{x}},\bar{\mathbf{p}}) + \sum_{j=1}^{n_x} g^{o,j}_{x_j}(\bar{\mathbf{x}},\bar{\mathbf{p}})(x_j - \bar{x}_j) + \sum_{j=1}^{n_p} g^{o,j}_{p_j}(\bar{\mathbf{x}},\bar{\mathbf{p}})(p_j - \bar{p}_j) \le 0 \quad \forall \mathbf{p} \in P_e, \end{aligned} \tag{5.15}$$

where g_{x_j} and g_{p_j} denote partial derivatives or, with an abuse of notation, subgradients. By the separability and linearity of the constraint in (5.16), a single inequality constraint is needed; i.e., P_e can be replaced by a single point $\mathbf{p}^*$ calculated by the following procedure:

- **FOR** $j = 1,\dots,n_p$ **DO**
  - **IF** $g^{o,j}_{p_j}(\bar{\mathbf{x}},\bar{\mathbf{p}}) \le 0$ **THEN** $p^*_j = p^L_j$ **ELSE** $p^*_j = p^U_j$.

  END

Indeed, by construction of $\mathbf{p}^*$ for each $j = 1,\dots,n_p$, we have

$$\max_{\mathbf{p}\in P} g^{o,j}_{p_j}(\bar{\mathbf{x}},\bar{\mathbf{p}})(p_j - \bar{p}_j) = \max_{\mathbf{p}\in P^e} g^{o,j}_{p_j}(\bar{\mathbf{x}},\bar{\mathbf{p}})(p_j - \bar{p}_j) = g^{o,j}_{p_j}(\bar{\mathbf{x}},\bar{\mathbf{p}})(p^*_j - \bar{p}_j).$$

Therefore, we also obtain for each $\mathbf{x} \in X$

$$\begin{aligned}
&\max_{\mathbf{p}\in P} g^{o,j}(\bar{\mathbf{x}},\bar{\mathbf{p}}) + \sum_{j=1}^{n_x} g^{o,j}_{x_j}(\bar{\mathbf{x}},\bar{\mathbf{p}})(x_j - \bar{x}_j) + \sum_{j=1}^{n_p} g^{o,j}_{p_j}(\bar{\mathbf{x}},\bar{\mathbf{p}})(p_j - \bar{p}_j)\\
&= \max_{\mathbf{p}\in P^e} g^{o,j}(\bar{\mathbf{x}},\bar{\mathbf{p}}) + \sum_{j=1}^{n_x} g^{o,j}_{x_j}(\bar{\mathbf{x}},\bar{\mathbf{p}})(x_j - \bar{x}_j) + \sum_{j=1}^{n_p} g^{o,j}_{p_j}(\bar{\mathbf{x}},\bar{\mathbf{p}})(p_j - \bar{p}_j)\\
&= g^{o,j}(\bar{\mathbf{x}},\bar{\mathbf{p}}) + \sum_{j=1}^{n_x} g^{o,j}_{x_j}(\bar{\mathbf{x}},\bar{\mathbf{p}})(x_j - \bar{x}_j) + \sum_{j=1}^{n_p} g^{o,j}_{p_j}(\bar{\mathbf{x}},\bar{\mathbf{p}})(p^*_j - \bar{p}_j).
\end{aligned}$$

Therefore the following NLP with a single linear inequality constraint,

$$\begin{aligned}
&\min_{\mathbf{x}\in X} f(\mathbf{x})\\
&\text{s.t. } g^{o,j}(\bar{\mathbf{x}},\bar{\mathbf{p}}) + \sum_{j=1}^{n_x} g^{o,j}_{x_j}(\bar{\mathbf{x}},\bar{\mathbf{p}})(x_j - \bar{x}_j) + \sum_{j=1}^{n_p} g^{o,j}_{p_j}(\bar{\mathbf{x}},\bar{\mathbf{p}})(p^*_j - \bar{p}_j) \le 0,
\end{aligned}$$

provides a valid upper bound for (1.1).

This approach will obviously furnish looser upper bounds than the ones produced by the MPEC and linearization approaches that rely on the concave overestimation of g only with respect to the parameters $\mathbf{p}$. However, a single linear inequality is required, compared to the polynomial or exponential number of nonlinear constraints. Again, the choice of $\bar{\mathbf{p}}$ greatly affects the tightness of the proposed upper bound.

5.4.2. MPEC formulation. Similar to the MPEC approach that was described in section 5.2, a possible bounding problem is to replace the lower-level problem of (5.14) with its equivalent KKT conditions and solve the resulting problem to obtain an upper bound. Although this method would produce valid upper bounds, there are two distinct drawbacks compared to the MPEC approach that relies on concave relaxation of g only with respect to $\mathbf{p}$. First, the process of creating a concave overestimator of g on $X \times P$ will replace convex and nonconvex, with respect to $\mathbf{x}$, terms by concave ones which does not seem to simplify the solution of the resulting problem. Secondly, the generated relaxation will be weaker. Note that, even using αBB techniques, the value of α would be greater than or equal to the value of α that corresponds to the concave relaxation only on P because the Hessian increases in size. In conclusion, this method does not seem to produce either tighter bounds or simpler constraint expressions and will, therefore, not be analyzed further.

5.5. Convergence of upper bounding problems. The various alternatives described restrict the SIP (1.1) by overestimating the function $g(\mathbf{x},\cdot)$ pointwise in X. In the nonsmooth interpretation of SIP (1.2) the parametric optimal solution value of the lower-level program $g^U(\mathbf{x})$ is overestimated, obtaining

$$\begin{aligned}
&\min_{\mathbf{x}\in X} f(\mathbf{x})\\
&\text{s.t.} \quad g^{o,U}(\mathbf{x}) \le 0,
\end{aligned}$$

with $g^{o,U}(\mathbf{x}) \ge g^U(\mathbf{x})$. As described in section 3, this relaxation of the lower-level program leads to a restriction of the SIP. In general, this restriction excludes some feasible points and may even render the upper bounding problem infeasible. To ensure

that the upper bound converges to the optimal solution value, a subdivision of the parameter host set P, as in [13], is deemed necessary. As in the lower bounding problem, for the subdivision additional variables and/or constraints will be introduced. Methods for efficient convergence are outside the scope of this paper, and here we only briefly discuss basic convergence properties.

Similarly to the ICR of Bhattacharjee and coworkers [12, 13] and the proposal by Floudas and Stein [19], an exhaustive subdivision of the parameter set P leads to a pointwise convergence of $g^{o,U}$ to g^U. Therefore, points $\bar{\mathbf{x}}$ satisfying $\max_{\mathbf{p}\in P} g(\bar{\mathbf{x}}, \mathbf{p}) < 0$, i.e., SIP Slater points, become feasible in the upper bounding problems for a sufficiently fine subdivision. As a consequence, if the upper bounding problems are solved to global optimality and SIP Slater points exist arbitrarily close to a global minimum of (1.1), the upper bound converges to the optimal solution value.

6. Implementation and numerical results.

6.1. Implementation. The proposed lower and upper bounding problems potentially contain nonconvex objective function and/or constraints. Aiming to obtain the best possible bounds, we solve all the problems globally with BARON version 7.5 [44], available through GAMS version 22.1 [15], on a 64-bit Xeon 3.2GHz processor running Linux 2.6.13.

As is typical in NLP and MINLP solvers, BARON allows the violation of inequality and equality tolerances by a positive tolerance. For the lower bounding problem this is a further relaxation and thus of no concern, but it is a limitation for the upper bounding problems involving equality constraints. Note that the inequality constraint $g(\mathbf{x}, \mathbf{p}) \leq 0$ can be further restricted to $g(\mathbf{x}, \mathbf{p}) \leq -\varepsilon$, for an ε equal to the constraint violation of the NLP solver, and therefore does not pose a significant problem. To obtain good estimates we set the smallest possible value (10^{-9}) for the relevant tolerances (`conttol, boxtol, inttol`). The absolute and relative termination criteria, i.e., the difference between the lower and upper bounds in the subproblems, are set to 10^{-4}. Our previous numerical experiments with similar programs have shown slow convergence for problems involving third-order monomials, e.g., x^3, and for consistency purposes we systematically encode third-order monomials as a product of a square and a linear term, e.g., $x^2\, x$, and fourth-order monomials as the product of two squares, e.g., $x^2\, x^2$.

The complementary slackness conditions in the lower bounding problem are formulated using the big-M formulation, since binary variables are needed for the second-order conditions. In the upper bounding problem the complementary slackness conditions are left as nonlinear equations. For the discretization approach in the lower bounding problems we follow the heuristic in [13] and define the discretization set P_D as the upper right endpoints, i.e., $P_D = \{\mathbf{p}^U\}$. For the linearizations the midpoint of P is used.

Since the problems considered are relatively small, for the αBB relaxations we obtain the smallest possible α through the solution of a global optimization problem. This is done in the spirit of obtaining the tightest possible bounds. On the other hand, the bounds on the KKT multipliers and the second derivatives are estimated using the natural interval extensions capabilities of DAEPACK [53, 54]. For the MPEC-based upper bound using relaxation with extra variables, the upper bound for the KKT multipliers is set to 10^3. Note that overestimating the bounds for the multipliers typically increases the computational requirements (in CPU seconds) to solve the problems.

TABLE 6.1
Numerical results.

Problem		Lower bounds				Upper bounds						
label	f^*	P_D	KKT-1	KKT-2	P_D+KKT	5.2.1	5.2.2	5.3.1	5.3.2	5.3.3	5.4	ICR
2	0.194	**0.16**	0.08	**0.16**	**0.16**	$+\infty$	**0.194**	$+\infty$	0.28	**0.194**	50.58	0.38
5	4.30	3.54	3.72	3.72	**3.84**	20.2	**4.32**	27.7	4.64	**4.32**	7890	4.72
6	97.2	86.26	**97.2**	**97.2**	**97.2**	$+\infty$	**97.2**	$+\infty$	306	**97.2**	$+\infty$	**97.2**
7	1.00	0.0556	**1.00**	**1.00**	**1.00**	86.1	**1.00**	$+\infty$	1.60	**1.00**	$+\infty$	**1.00**
8	2.44	-7.14	-8.17	-8.17	**-4.16**	$+\infty$	**3.13**	$+\infty$	4.20	**3.13**	$+\infty$	7.39
9	-12	**-53.3**	**-53.3**	**-53.3**	**-53.3**	$+\infty$	**-12.0**	$+\infty$	**-12.0**	**-12.0**	$+\infty$	**-12.0**
N	0.00	-1.00	**0.00**	**0.00**	**0.00**	$+\infty$	**0.00**	$+\infty$	**0.00**	**0.00**	$+\infty$	**0.00**

TABLE 6.2
Computational requirements.

Problem	CPU time (s) lower bounds				CPU time (s) upper bounds						
label	P_D	KKT-1	KKT-2	P_D+KKT	5.2.1	5.2.2	5.3.1	5.3.2	5.3.3	5.4	ICR
2	0.03	0.13	0.2	0.05	0.01	0.48	0.01	0.02	0.02	0.01	0.01
5	0.02	0.12	0.11	0.06	0.13	0.54	0.02	0.02	0.02	0.01	0.01
6	0.03	0.06	0.05	0.03	0.05	0.67	0.01	0.03	0.04	0.01	0.04
7	0.02	0.76	0.82	0.36	0.28	0.18 (121)	0.01	0.02	0.02	0.01	0.02
8	0.01	0.35	0.56	0.16	0.07	0.01 (273)	0.01	0.02	0.02	0.01	0.01
9	0.01	0.03	0.02	0.05	3.30	0.41 (1000)	0.01	0.01	0.01	0.01	0.01
N	0.01	0.06	0.05	0.07	0.01	0.04	0.01	0.01	0.02	0.01	0.01

6.2. Numerical results. As a test set we use the well-established problems by Watson [57], summarized in Appendix A. Since BARON and DAEPACK currently do not support trigonometric functions, we only use those examples that do not involve trigonometric functions. For all problems we used $\mathbf{x} \in [-10, 10]^{n_x}$.

Tables 6.1 and 6.2 respectively contain the computational requirement as reported by BARON (through the GAMS attribute `resusd`) and the bounding values obtained. No distinction is made for times below 0.01s. In three cases (all KKT-based upper bounding problems) we distinguish between the time to find the optimal solution value and to confirm it (number in brackets) because the two computational requirements differ dramatically. The first column (label) has the label of the problem, while the column labeled f^* contains the best known solution for the problem. The following four columns show the results from the lower bounding problems. P_D denotes discretization, KKT-1 results by applying the necessary first-order KKT conditions, KKT-2 by applying the necessary first- and second-order KKT conditions, and P_D+KKT by combining the necessary first- and second-order KKT conditions with discretization. The next six columns contain the upper bounds obtained by our upper bounding proposals, labeled by the corresponding sections. The final column (ICR) is the interval constrained reformulation by Bhattacharjee, Green, and Barton [12]. To ensure that we have comparable solution times we reproduced the ICR from [12]. In Table 6.1 the best lower and upper bounds for each problem are highlighted by a bold font.

6.3. Conclusions from numerical experiments. The proposed lower and upper bounding problems bracket the optimal solution value quite successfully. Furthermore, the bounds furnished are often exact, and in some cases both the upper and the lower bounds are exact. The computational requirement for obtaining upper and lower bounds is quite low for the small-scale problems considered; note that for the case of a KKT-based upper bound with extra variables, the computational requirement for confirming the global solution is quite high for three problems involving two

parameters. Recall also from sections 4.4 and 5.5 and [12, 13, 19] that subdivision methods would tighten the bounds in the cases for which they are not exact.

The combination of discretization and KKT conditions produces tighter lower bounds than either of the methods taken alone. Moreover, the additional constraints accelerate convergence compared to when only the KKT conditions are used. Similarly the addition of the second-order consideration can help accelerate convergence and/or furnish a tighter bound.

As expected, the KKT-based upper bounds using extra variables (section 5.2.2) can be significantly tighter than the ICR-based, at the expense of a higher computational cost. Surprisingly, the linearization-based bounds using auxiliary variables (section 5.3.3) produced bounds as tight as the ones based on the KKT conditions; we believe that this is due to the problem structure. The bounds based on smooth relaxation without extra variables (sections 5.2.1 and 5.3.1) are relatively weak. We want to again point out that we deviated from the αBB relaxation described by Adjiman and Floudas [4] and considered the constraint as a whole. Note finally that the number of parameters in the problems considered is small ($n_p \in \{1, 2\}$), and therefore the effect of the exponential number of constraints in the linearization (sections 5.3.1, 5.3.2, and 5.3.3) is not apparent.

7. Conclusions and future work. We consider SIPs that involve nonconvex functions and present lower and upper bounding problems. For the lower bounding problem we combine and extend literature ideas based on discretization and the necessary KKT conditions of the lower-level program. The upper bounding problem is constructed based on a convex relaxation of the lower-level program which results in a restriction of the SIP, similarly to the ICR of Bhattacharjee and coworkers [12, 13]. The resulting lower-level programs are replaced by the sufficient KKT conditions or further relaxed by linearization. The proposed upper bounding methodology is more expensive than ICR, but often leads to tighter upper bounds. What the method of choice is will most likely depend on the problem size and structure. Therefore further experimentation with a large set of problems of various sizes is of interest. To that extent an automation of the convex relaxation such as the one proposed by Gatzke, Tolsma, and Barton [21] is deemed necessary. Furthermore, the comparison of different solvers is also of interest.

Many of the ideas proposed can be refined, such as the second-order check for the lower bound, which, currently, is relatively weak. Another example is to produce good heuristics for the parameter values around which to linearize. A third example is that for the αBB method we calculated α uniformly over X by taking interval extensions on $X \times P$; instead it would be interesting to consider a method similar to what we developed for the McCormick relaxations, and calculate different α for different subsets of X.

We considered global solution of all bounding problems, which is computationally expensive. For the lower bounding problem a further convex relaxation could be solved with a local solver. For the upper bounding problem local methods could be applied, but these may fail. For implementation within a branch-and-bound framework for the global solution of SIP there is a trade-off between expensive and tight versus cheap and loose bounds. A promising heuristic is to solve these problems with a global solver such as BARON [44] and a loose termination criterion.

It seems very promising to combine some of the ideas presented. A simple combination is to periodically employ the tighter and more expensive bounding problems. A more elaborate combination is to use a KKT-based upper bounding problem but

solve the resulting MPEC only approximately to obtain a point $\bar{\mathbf{x}}$ and an estimate for the corresponding optimal solution of the relaxed lower-level problem $\bar{\mathbf{p}}$. Then the feasibility of $\bar{\mathbf{x}}$ can be probed by linearizing the concave lower-level problem around $\bar{\mathbf{p}}$. This approach is difficult to implement with black-box NLP solvers, but could be easily implemented in a framework such as the NCP approach by Floudas and Stein [19]. The promise of the combination is that an approximate solution of the MPEC will provide a point $\bar{\mathbf{p}}$ which is suitable for linearization.

In this paper we consider a single semi-infinite constraint. An interesting extension is to consider many semi-infinite constraints, for which several alternatives for extending our proposals could be applied. One possibility is to consider many lower-level programs by the introduction of as many sets of parameters $\mathbf{p}^i$ as there are constraints. Another alternative is to introduce an SOS-1 set of binary variables

$$\mathbf{y} \in \{0,1\}^{n_g}, \qquad \sum_{i=1}^{n_g} y_i = 1, \qquad y_j = 1 \Leftrightarrow g_i(\mathbf{x},\mathbf{p}) \geq \max_i g_i(\mathbf{x},\mathbf{p}),$$

where n_g is the number of constraints. Then a logical constraint

$$(y_j = 1) \Rightarrow \mathbf{p} \text{ is a KKT point of } g_j$$

can be implemented.

In this paper we assumed that the parameter host set is explicitly given as a box. An interesting extension is to consider general host sets. Another interesting extension is to generalized semi-infinite constraints. In a forthcoming publication [27] we analyze which of the proposed methods can be extended to GSIP and provide numerical results from an extensive test set.

In discretization approaches tightening of lower bounds can be achieved by gradually increasing the cardinality of the finite set of parameters considered, and therefore increasing the number of constraints [13]. Bhattacharjee et al. [13] employed a subdivision approach to tighten the upper bound, which also leads to an increase in the number of constraints. For our proposed lower and upper bounding problems a subdivision of P is also deemed necessary for convergence, and this will lead to an increase in the number of constraints. Moreover, for the bounding problems based on MPEC, an increase in the number of variables is also expected, making the solution of the MPEC increasingly expensive for finer subdivisions. A possible heuristic is to include the KKT-based upper bound only for a subset of the sets P^i and to use the interval extensions for the rest.

Appendix A. Test set. For consistency purposes we use the problem labels of Watson [57].

2.

$$\begin{aligned} P &= [0,1], \\ f(\mathbf{x}) &= \frac{1}{3}x_1^2 + x_2^2 + \frac{1}{2}x_1, \\ g(\mathbf{x},p) &= (1 - x_1^2\,p^2)^2 - x_1\,p^2 - x_2^2 + x_2. \end{aligned}$$

5.

$$\begin{aligned} P &= [0,1], \\ f(\mathbf{x}) &= e^{x_1} + e^{x_2} + e^{x_3}, \\ g(\mathbf{x},p) &= \frac{1}{1+p^2} - x_1 - x_2\,p - x_3\,p^2. \end{aligned}$$

6.

$$\begin{aligned} P &= [0,1], \\ f(\mathbf{x}) &= (x_1 - 2\,x_2 + 5\,x_2^2 - x_2^2\,x_2 - 13)^2 + (x_1 - 14x_2 + x_2^2 + x_2^3 - 29)^2, \\ g(\mathbf{x},p) &= x_1^2 + 2\,x_2\,p^2 + e^{x_1+x_2} - e^p. \end{aligned}$$

Note that in [12] the exponent is missing in the first term of the objective function.

7.

$$\begin{aligned} P &= [0,1]^2, \\ f(\mathbf{x}) &= x_1^2 + x_2^2 + x_3^2, \\ g(\mathbf{x},\mathbf{p}) &= x_1\,(p_1 + p_2^2 + 1) + x_2\,(p_1\,p_2 - p_2^2) + x_3\,(p_1\,p_2 + p_2^2 + p_2) + 1. \end{aligned}$$

8.

$$\begin{aligned} P &= [0,1]^2, \\ f(\mathbf{x}) &= x_1 + \frac{1}{2}x_2 + \frac{1}{2}x_3 + \frac{1}{3}x_4 + \frac{1}{4}x_5 + \frac{1}{3}x_6, \\ g(\mathbf{x},\mathbf{p}) &= e^{p_1^2+p_2^2} - x_1 - x_2\,p_1 - x_3\,p_2 - x_4\,p_1^2 - x_5\,p_1\,p_2 - x_6\,p_2^2. \end{aligned}$$

Note that presumably in Watson's collection [57] the coefficient of x_4 in the objective function is mistyped. This is suggested by the optimal solution value reported in [57] and by the symmetry of the problem with respect to the variables x_4 and x_6.

9.

$$\begin{aligned} P &= [-1,1]^2, \\ f(\mathbf{x}) &= -4\,x_1 - \frac{2}{3}(x_4 + x_6), \\ g(\mathbf{x},\mathbf{p}) &= x_1 + x_2\,p_1 + x_3\,p_2 + x_4\,p_1^2 + x_5\,p_1\,p_2 + x_6\,p_2^2 - 3 - (p_1 - p_2)^2\,(p_1 + p_2)^2. \end{aligned}$$

N.

$$\begin{aligned} P &= [-1,1], \\ f(\mathbf{x}) &= x_2, \\ g(\mathbf{x},p) &= 2\,x_1^2\,p^2 - p^4 + x_1^2 - x_2. \end{aligned}$$

Appendix B. McCormick relaxations of the product of two functions. In this section we will show the convex and concave relaxations for a product of two functions $g_1(\mathbf{z})g_2(\mathbf{z})$ on $Z \subset \mathbb{R}^{n_z}$; compare also the treatment of trilinear terms in [29].

Assume that there exist convex functions g_1^u and g_2^u and concave functions g_1^o and g_2^o that satisfy

$$\begin{aligned} g_1^u(\mathbf{z}) \le g_1(\mathbf{z}) \le g_1^o(\mathbf{z}) \quad \forall \mathbf{z} \in Z, \\ g_2^u(\mathbf{z}) \le g_2(\mathbf{z}) \le g_2^o(\mathbf{z}) \quad \forall \mathbf{z} \in Z. \end{aligned}$$

Furthermore, let $G_1^L, G_1^U, G_2^L, G_2^U$ satisfy

$$\begin{aligned} G_1^L \le g_1(\mathbf{z}) \le G_1^U \quad \forall \mathbf{z} \in Z, \\ G_2^L \le g_2(\mathbf{z}) \le G_2^U \quad \forall \mathbf{z} \in Z. \end{aligned}$$

Then using the following definitions,

$$\alpha_1(\mathbf{z}) = \begin{cases} G_2^L g_1^u(\mathbf{z}) & \text{if } G_2^L \ge 0, \\ G_2^L g_1^o(\mathbf{z}) & \text{otherwise,} \end{cases}$$

$$\alpha_2(\mathbf{z}) = \begin{cases} G_1^L g_2^u(\mathbf{z}) & \text{if } G_1^L \ge 0, \\ G_1^L g_2^o(\mathbf{z}) & \text{otherwise,} \end{cases}$$

$$\beta_1(\mathbf{z}) = \begin{cases} G_2^U g_1^u(\mathbf{z}) & \text{if } G_2^U \ge 0, \\ G_2^U g_1^o(\mathbf{z}) & \text{otherwise,} \end{cases}$$

$$\beta_2(\mathbf{z}) = \begin{cases} G_1^U g_2^u(\mathbf{z}) & \text{if } G_1^U \ge 0, \\ G_1^U g_2^o(\mathbf{z}) & \text{otherwise,} \end{cases}$$

$$\gamma_1(\mathbf{z}) = \begin{cases} G_2^L g_1^u(\mathbf{z}) & \text{if } G_2^L \le 0, \\ G_2^L g_1^o(\mathbf{z}) & \text{otherwise,} \end{cases}$$

$$\gamma_2(\mathbf{z}) = \begin{cases} G_1^U g_2^u(\mathbf{z}) & \text{if } G_1^U \le 0, \\ G_1^U g_2^o(\mathbf{z}) & \text{otherwise,} \end{cases}$$

$$\delta_1(\mathbf{z}) = \begin{cases} G_2^U g_1^u(\mathbf{z}) & \text{if } G_2^U \le 0, \\ G_2^U g_1^o(\mathbf{z}) & \text{otherwise,} \end{cases}$$

$$\delta_2(\mathbf{z}) = \begin{cases} G_1^L g_2^u(\mathbf{z}) & \text{if } G_1^L \le 0, \\ G_1^L g_2^o(\mathbf{z}) & \text{otherwise,} \end{cases}$$

valid convex and concave (g^u and g^o) relaxations of g on Z are given by

$$\begin{aligned} g^u(\mathbf{z}) \ge \max\{\alpha_1(\mathbf{z}) + \alpha_2(\mathbf{z}) - G_1^L G_2^L, \beta_1(\mathbf{z}) + \beta_2(\mathbf{z}) - G_1^U G_2^U\}, \\ g^o(\mathbf{z}) \le \min\{\gamma_1(\mathbf{z}) + \gamma_2(\mathbf{z}) - G_1^U G_2^L, \delta_1(\mathbf{z}) + \delta_2(\mathbf{z}) - G_1^L G_2^U\}. \end{aligned}$$

Acknowledgments. We would like to thank Benoît Chachuat for fruitful discussions. We are also grateful to the anonymous reviewers for the thorough analysis and detailed comments that considerably improved this paper.

REFERENCES

[1] C. S. Adjiman, I. P. Androulakis, and C. A. Floudas, *A global optimization method, αBB, for general twice-differentiable constrained NLPs—*II. *Implementation and computational results*, Computers & Chemical Engineering, 22 (1998), pp. 1159–1179.

[2] C. S. Adjiman, I. P. Androulakis, C. D. Maranas, and C. A. Floudas, *A global optimization method, αBB, for process design*, Computers & Chemical Engineering, 20 (1996), pp. S419–S424.

[3] C. S. Adjiman, S. Dallwig, C. A. Floudas, and A. Neumaier, *A global optimization method, αBB, for general twice-differentiable constrained NLPs—*I. *Theoretical advances*, Computers & Chemical Engineering, 22 (1998), pp. 1137–1158.

[4] C. S. Adjiman and C. A. Floudas, *Rigorous convex underestimators for general twice-differentiable problems*, J. Global Optim., 9 (1996), pp. 23–40.

[5] I. G. Akrotirianakis and C. A. Floudas, *A new class of improved convex underestimators for twice continuously differentiable constrained NLPs*, J. Global Optim., 30 (2004), pp. 367–390.

[6] G. Alefeld and G. Mayer, *Interval analysis: Theory and applications*, J. Comput. Appl. Math., 121 (2000), pp. 421–464.

[7] J. F. Bard, *Practical Bilevel Optimization: Algorithms and Applications*, Nonconvex Optim. Appl., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.

[8] P. I. Barton, *Subgradients for McCormick Relaxations*, technical report, Massachusetts Institute of Technology, Cambridge, MA, 2006; available online from http://yoric.mit.edu/download/Reports/barton06mc.pdf.

[9] A. A. Ben-Tal and A. Nemirovski, *Robust optimization—Methodology and applications*, Math. Programming, 92 (2002), pp. 453–480.

[10] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed., Athena Scientific, Belmont, MA, 1999.

[11] B. Bhattacharjee, *Kinetic Model Reduction Using Integer and Semi-Infinite Programming*, Ph.D. thesis, Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, 2003.

[12] B. Bhattacharjee, W. H. Green, Jr., and P. I. Barton, *Interval methods for semi-infinite programs*, Comput. Optim. Appl., 30 (2005), pp. 63–93.

[13] B. Bhattacharjee, P. Lemonidis, W. H. Green, Jr., and P. I. Barton, *Global solution of semi-infinite programs*, Math. Program., 103 (2005), pp. 283–307.

[14] J. W. Blankenship and J. E. Falk, *Infinitely constrained optimization problems*, J. Optim. Theory Appl., 19 (1976), pp. 261–281.

[15] A. Brooke, D. Kendrick, and A. Meeraus, *GAMS: A User's Guide*, The Scientific Press, Redwood City, CA, 1988.

[16] I. D. Coope and C. J. Price, *Exact penalty function methods for nonlinear semi-infinite programming*, in Semi-Infinite Programming, R. Reemtsen and J. J. Rückmann, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 137–157.

[17] I. D. Coope and G. A. Watson, *A projected Lagrangian algorithm for semi-infinite programming*, Math. Programming, 32 (1985), pp. 337–356.

[18] V. Dua, K. P. Papalexandri, and E. N. Pistikopoulos, *Global optimization issues in multiparametric continuous and mixed-integer optimization problems*, J. Global Optim., 30 (2004), pp. 59–89.

[19] C. A. Floudas and O. Stein, *The adaptive convexification algorithm: A feasible point method for semi-infinite programming*, SIAM J. Optim., 18 (2007), pp. 1187–1208.

[20] J. Fortuny-Amat and B. McCarl, *A representation and economic interpretation of a two-level programming problem*, J. Oper. Res. Soc., 32 (1981), pp. 783–792.

[21] E. P. Gatzke, J. E. Tolsma, and P. I. Barton, *Construction of convex function relaxations using automated code generation techniques*, Optim. Engrg., 3 (2002), pp. 305–326.

[22] A. M. Geoffrion and R. Nauss, *Parametric and postoptimality analysis in integer linear programming*, Management Sci., 23 (1977), pp. 453–466.

[23] M. A. Goberna and M. A. Lopez, *Linear Semi-Infinite Optimization*, John Wiley and Sons, New York, 1998.

[24] S. Görner, A. Potchinkov, and R. Reemtsen, *The direct solution of nonconvex nonlinear FIR filter design problems by a SIP method*, Optim. Engrg., 1 (2000), pp. 123–154.

[25] R. Hettich, *An implementation of a discretization method for semi-infinite programming*, Math. Programming, 34 (1986), pp. 354–361.

[26] R. Hettich and K. O. Kortanek, *Semi-infinite programming: Theory, methods, and applications*, SIAM Rev., 35 (1993), pp. 380–429.

[27] P. Lemonidis, A. Mitsos, and P. I. Barton, *Relaxation-based Bounds for Generalized Semi-infinite Programs*, manuscript, 2006.

[28] Z. Liu and Y. H. Gong, *Semi-infinite quadratic optimization method for the design of robust adaptive array processors*, IEEE Proc. Radar and Signal Processing, 137 (1990), pp. 177–182.

[29] C. D. Maranas and C. A. Floudas, *Finding all solutions of nonlinearly constrained systems of equations*, J. Global Optim., 7 (1995), pp. 143–182.

[30] G. P. McCormick, *Computability of global solutions to factorable nonconvex programs: Part I. Convex underestimating problems*, Math. Programming, 10 (1976), pp. 147–175.

[31] A. Mitsos and P. I. Barton, *Issues in the Development of Global Optimization Algorithms for Bilevel Programs with a Nonconvex Inner Program*, techical report, Massachusetts Institute of Technology, Cambridge, MA, 2006; available online from http://yoric.mit.edu/download/Reports/bilevelissues.pdf.

[32] R. E. Moore, *Methods and Applications of Interval Analysis*, SIAM Stud. Appl. Math. 2, SIAM, Philadelphia, 1979.

[33] M. Nees, *Chebyshev approximation by discrete superposition. Application to neural networks*, Adv. Comput. Math., 5 (1996), pp. 137–151.

[34] S. Nordebo and Z. Q. Zang, *Semi-infinite linear programming: A unified approach to digital filter design with time- and frequency-domain specifications*, IEEE Trans. Circuits Systems II Analog Digital Signal Process., 46 (1999), pp. 765–775.

[35] Y. Ohtake and N. Nishida, *A branch-and-bound algorithm for* 0–1 *parametric mixed integer programming*, Oper. Res. Lett., 4 (1985), pp. 41–45.

[36] O. O. Oluwole, B. Bhattacharjee, J. E. Tolsma, P. I. Barton, and W. H. Green, Jr., *Rigorous valid ranges for optimally reduced kinetic models*, Combustion and Flame, 146 (2006), pp. 348–365.

[37] E. R. Panier and A. L. Tits, *A globally convergent algorithm with adaptively refined discretization for semi-infinite optimization problems arising in engineering design*, IEEE Trans. Automat. Control, 34 (1989), pp. 903–908.

[38] E. Polak and D. M. Stimler, *Majorization—A computational-complexity reduction technique in control-system design*, IEEE Trans. Automat. Control, 33 (1988), pp. 1010–1021.

[39] C. J. Price and I. D. Coope, *An exact penalty-function algorithm for semi-infinite programs*, BIT, 30 (1990), pp. 723–734.

[40] C. J. Price and I. D. Coope, *Numerical experiments in semi-infinite programming*, Comput. Optim. Appl., 6 (1995), pp. 169–189.

[41] R. R. Reemtsen and S. Görner, *Numerical methods for semi-infinite programming: A survey*, in Semi-Infinite Programming, R. Reemtsen and J. J. Rückmann, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 195–275.

[42] H. Ratschek and J. Rokne, *Computer Methods for the Range of Functions*, Ellis Horwood Ser. Math. Appl., Halsted Press, New York, 1984.

[43] R. Reemtsen, *Discretization methods for the solution of semi-infinite programming-problems*, J. Optim. Theory Appl., 71 (1991), pp. 85–103.

[44] N. Sahinidis and M. Tawarmalani, *BARON* (software manual), available online at http://www.gams.com/solvers/baron.pdf, 2005.

[45] A. B. Singer and P. I. Barton, *Global optimization with nonlinear ordinary differential equations*, J. Global Optim., 34 (2006), pp. 159–190.

[46] E. M. B. Smith and C. C. Pantelides, *A symbolic reformulation/spatial branch-and-bound algorithm for the global optimisation of nonconvex MINLPs*, Computers & Chemical Engineering, 23 (1999), pp. 457–478.

[47] O. Stein and G. Still, *On generalized semi-infinite optimization and bilevel optimization*, European J. Oper. Res., 142 (2002), pp. 444–462.

[48] O. Stein and G. Still, *Solving semi-infinite optimization problems with interior point techniques*, SIAM J. Control Optim., 42 (2003), pp. 769–788.

[49] G. Still, *Discretization in semi-infinite programming: The rate of convergence*, Math. Programming, 91 (2001), pp. 53–69.

[50] Y. Tanaka, M. Fukushima, and T. Ibaraki, *A globally convergent SQP method for semi-infinite nonlinear optimization*, J. Comput. Appl. Math., 23 (1988), pp. 141–153.

[51] M. Tawarmalani and N. V. Sahinidis, *Convexification and Global Optimization in Continuous and Mixed-Integer Nonlinear Programming*, Nonconvex Optim. Appl., Kluwer Academic Publishers, Boston, 2002.

[52] M. Tawarmalani and N. V. Sahinidis, *Global optimization of mixed-integer nonlinear programs: A theoretical and computational study*, Math. Program., 99 (2004), pp. 563–591.

[53] J. Tolsma and P. I. Barton, *DAEPACK: An open modeling environment for legacy models*, Industrial and Engineering Chemistry Research, 39 (2000), pp. 1826–1839.

[54] J. E. Tolsma and P. I. Barton, *DAEPACK: A Symbolic and Numeric Library*, website at http://yoric.mit.edu/daepack/daepack.html.

[55] A. I. F. Vaz, E. M. G. P. Fernandes, and M. P. S. F. Gomes, *SIPAMPL: Semi-infinite programming with AMPL*, ACM Trans. Math. Software, 30 (2004), pp. 47–61.
[56] X. J. Wang and T. S. Chang, *An improved univariate global optimization algorithm with improved linear lower bounding functions*, J. Global Optim., 8 (1996), pp. 393–411.
[57] G. Watson, *Numerical experiments with globally convergent methods for semi-infinite programming problems*, Lecture Notes in Econ. Math. Systems, 215 (1983), pp. 193–205.
[58] S. Zakovic, B. Rustem, and S. P. Asprey, *A parallel algorithm for semi-infinite programming*, Comput. Statist. Data Anal., 44 (2003), pp. 377–390.

SIAM J. OPTIM.
Vol. 19, No. 1, pp. 114–143

PRIMAL-DUAL AFFINE SCALING INTERIOR POINT METHODS FOR LINEAR COMPLEMENTARITY PROBLEMS*

FLORIAN A. POTRA†

Abstract. A first order affine scaling method and two mth order affine scaling methods for solving monotone linear complementarity problems (LCPs) are presented. All three methods produce iterates in a wide neighborhood of the central path. The first order method has $O(nL^2(\log nL^2)(\log\log nL^2))$ iteration complexity. If the LCP admits a strict complementary solution, then both the duality gap and the iteration sequence converge superlinearly with Q-order two. If $m = \Omega(\log(\sqrt{n}L))$, then both higher order methods have $O(\sqrt{n})L$ iteration complexity. The Q-order of convergence of one of the methods is $(m+1)$ for problems that admit a strict complementarity solution, while the Q-order of convergence of the other method is $(m+1)/2$ for general monotone LCPs.

Key words. linear complementarity, interior point, affine scaling, large neighborhood, superlinear convergence

AMS subject classifications. 90C51, 65K05, 49M15, 90C05, 90C20

DOI. 10.1137/060670341

1. Introduction.

The primal-dual affine scaling direction plays a special role in the theory and practice of interior point methods. It turns out that the search direction used by most primal-dual interior point methods is a convex combination $(1-\gamma)w + \gamma\overline{w}$ of the primal-dual affine scaling method w and the centering direction $\overline{w}$ (see the monographs [35, 48, 49]). Optimality is improved along the affine scaling direction, while centrality is improved on the centering direction. The first interior point method of this type was proposed, in the context of linear programming (LP), by Kojima, Mizuno, and Yoshise [15]. They proved that the algorithm had $O(nL)$ iteration complexity, the same as Karmarkar's algorithm [12]. Shortly after that, they improved the algorithm and generalized it for monotone linear complementarity problems (LCP) [14]. The improved algorithm has $O(\sqrt{n}L)$ iteration complexity. This iteration complexity was first obtained by Renegar [34] for an interior point method that follows the primal central path of LP, and it remains the best iteration complexity known to date. The algorithm of [14] follows the primal-dual central path. Starting with a point $z^0 \in \mathcal{N}_2(\alpha)$ in a (small) neighborhood of the primal-dual central path, the algorithm takes at each iteration a unit step along the direction $(1-\gamma)w + \gamma\overline{w}$, where $\gamma = 1 - \alpha/\left((1-\alpha)\sqrt{n}\right)$, producing a sequence of iterates $\left(z^k\right)$ that remains in $\mathcal{N}_2(\alpha)$. A similar algorithm was independently proposed and investigated by Monteiro and Adler for LP [21], and for quadratic programming (QP) [22].

While the algorithms mentioned above attain the best known iteration complexity, their practical performance is not satisfactory because they generate points in a small neighborhood of the central path and use a fixed stepsize (of one) along a direction that is dominated by the centering direction. In order to alleviate some of these problems, Kojima, Mizuno, and Yoshise [16] proposed a potential reduction algorithm for LCPs

*Received by the editors September 20, 2006; accepted for publication (in revised form) August 27, 2007; published electronically February 6, 2008. This work was supported by the National Science Foundation under grant 0139701, and by the National Institute of Health under grant R01GM075298-01.

http://www.siam.org/journals/siopt/19-1/67034.html

†Department of Mathematics and Statistics, University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore, MD 22150 (potra@math.umbc.edu).

with $O(\sqrt{n}L)$ iteration complexity. The search direction is of the same type with $\gamma = 1/(1+\sqrt{n})$. However, instead of taking a fixed stepsize along this direction, a line search is performed in order to ensure that the iterates remain strictly feasible, and that the Tanabe–Todd–Ye potential function [43, 44] is reduced by a fixed constant at each iteration.

Although the potential reduction method of [16] is more flexible than the short step path following methods of [14, 21, 22], it cannot attain superlinear convergence since its search direction has a fixed centering component. There are basically three ways to obtain superlinear convergence: to take $\gamma = 0$ every second step, to use a sequence of γ's that converges to zero, or to take $\gamma = 0$ at every step. The methods from the first category are called predictor-corrector methods, while the methods from the last category are called affine scaling methods and form the subject of the present paper.

Zhang, Tapia, and Dennis [52] gave sufficient conditions for a class of interior point methods to produce a sequence of iterates with duality gap converging superlinearly to zero. However, no example of an algorithm satisfying those conditions and having polynomial complexity was given. The results of [52] were generalized for LCP in [53]. The first interior point method having both polynomial complexity and superlinear convergence was the predictor-corrector method of Mizuno, Todd, and Ye (MTY). This method was proposed for LP in [20], where it was shown to have $O(\sqrt{n}L)$ iteration complexity. Shortly after that, Ye et al. [51] proved that the duality gap of the iterates produced by MTY converges quadratically to zero. MTY was generalized to LCP in [11], and the resulting algorithm was proved to have $O(\sqrt{n}L)$ iteration complexity under general conditions, and superlinear convergence under the assumption that the LCP has a (perhaps not unique) strictly complementary condition (i.e., the LCP is nondegenerate) and the iteration sequence converges. From [3] it follows that the latter assumption always holds. Subsequently, Ye and Anstreicher [50] proved that MTY converges quadratically, assuming only that the LCP is nondegenerate. The nondegeneracy assumption is not restrictive, since according to [24] a large class of interior point methods, which contains MTY, can have only linear convergence if this assumption is violated.

The largest step path following method (LSPF) determines γ at each iteration such that the duality gap is minimized under the constraint that the point obtained by taking a unit step along the search direction $(1-\gamma)w + \gamma\overline{w}$ remains in $\mathcal{N}_2(\alpha)$. McShane [17] proved that LSPF has $O(\sqrt{n}L)$ iteration complexity for general (monotone) LCPs and superlinear convergence under the assumption that the LCP is nondegenerate and the iteration sequence converges. The superlinear convergence is a consequence of the fact that the γ's converge to zero. Gonzaga [6] proved superlinear convergence assuming only nondegeneracy and showed that, with the addition of a computationally trivial safeguard, LSPF achieves Q-quadratic convergence. We mention that the convergence of the iteration sequence generated by LSPF follows from the general results of Bonnans and Gonzaga [3].

It took considerable effort to obtain superlinear convergence results for potential reduction algorithms. One of the first results of this type was obtained for LP by Tunçel [46]. His algorithm uses a direction of the form $(1-\gamma)w + \gamma\overline{w}$, where γ converges to zero at the same rate as the duality gap. The stepsize along this direction is obtained by minimizing a special, nonsmooth, potential function. The algorithm has $O(nL)$ iteration complexity for general LPs and Q-quadratic convergence of the duality gap under the assumption that the LP is nondegenerate. This

assumption is rather strong since it implies that the Jacobian of the corresponding LCP is nonsingular at the solution. This strong restriction was removed by Tütüncü [47], who proposed a potential reduction method with $O(nL)$ iteration complexity and Q-quadratic convergence of both the duality gap and the iteration sequence for general LPs. The search direction is the reduced Newton direction of a multiplicative variant of the Tanabe–Todd–Ye potential function, which is not exactly of the form $(1-\gamma)w+\gamma\overline{w}$, but it is shown that it asymptotically approaches the affine scaling direction, thus ensuring superlinear convergence. The algorithm uses the largest stepsize along this direction such that the feasibility of the iterates is maintained and the original Tanabe–Todd–Ye potential function is reduced by a fixed quantity at each iteration.

All the superlinear convergence results mentioned above are obtained either for LP, which always has a strictly complementary solution, or for LCPs that admit such a solution. As mentioned before, a large class of first order interior point methods cannot have superlinear convergence without this assumption [24]. In order to obtain superlinear convergence for degenerate LCPs one has to either use nonstandard interior point methods or to consider higher order methods. The first approach was first taken by Mizuno [18], who used the so-called step variant of the Tapia indicator [5] to identify the variables that are not strictly complementary, and modified the MTY algorithm in order to accelerate the convergence to zero of those variables. Mizuno's result was refined in [33, 32]. The drawback of this approach is that it needs the estimation of the partition of the optimal face in order to achieve superlinear convergence. The second approach was taken by Sturm [42], who constructed a second order MTY-type algorithm that has $O(\sqrt{n}L)$ iteration complexity and Q-superlinear convergence of order 1.5 for degenerate LCPs. By using mth order derivatives, Stoer, Wechs, and Mizuno [41] constructed higher order MTY-type algorithms with Q-order $m+1$ for nondegenerate LCPs and $(m+1)/2$ for degenerate LCPs. The complexity of the predictor-corrector algorithm for degenerate LCPs from [41] is analyzed in [38]. It follows that for monotone LCPs with feasible starting points the algorithm has $O(\sqrt{n}L)$ iteration complexity.

The path following interior point algorithms described in the paragraph above use a small neighborhood of the central path. However, extensive numerical experiments show convincingly that predictor-corrector methods perform better when using large neighborhoods of the central path. Predictor-corrector methods of MTY type are more difficult to develop and analyze in such neighborhoods, since the correctors are rather inefficient there. For example, it is known that one needs $O(n)$ corrector steps in order to reduce the δ_∞ proximity measure by a factor of .5 (see [2]). Therefore a straightforward generalization of the MTY algorithm would have $O(n^{1.5}L)$ iteration complexity. Gonzaga [7] proposed an interior point method in the $\mathcal{N}_\infty(\alpha)$ neighborhood of the central path, where a predictor is followed by an a priori unknown number of correctors, with $O(nL)$-iteration complexity [7]. No superlinear convergence results are known for this algorithm. An interior point method for LP, acting in a large neighborhood of the central path defined by a self-regular proximity measure, with $O(\sqrt{n}L\log n)$ iteration complexity and superlinear convergence, was proposed in [26]. The first order predictor-corrector method from [30] uses the wide $\mathcal{N}_\infty^-(\alpha)$ neighborhood, takes one predictor followed by exactly one corrector at each iteration, and has $O(nL)$ iteration complexity for general monotone LCPs and Q-quadratic convergence for nondegenerate LCPs. The mth order predictor-corrector methods of [30] also use the $\mathcal{N}_\infty^-(\alpha)$ neighborhood, and their Q-order of convergence

is $m+1$ for nondegenerate LCPs and $(m+1)/2$ for degenerate LCPs. If $m = n^{\omega}$, for some $\omega > 0$, then they have $O(\sqrt{n}L)$ iteration complexity. We note that the mth order method uses two matrix factorizations and m backsolves per iteration, so that if $\omega < 1$, then the computational cost per iteration is dominated by the cost of the two matrix factorizations. The results of [30] were generalized for sufficient LCPs in [31]. An mth order interior point algorithm for sufficient linear complementarity problems in the $\mathcal{N}_{\infty}^{-}(\alpha)$ neighborhood, that requires only one matrix factorization and m backsolves per iteration, was proposed by Stoer [37]. The Q-order of convergence is m for nondegenerate problems. The algorithm is modified in [36] so that its asymptotic Q-order of convergence is $m+1$ for nondegenerate LCPs and $(m+1)/2$ for degenerate LCPs. No complexity results have been proved for the algorithms in [36, 37].

Primal-dual affine scaling methods use the pure primal-dual affine scaling method at each iteration. The first method of this type was proposed by Monteiro, Adler, and Resende [23]. Their algorithm takes a small fixed step (of length $1/(nL)$) along this direction and has $O(nL^2)$ iteration complexity. Because of the fixed stepsize, the duality gap is reduced only by a factor of $(1 - 1/(nL))$ at each iteration, and therefore the algorithm cannot be superlinearly convergent. Moreover, only the first nL^2 iterates are guaranteed to be feasible, so that the algorithm may not produce an infinite sequence. In the same paper, Monteiro, Adler, and Resende [23] propose an mth order affine scaling method for QP with $O(n^{\frac{m+1}{2m}} L^{\frac{m+1}{2m}})$ iteration complexity. They again use a small fixed stepsize (depending on n, m, and L) so that the algorithm cannot have superlinear convergence. In order to overcome the inherent inefficiency caused by the use of a fixed stepsize, Mizuno and Nagasawa [19] proposed an affine scaling method for LP, where at each iteration the stepsize is determined by a line search that ensures the feasibility of the iterates and the fact that the Tanabe–Todd–Ye potential function remains bounded. By choosing the parameter defining the potential function equal to $1/L$, they obtain $O(nL^2)$ iteration complexity. However, no superlinear convergence results are obtained. Tunçel [45] analyzes a variant of this affine scaling where Tunçel's potential function [46] is used instead of the Tanabe–Todd–Ye potential function. The iteration complexity of the resulting algorithm is again $O(nL^2)$ when the parameter δ defining this potential function is equal to $1/(2L)$. The first affine scaling method with polynomial complexity and superlinear convergence was obtained, to our knowledge, by Monteiro and Wright [25] for LCPs. The stepsize along the affine scaling direction is obtained in such a way that the iterates are feasible and remain in neighborhood of the central path depending on two parameters η and δ. This neighborhood, which is closely related to Tunçel's potential function, is relatively narrow for large values of the duality gap, but it widens considerably as the duality gap approaches zero. It is shown that by choosing $\delta = \Theta(1/L)$, one obtains $O(nL^2)$ iteration complexity for monotone LCPs. It is shown that if the LCP has a strictly complementary solution, then the Q-order of convergence of the duality gap is $2 - \delta\xi$, where $\xi = 1$ for skew-symmetric problems and $\xi = 2$ otherwise. As a byproduct Monteiro and Wright show that Tunçel's affine scaling method [45] for LP has Q-order $2 - \delta$.

We note that the primal-dual affine scaling direction considered in this paper is sometime called the classical primal-dual affine scaling direction, in order to distinguish it from the so-called Dikin primal-dual affine scaling direction considered by Jansen, Roos, and Terlaky [8] for LP and generalized for LCP in [9]. The corresponding Dikin-type primal-dual affine scaling algorithms have $O(nL)$ iteration complexity. By using an mth order Dikin primal-dual affine scaling direction, Jansen et al. [10]

have obtained an algorithm with $O(n^{\frac{m+1}{2m}}L)$ iteration complexity. Unlike the algorithms from [8, 9], which used a small fixed stepsize, the algorithm from [10] may take a larger stepsize, which is determined at each iteration such that the new point remains in the $\mathcal{N}_\infty(\alpha)$ neighborhood of the central path. No superlinear results have been proved for Dikin-type primal-dual affine scaling algorithms. In fact, since the Dikin primal-dual affine scaling direction contains an important centering component, no superlinear convergence is expected for such algorithms.

In the present paper we present an affine scaling algorithm that uses the $\mathcal{N}_\infty^-(\alpha)$ neighborhood of the central path. It has $O(nL^2\phi(nL^2))$ iteration complexity for general monotone LCPs, where $\phi(t) = (\log t)(\log\log t)$. If the LCP admits a strictly complementary solution, then the duality gap converges Q-superlinearly to zero and the iteration sequence converges Q-superlinearly to a strictly complementary solution. The Q-orders of convergence of both sequences are equal to two. We also present two mth order affine scaling methods having $O((\sqrt{n}L)^{\frac{m+1}{m}}(\phi(\sqrt{n}L))^{\frac{1}{m}})$ iteration complexity for monotone LCPs. By taking $m = \Omega(\log(\sqrt{n}L))$, we obtain $O(\sqrt{n}L)$ iteration complexity, the best iteration complexity known so far for LP, QP, and monotone LCP. We note that both algorithms use one matrix factorization and m backsolves per iteration. This requires $O(n^3 + mn^2)$ arithmetic operations per iteration. Therefore if $\log(\sqrt{n}L) \ll n$, then the cost of an iteration is dominated by the cost of the matrix factorization.

If the LCP admits a strictly complementary solution, then the iterates produced by the first algorithm converge Q-superlinearly to a strictly complementary solution, and the corresponding duality gaps converge Q-superlinearly to zero. The Q-orders of convergence of both sequences are equal to $m+1$. If $m \geq 2$, then the second algorithm produces a sequence of iterates that converges Q-superlinearly to a maximal complementarity solution, even for degenerate LCPs, with Q-order $(m+1)/2$. The sequence of duality gaps converges Q-superlinearly to zero with the same Q-order. To our knowledge this is the first affine scaling method acting in the $\mathcal{N}_\infty^-(\alpha)$ neighborhood of the central path that has $O(\sqrt{n}L)$ iteration complexity and Q-superlinear convergence in the absence of strict complementarity.

Conventions. We denote by $\mathbb{N}$ the set of all nonnegative integers. $\mathbb{R}$, $\mathbb{R}_+$, $\mathbb{R}_{++}$ denote the sets of real, nonnegative real, and positive real numbers, respectively. For any real number κ, $\lceil\ \kappa\ \rceil$ denotes the smallest integer greater than or equal to κ. Given a vector x, the corresponding upper case symbol denotes, as usual, the diagonal matrix X defined by the vector. The symbol e represents the vector of all ones, with dimension given by the context.

We denote componentwise operations on vectors by the usual notation for real numbers. Thus, given two vectors u, v of the same dimension, uv, u/v, etc. will denote the vectors with components u_iv_i, u_i/v_i, etc. This notation is consistent as long as componentwise operations always have precedence in relation to matrix operations. Note that $uv \equiv Uv$, and if A is a matrix, then $Auv \equiv AUv$, but in general $A(uv) \neq (Au)v$. Also if f is a scalar function and v is a vector, then $f(v)$ denotes the vector with components $f(v_i)$. For example, if $v \in \mathbb{R}^n_+$ and $\lambda \in \mathbb{R}$, then $\sqrt{v}$ denotes the vector with components $\sqrt{v_i}$, and $\lambda - v$ denotes the vector with components $\lambda - v_i$. Traditionally the vector $\lambda - v$ is written as $\lambda e - v$, where e is the vector of all ones. Inequalities are to be understood in a similar fashion. For example, if $v \in \mathbb{R}^n$, then $v \geq 3$ means that $v_i \geq 3$, $i = 1, \ldots, n$. Traditionally this is written as $v \geq 3\,e$. If $\|.\|$ is a vector norm on $\mathbb{R}^n$ and A is a matrix, then the operator norm induced by $\|.\|$ is defined by $\|A\| = \max\{\|Ax\|\ ;\ \|x\| = 1\}$. As a particular case we note that if U is the

diagonal matrix defined by the vector u, then $\| U \|_2 = \| u \|_\infty$.

We frequently use the $O(\cdot)$ and $\Omega(\cdot)$ notation to express asymptotic relationships between functions. The most common usage will be associated with a sequence $\{x^k\}$ of vectors and a sequence $\{\tau_k\}$ of positive real numbers. In this case $x^k = O(\tau_k)$ means that there is a constant K (dependent on problem data) such that for every $k \in \mathbb{N}$, $\left\|x^k\right\| \leq K\tau_k$. Similarly, if $x^k > 0$, $x^k = \Omega(\tau_k)$ means that $(x^k)^{-1} = O(1/\tau_k)$. If we have both $x^k = O(\tau_k)$ and $x^k = \Omega(\tau_k)$, we write $x^k = \Theta(\tau_k)$.

If $x, s \in \mathbb{R}^n$, then the vector $z \in \mathbb{R}^{2n}$ obtained by concatenating x and s is denoted by $z = \lceil\, x,\, s \,\rfloor = \left[\, x^T,\, s^T \,\right]^T$, and the mean value of xs is denoted by $\mu(z) = \frac{x^T s}{n}$.

2. The linear complementarity problem and its central path.

2.1. The horizontal linear complementarity problem. Given two matrices $Q, R \in \mathbb{R}^{n \times n}$ and a vector $b \in \mathbb{R}^n$, the horizontal linear complementarity problem (HLCP) consists of finding a pair of vectors $z = \lceil\, x,\, s \,\rfloor$ such that

$$\begin{aligned} xs &= 0, \\ Qx + Rs &= b, \\ x, s &\geq 0. \end{aligned} \tag{2.1}$$

The standard (monotone) LCP is obtained by taking $R = -I$ and Q positive semidefinite. There are other formulations of the LCP as well but, as shown in [1], all popular formulations are equivalent, and the behavior of a large class of interior point methods is identical on those formulations, so that it is sufficient to prove results for only one of the formulations. We have chosen HLCP because of its symmetry. The LP problem, and the QP problem, can be formulated as HLCPs. Therefore, the HLCP provides a convenient general framework for studying interior point methods.

Throughout this paper we assume that the HLCP is monotone, in the sense that

$$Qu + Rv = 0 \quad \text{implies } u^T v \geq 0 \text{ for any } u, v \in \mathbb{R}^n\,.$$

This condition is satisfied if the HLCP is a reformulation of a QP [4]. If the HLCP is a reformulation of an LP, then the following stronger condition holds,

$$Qu + Rv = 0 \quad \text{implies } u^T v = 0 \text{ for any } u, v \in \mathbb{R}^n\,,$$

and HLCP is called skew-symmetric in this case. The following proposition contains two simple properties of a monotone HLCP.

PROPOSITION 2.1. *If HLCP is monotone, then the $n \times 2n$-matrix (Q, R) has full rank, and*

$$(u^T \overline{v} + \overline{u}^T v)^2 \leq 4 \left(u^T v\right) \left(\overline{u}^T \overline{v}\right) \qquad \forall \lceil\, u,\, v \,\rfloor, \lceil\, \overline{u}\,,\, \overline{v} \,\rfloor \in \operatorname{Ker}(Q, R)\,.$$

Proof. The full rank property follows from [3]. If $Qu + Rv = Q\overline{u} + R\overline{v} = 0$, then $Q(u + \lambda\overline{u}) + R(v + \lambda\overline{v}) = 0$ for any $\lambda \in \mathbb{R}$, so that

$$0 \leq (u + \lambda\overline{u})^T (v + \lambda\overline{v}) = u^T v + \lambda(u^T \overline{v} + \overline{u}^T v) + \lambda^2 \overline{u}^T \overline{v} \qquad \forall \lambda \in \mathbb{R}\,.$$

Since the right-hand side above is a nonnegative quadratic function in λ, its discriminant must be nonpositive, which proves the second part of the proposition. ∎

In the skew-symmetric case we can often obtain sharper estimates, due to the following consequence of Proposition 2.1.

COROLLARY 2.2. *If HLCP is skew-symmetric, then* $u^T\overline{v}+\overline{u}^Tv=0$ *for all* $u,v,\overline{u},\overline{v}$ *satisfying* $Qu+Rv=Q\overline{u}+R\overline{v}=0$.

Let us denote the set of all feasible points of HLCP by

$$\mathcal{F}=\{z=\lceil\, x,\, s\,\rfloor\in\mathbb{R}^{2n}_{+}\ :\ Qx+Rs=b\},$$

and the solution set (or the optimal face) of HLCP by

$$\mathcal{F}^*=\{z^*=\lceil\, x^*,\, s^*\,\rfloor\in\mathcal{F}\ :\ x^*s^*=0\}.$$

The structure of $\mathcal{F}^*$ is very important in the analysis of interior point methods. Let us define three subsets $\mathcal{B}$, $\mathcal{N}$, and $\mathcal{J}$ of the index set $\{1,\dots,n\}$ by

$$\mathcal{B}=\{i=1,\dots,n\,|\,x_i^*>0 \text{ for at least one } \lceil\, x^*,\, s^*\,\rfloor\in\mathcal{F}^*\},$$

$$\mathcal{N}=\{i=1,\dots,n\,|\,s_i^*>0 \text{ for at least one } \lceil\, x^*,\, s^*\,\rfloor\in\mathcal{F}^*\},$$

$$\mathcal{J}=\{i=1,\dots,n\,|\,x_i^*=s_i^*=0\ \forall\lceil\, x^*,\, s^*\,\rfloor\in\mathcal{F}^*\}.$$

One can prove that $\mathcal{B}$, $\mathcal{N}$, and $\mathcal{J}$ form a partition of $\{1,\dots,n\}$ and that there exists a solution $\lceil\, x^*,\, s^*\,\rfloor\in\mathcal{F}^*$ such that $x^*_{\mathcal{B}}>0$ and $s^*_{\mathcal{N}}>0$. Such a solution is called a *maximal complementarity solution*, since one can prove that, for any $\lceil\, x^*,\, s^*\,\rfloor\in\mathcal{F}^*$, $x_i^*>0\Rightarrow i\in\mathcal{B}$ and $s_j^*>0\Rightarrow j\in\mathcal{N}$. If the solution of the HLCP is unique, then it is a maximal complementarity solution. Otherwise it can be shown that the relative interior of $\mathcal{F}^*$ is composed of maximal complementarity solutions.

If the set $\mathcal{J}$ is empty, then a maximal complementarity solution is called a strictly complementary solution. Let us denote by $\mathcal{F}^c$ the set of all such solutions, i.e.,

$$\mathcal{F}^c=\{z^*=\lceil\, x^*,\, s^*\,\rfloor\in\mathcal{F}^*:x^*+s^*>0\}\,.$$

We say that the HLCP is nondegenerate if it has a strictly complementary solution. If the set J is nonempty, then we say that the HLCP is degenerate.

The set

$$\mathcal{F}^0=\mathcal{F}\bigcap\mathbb{R}^{2n}_{++}$$

is called the set of strictly feasible points, or the set of interior points.

2.2. The central path and its analyticity. It is known (see [13]) that if $\mathcal{F}^0$ is nonempty, then for any vector $p\in\mathbb{R}^n_{++}$ and any parameter $\tau>0$ the nonlinear system

$$\begin{aligned} xs&=\tau p,\\ Qx+Rs&=b \end{aligned}\tag{2.2}$$

has a unique positive solution $z(\tau,p)=\lceil\, x(\tau,p),\, s(\tau,p)\,\rfloor$. For a fixed p, the curve

$$\mathcal{C}(p)=\{z(\tau,p)\ :\ \tau>0\}$$

is called the weighted central path of the HLCP with weight vector p. It turns out that if the HLCP is nondegenerate, then $z(\tau,p)$ is an analytic function in $\tau>0$ and $p>0$, which has an analytic continuation at $\tau=0$. In case the HLCP is degenerate, then $z(\tau,p)$ is analytic in $\rho=\sqrt{\tau}$, which also has an analytic continuation at $\rho=0$.

Moreover, all the derivatives of z are bounded on any compact set contained in the domain of analyticity of z. More precisely, the following results follow from the more general results of [39, 40].

THEOREM 2.3. *If HLCP is monotone and $\mathcal{F}^0$ is nonempty, then (2.2) has a unique positive solution $z(\tau,p) = \lceil\, x(\tau,p),\, s(\tau,p)\,\rfloor$ for any $(\tau,p) \in \mathbb{R}^{n+1}_{++}$, and the following properties hold:*

A. *If the HLCP is nondegenerate, then $z(\tau,p)$ can be analytically extended to an open neighborhood of $\mathbb{R}_+ \times \mathbb{R}^n_{++}$, and for any compact set $\mathcal{K} \subset \mathbb{R}_+ \times \mathbb{R}^n_{++}$ and any integer $i \in \mathbb{N}$ there are constants $c(\mathcal{K},i)$ such that*

$$\left\| \frac{\partial^i z(\tau,p)}{\partial \tau^i} \right\|_2 \le c(\mathcal{K},i) \qquad \forall (\tau,p) \in \mathcal{K},\ i = 0,1,2,\ldots.$$

B. *If the HLCP is degenerate, then $\overline{z}(\rho,p) := z(\rho^2,p)$ can be analytically extended to an open neighborhood of $\mathbb{R}_+ \times \mathbb{R}^n_{++}$, and for any compact set $\mathcal{K} \subset \mathbb{R}_+ \times \mathbb{R}^n_{++}$ and any integer $i \in \mathbb{N}$ there are constants $\overline{c}(\mathcal{K},i)$ such that*

$$\left\| \frac{\partial^i \overline{z}(\rho,p)}{\partial \rho^i} \right\|_2 \le \overline{c}(\mathcal{K},i) \qquad \forall (\tau,p) \in \mathcal{K},\ i = 0,1,2,\ldots.$$

If $p = e$, the vector of all ones, then $\mathcal{C} = \mathcal{C}(e)$ is simply called the central path of the HLCP. The distance of a point $z \in \mathcal{F}$ to the central path $\mathcal{C}$ can be quantified by different proximity measures. The following proximity measures have been extensively used in the interior point literature:

$$\delta_2(z) := \left\| \frac{xs}{\mu(z)} - e \right\|_2, \quad \delta_\infty(z) := \left\| \frac{xs}{\mu(z)} - e \right\|_\infty, \quad \delta^-_\infty(z) := \left\| \left[\frac{xs}{\mu(z)} - e \right]^- \right\|_\infty,$$

where $[v]^-$ denotes the negative part of the vector v, i.e., $[v]^- = -\max\{-v, 0\}$.

By using the above proximity measures, we can define the following neighborhoods of the central path:

$$\mathcal{N}_2(\eta) = \{ z \in \mathcal{F}^0 \ :\ \delta_2(z) \le \eta \},$$
$$\mathcal{N}_\infty(\eta) = \{ z \in \mathcal{F}^0 \ :\ \delta_\infty(z) \le \eta \},$$
$$\mathcal{N}^-_\infty(\eta) = \{ z \in \mathcal{F}^0 \ :\ \delta^-_\infty(z) \le \eta \},$$

where $0 < \eta < 1$ is a given parameter. We have

$$\mathcal{C} \subset \mathcal{N}_2(\eta) \subset \mathcal{N}_\infty(\eta) \subset \mathcal{N}^-_\infty(\eta), \quad \lim_{\eta \downarrow 0} \mathcal{N}^-_\infty(\eta) = \mathcal{C}, \quad \lim_{\eta \uparrow 1} \mathcal{N}^-_\infty(\eta) = \mathcal{F}. \tag{2.3}$$

3. The first order affine scaling method.

3.1. The algorithm. In the remainder of this paper we will work with $\mathcal{N}^-_\infty(\eta)$. We note that this neighborhood can be written under the form

$$\mathcal{N}^-_\infty(\eta) = \mathcal{D}(1-\eta), \text{ where } \mathcal{D}(\beta) = \{ z \in \mathcal{F}^0 \ :\ xs \ge \beta \mu(z) \}.$$

At each step of our algorithm we are given a point $z = \lceil\, x,\, s\,\rfloor \in \mathcal{D}(\beta)$, and we compute the affine scaling direction $w = \lceil\, u,\, v\,\rfloor$ by solving the following linear system:

$$\begin{cases} su + xv = -xs, \\ Qu + Rv = 0. \end{cases} \tag{3.1}$$

This search direction coincides, up to a scalar factor, with the first derivative to central path $z(\tau,p)$ that passes through $p=xs/\mu$ at $\tau=\mu$. Indeed, by differentiating (2.2), we obtain

$$\left\{\begin{array}{l} s(\tau,p)\frac{\partial}{\partial\tau}x(\tau,p)+x(\tau,p)\frac{\partial}{\partial\tau}s(\tau,p)=\frac{xs}{\mu}, \\ \qquad\qquad Q\frac{\partial}{\partial\tau}x(\tau,p)+R\frac{\partial}{\partial\tau}s(\tau,p)=0. \end{array}\right. \tag{3.2}$$

Given that $x(\mu,p)=x$ and $s(\mu,p)=s$, we obtain

$$u=-\mu\frac{\partial}{\partial\tau}x(\tau,p)\,,\qquad v=-\mu\frac{\partial}{\partial\tau}s(\tau,p)\,. \tag{3.3}$$

It turns out that the duality gap has the fastest local decrease along this direction. Indeed if we denote by

$$z(\theta)=z+\theta w \tag{3.4}$$

a point along this direction, we have

$$\begin{aligned} x(\theta)s(\theta)&=(x+\theta u)(s+\theta v)=(1-\theta)xs+\theta^2uv, \\ \mu(\theta)&=\frac{(x+\theta u)^T(s+\theta v)}{n}=(1-\theta)\mu+\theta^2\frac{u^Tv}{n}\,. \end{aligned} \tag{3.5}$$

In the skew-symmetric case we have $u^Tv=0$, while in the monotone case, according to Lemma 3.1, we have $0\le u^Tv\le .25n\mu$, so that we can write

$$\mu(\theta)\le(1-\theta+.25\theta^2)\mu\le(1-.5\theta)^2\mu\,,\qquad \mu'(\theta)\le(.5\theta-1)\mu\,. \tag{3.6}$$

Given that $\mu(\theta)$ is decreasing on the interval $[0,2]$, we compute the steplength $\overline{\theta}$ as

$$\overline{\theta}=\max\left\{\hat{\theta}\in[0,1]\ :\ z(\theta)\in\mathcal{D}(\beta_+)\ \forall\theta\in[0,\hat{\theta}\,]\right\}\,, \tag{3.7}$$

where

$$\beta_+=\beta-\alpha. \tag{3.8}$$

We obtain thus a new point satisfying

$$z^+=z(\overline{\theta}\,)\in\mathcal{D}(\beta_+), \tag{3.9}$$

and the process can be repeated. The initial value for β is $\overline{\beta}$, and the α's are chosen as the terms of a monotone sequence of nonnegative numbers satisfying the following two conditions:

$$\sum_{k=0}^{\infty}\alpha_k\le\overline{\beta}\,-\underline{\beta}\,,\qquad \sum_{k=0}^{\infty}\sqrt{\alpha_k}=\infty.$$

This ensures the global convergence of the algorithm. In order to obtain good polynomial complexity we have chosen

$$\alpha_k=\frac{\nu(\overline{\beta}\,-\underline{\beta})}{(\exp(1)+k+1)\log^{1+\nu}(\exp(1)+k+1)}\,,\quad k=0,1,\ldots, \tag{3.10}$$

where $0 < \nu \leq 1$ is a given parameter. For this choice we have

$$\sum_{k=0}^{\infty} \alpha_k < \nu(\overline{\beta} - \underline{\beta}) \int_e^{\infty} \frac{dt}{t \log^{1+\nu} t} = \overline{\beta} - \underline{\beta}\,,$$

and, as we will see in the next section,

$$\sum_{k=0}^{K} \sqrt{\alpha_k} = \Omega\left(\frac{\sqrt{K}}{\log^{\frac{1+\nu}{2}} K}\right).$$

We end this section by formally defining our iterative procedure, as follows.

ALGORITHM 1. Given real parameters $0 < \underline{\beta} < \overline{\beta} < 1$, $0 < \nu \leq 1$, and a vector $z^0 \in \mathcal{D}(\overline{\beta}\,)$:

Set $k \leftarrow 0\,,\ \beta_0 \leftarrow \overline{\beta}$ and consider the sequence (3.10);
repeat
Set $z \leftarrow z^k\,,\ \alpha \leftarrow \alpha_k\,,\ \beta \leftarrow \beta_k\,,\ \beta_+ \leftarrow \beta - \alpha\,;$
Compute direction $w = \lceil\, u,\, v \,\rfloor$ by solving (3.1);
Compute steplength $\overline{\theta}$ from (3.7);
Compute z^+ from (3.9);
Set $\theta_k \leftarrow \overline{\theta}\,,\ z^{k+1} \leftarrow z^+\,,\ \mu_{k+1} \leftarrow \mu(z^+)\,,\ \beta_{k+1} \leftarrow \beta_+;$
Set $k \leftarrow k + 1$.
continue

3.2. Technical results. In order to analyze Algorithm 1 we need to establish some properties of the solution of a linear system of the form

$$\left\{ \begin{array}{r} su + xv = a, \\ Qu + Rv = 0. \end{array} \right. \tag{3.11}$$

By using the notation

$$D = X^{-1/2} S^{1/2}\,, \tag{3.12}$$

this system can be written as

$$\left\{ \begin{array}{r} Du + D^{-1}v = (xs)^{-1/2} a, \\ QD^{-1}(Du) + RD(D^{-1}v) = 0. \end{array} \right. \tag{3.13}$$

Since the pair (QD^{-1}, RD) is monotone, one can easily prove the following results (see, for example, [29, Lemma 3.1]).

LEMMA 3.1. *If HLCP is monotone, then for any $z = \lceil\, x,\, s \,\rfloor \in \mathbb{R}^{2n}_{++}$ and any $a \in \mathbb{R}^n$ the linear system* (3.11) *has a unique solution $w = \lceil\, u,\, v \,\rfloor$, and the following properties are satisfied:*

(i) $\| Du \|_2^2 + \left\| D^{-1}v \right\|_2^2 \leq \left\| (xs)^{-1/2} a \right\|_2^2\,;$
(ii) $\|uv - \frac{u^T v}{n}\, e\|_2 \leq \| uv \|_2 \leq \frac{1}{2\sqrt{2}} \left\| (xs)^{-1/2} a \right\|_2^2\,;$
(iii) $\| uv \|_\infty \leq \frac{1}{4} \left\| (xs)^{-1/2} a \right\|_2^2\,;$
(iv) $u^T v \leq \frac{1}{4} \left\| (xs)^{-1/2} a \right\|_2^2\,.$

The proof of global convergence of Algorithm 1 requires explicit lower bounds for the series $\sum_{k=0}^{K-1} \alpha_k^{1/(m+1)}$ for K sufficiently large. Given that

$$\sum_{k=0}^{K-1} \alpha_k^{\frac{1}{m+1}} > \left(\nu(\overline{\beta} - \underline{\beta})\right)^{\frac{1}{m+1}} \int_{\exp(1)+1}^{K+\exp(1)+1} \frac{dt}{t^{\frac{1}{m+1}} \log^{\frac{1+\nu}{m+1}} t}\,, \tag{3.14}$$

it remains to find an explicit lower bound depending on K of the integral appearing on the right-hand-side of the above inequality. To this effect we will use the following results.

LEMMA 3.2. *Let f and g be two continuous, nonnegative, and nonincreasing real functions defined on the interval $[\gamma, \delta]$. If f is convex, then*

$$\int_\gamma^\delta f(t)g(t)dt \geq f\left(\frac{\gamma+\delta}{2}\right) \int_\gamma^\delta g(t)dt\,.$$

Proof. Let us define $\zeta = (\gamma+\delta)/2$. From the convexity of f it follows that f has lateral derivatives $f'(\zeta-0) \leq f'(\zeta+0) \leq 0$ at ζ and that

$$f(t) \geq f(\zeta) + m(t-\zeta) \qquad \forall m \in [f'(\zeta-0), f'(\zeta+0)]\,,\ \forall t \in [\gamma,\delta]\,.$$

Since g is nonnegative and nonincreasing we deduce that

$$f(t)g(t) \geq f(\zeta)g(t) + mg(\zeta)(t-\zeta) \qquad \forall m \in [f'(\zeta-0), f'(\zeta+0)]\,,\ \forall t \in [\gamma,\delta]\,,$$

and the statement of our lemma follows by fixing m and taking the integral with respect to t of both sides of the above inequality. ∎

COROLLARY 3.3. *If $0 < \omega < 1$, $\eta > 0$, and $\varsigma \geq 2\gamma > 2$, then*

$$\int_\gamma^{\gamma+\varsigma} \frac{dt}{t^\omega \log^\eta t} \geq \frac{\varsigma^{1-\omega}}{\log^\eta \varsigma}\,.$$

Proof. Given that the function $t \to \log^{-\eta}(t)$ is convex for $t > 1$, we can apply Lemma 3.2 to obtain

$$\int_\gamma^{\gamma+\varsigma} \frac{dt}{t^\omega \log^\eta t} \geq \frac{(\gamma+\varsigma)^{1-\omega} - \gamma^{1-\omega}}{(1-\omega)\log^\eta(\gamma+.5\varsigma)} \geq \frac{(\gamma+\varsigma)^{1-\omega} - \gamma^{1-\omega}}{(1-\omega)\log^\eta \varsigma}\,.$$

Therefore it remains to prove that

$$(\gamma+\varsigma)^{1-\omega} - \gamma^{1-\omega} \geq (1-\omega)\varsigma^{1-\omega}\,.$$

Dividing both sides of the above inequality by $\varsigma^{1-\omega}$ and denoting $\sigma = \gamma/\varsigma$, $\tau = 1-\omega$, this reduces to proving that

$$\phi(\tau,\sigma) := (1+\sigma)^\tau - \sigma^\tau - \tau \geq 0\,,\ \forall \tau \in [0,1] \qquad \forall \sigma \in [0,.5]\,.$$

Since

$$\frac{\partial \phi}{\partial \sigma} = \tau\left((1+\sigma)^{\tau-1} - \sigma^{\tau-1}\right) \leq 0,$$

we deduce that

$$\phi(\tau,\sigma) \geq \psi(\tau) := \phi(\tau,.5) \qquad \forall \tau \in [0,1]\,,\ \forall \sigma \in [0,.5]\,.$$

We have

$$\psi(\tau) = 1.5^\tau - .5^\tau - \tau\,, \quad \psi(0) = 0\,,\ \psi(1) = 0,$$
$$\psi'(\tau) = 1.5^\tau \log 1.5 - .5^\tau \log .5 - 1\,, \quad \psi'(0) > 0\,,\ \psi'(1) < 0,$$
$$\psi''(\tau) = 1.5^\tau \log^2 1.5 - .5^\tau \log^2 .5\,, \quad \psi''(0) < 0\,,\ \psi''(1) > 0,$$
$$\psi'''(\tau) = 1.5^\tau \log^3 1.5 - .5^\tau \log^3 .5 > 0\,, \quad \forall \tau \in [0,1]\,.$$

The positivity of ψ''' implies that ψ'' has a unique zero $\tau_2 \in [0,1]$, which in turn shows that ψ' is decreasing on $[0,\tau_2]$ and increasing on $[\tau_2,1]$. Hence ψ' has a unique zero $\tau_1 \in [0,\tau_2]$. This shows that ψ increases on $[0,\tau_1]$ and decreases on $[\tau_1,1]$. Therefore ψ is positive on $(0,1)$. □

Using this corollary together with the following lemma, we will be able to obtain explicit upper bounds on the number of steps required by Algorithm 1 to reduce the duality gap below any desired tolerance.

LEMMA 3.4. *The function* $h : [\exp(1),\infty) \to [\exp(1),\infty)$ *given by* $h(t) = t/\log t$ *is increasing and bijective, and its inverse* $h^{-1} : [\exp(1),\infty) \to [\exp(1),\infty)$ *satisfies the inequalities*

$$t\log t < h^{-1}(t) < t\left(1 + \frac{\log\log t}{\log t - 1}\right)\log t < 2t\log t \qquad \forall t \in (e,\infty)\,.$$

Proof. The first part of the lemma follows immediately from the fact that

$$h'(t) = \frac{\log t - 1}{\log^2 t} > 0 \qquad \forall t \in (e,\infty)\,.$$

In order to prove the second part we denote by l and u the lower and upper bounds above and show that $h(l) < t < h(u)$. We have obviously

$$h(l) = \frac{t\log t}{\log(t\log t)} = \frac{t\log t}{\log t + \log\log t} < t\,,$$

which proves the first inequality. Finally, by using the inequality $\log(1+\alpha) < \alpha$, we obtain

$$h(u) = \frac{t\left(1 + \frac{\log\log t}{\log t - 1}\right)\log t}{\log(t\log t) + \log\left(1 + \frac{\log\log t}{\log t - 1}\right)} > \frac{t\left(1 + \frac{\log\log t}{\log t - 1}\right)\log t}{\log t + \log\log t + \frac{\log\log t}{\log t - 1}} = t\,. \qquad \square$$

COROLLARY 3.5. *If* $\sigma > \exp(1)\,,\ \nu > 0\,,\ m > 0\,,$ *then*

$$\tau \geq \left(2\left(1 + \frac{1+\nu}{m+1}\log\frac{1+\nu}{m}\right)\frac{m+1}{m}\log\sigma\right)^{\frac{1+\nu}{m}} \sigma^{\frac{m+1}{m}} \quad \textit{implies} \quad \frac{\tau^{\frac{m}{m+1}}}{\log^{\frac{1+\nu}{m+1}}\tau} \geq \sigma\,.$$

Proof. By noticing that

$$\frac{\tau^{\frac{m}{m+1}}}{\log^{\frac{1+\nu}{m+1}}\tau} = \left(\frac{\tau^{\frac{m}{1+\nu}}}{\frac{1+\nu}{m}\log\tau^{\frac{m}{1+\nu}}}\right)^{\frac{1+\nu}{m+1}}\,,$$

we can write the last inequality in the statement of Corollary 3.5 under the equivalent form

$$\frac{\tau^{\frac{m}{1+\nu}}}{\log \tau^{\frac{m}{1+\nu}}} \geq \frac{(1+\nu)\,\sigma^{\frac{m+1}{1+\nu}}}{m}\,.$$

According to Lemma 3.4 this inequality is satisfied if

$$\tau \geq \left(\frac{2(1+\nu)\,\sigma^{\frac{m+1}{1+\nu}}}{m}\,\log\frac{(1+\nu)\,\sigma^{\frac{m+1}{1+\nu}}}{m}\right)^{\frac{1+\nu}{m}} =: rhs\,.$$

The right-hand side of the above inequality can be majorized as follows:

$$\begin{aligned}
rhs &= \left(\frac{2(1+\nu)}{m}\right)^{\frac{1+\nu}{m}} \sigma^{\frac{m+1}{m}} \left(\log\frac{(1+\nu)\,\sigma^{\frac{m+1}{1+\nu}}}{m}\right)^{\frac{1+\nu}{m}} \\
&= \left(\frac{2(1+\nu)}{m}\right)^{\frac{1+\nu}{m}} \sigma^{\frac{m+1}{m}} \left(\left(1+\frac{\log\frac{1+\nu}{m}}{\log\sigma^{\frac{m+1}{1+\nu}}}\right)\log\sigma^{\frac{m+1}{1+\nu}}\right)^{\frac{1+\nu}{m}} \\
&\leq \left(\frac{2(1+\nu)}{m}\right)^{\frac{1+\nu}{m}} \sigma^{\frac{m+1}{m}} \left(\left(1+\frac{1+\nu}{m+1}\log\frac{1+\nu}{m}\right)\log\sigma^{\frac{m+1}{1+\nu}}\right)^{\frac{1+\nu}{m}} \\
&= \left(\frac{2(1+\nu)}{m}\right)^{\frac{1+\nu}{m}} \sigma^{\frac{m+1}{m}} \left(\left(1+\frac{1+\nu}{m+1}\log\frac{1+\nu}{m}\right)\frac{m+1}{1+\nu}\log\sigma\right)^{\frac{1+\nu}{m}} \\
&= \left(2\left(1+\frac{1+\nu}{m+1}\log\frac{1+\nu}{m}\right)\frac{m+1}{m}\log\sigma\right)^{\frac{1+\nu}{m}} \sigma^{\frac{m+1}{m}},
\end{aligned}$$

which proves the corollary. □

LEMMA 3.6. *The function* $g : [\exp(\exp(1)),\infty) \to [\exp(\exp(1)),\infty)$ *given by* $g(t) = t/\log\log t$ *is increasing and bijective, and its inverse satisfies*

$$t\log\log t < g^{-1}(t) < t\left(1+\frac{\log\log\log t}{(\log\log t)\log t - 1}\right)\log\log t < 2.072\,t\log\log t\,.$$

Proof. The first part of the lemma follows immediately from the fact that

$$g'(t) = \frac{-1+(\log\log t)\log t}{(\log\log t)^2\log t} > 0 \qquad \forall t \in (5.8313,\infty)\,.$$

In order to prove the second part we denote by l and u the lower and upper bounds above and show that $g(l) < t < g(u)$. We have obviously

$$g(l) = \frac{t\log\log t}{\log\log(t\log\log t)} = \frac{t\log\log t}{\log\log t + \log\log\log t} < t\,,$$

which proves the first inequality. By using the inequality $\log(1+\alpha) < \alpha$, we obtain

$$g(u) = \frac{t\left(1+\frac{\log\log\log t}{(\log\log t)\log t-1}\right)\log\log t}{\log\left(\log t+\log\log\log t+\log\left(1+\frac{\log\log\log t}{(\log\log t)\log t-1}\right)\right)}$$

$$> \frac{t\left(1+\frac{\log\log\log t}{(\log\log t)\log t-1}\right)\log\log t}{\log\left(\log t+(\log\log\log t)\left(1+\frac{1}{(\log\log t)\log t-1}\right)\right)}$$

$$> \frac{t\left(1+\frac{\log\log\log t}{(\log\log t)\log t-1}\right)\log\log t}{\log\log t+\frac{\log\log\log t}{\log t}\frac{(\log\log t)\log t}{(\log\log t)\log t-1}} = t\,.$$

Finally, it is easily checked that

$$\max_{t\geq\exp(\exp(1))}\frac{\log\log\log t}{(\log\log t)\log t-1} < 0.072\,. \qquad \square$$

3.3. Polynomial complexity. In the next theorem we show that Algorithm 1 is well defined, and we give bounds on the decrease of the duality gap at each iteration.

THEOREM 3.7. *If HLCP is monotone, then Algorithm* 1 *is well defined and produces a sequence of points* $\left(z^k\right)_{k=0}^{\infty}$, *with* $z^k \in \mathcal{D}(\beta_k) \subset \mathcal{D}(\underline{\beta})$. *If* $n \geq 5$, *then the following relations hold:*

$$\theta_k \geq \sqrt{2\alpha_k/n}\,, \quad \mu_{k+1} \leq \left(1-\sqrt{\alpha_k/n}\right)\mu_k\,, \qquad k=0,1,\ldots,$$

$$\mu_k \leq \mu_0 \exp\left(-\sqrt{\frac{\nu\left(\overline{\beta}-\underline{\beta}\right)k}{n\log^{1+\nu}k}}\right)\,, \qquad k=8,9,\ldots.$$

Proof. The relation $z^k \in \mathcal{D}(\beta_k)$ is ensured by the line search (3.7). Now let $z \in \mathcal{D}(\beta)$ be given, and define

$$p = \frac{xs}{\mu}\,, \qquad q = \frac{uv}{\mu}\,.$$

From our hypothesis and from Lemma 3.1 it follows that

$$p \geq \beta e\,, \quad \|\,q\,\|_\infty \leq \frac{n}{4}\,, \quad \left\|\,q-(e^Tq/n)e\,\right\|_2 \leq \|\,q\,\|_2 \leq \frac{n}{2\sqrt{2}}\,, \quad \frac{u^Tv}{\mu} = e^Tq \leq \frac{n}{4}\,.$$

Using (3.5) and (3.6), we deduce that

$$\frac{x(\theta)s(\theta)}{\mu(\theta)} = \frac{(1-\theta)p+\theta^2q}{1-\theta+(e^Tq/n)\theta^2}$$

$$= \frac{(1-\theta)p+\beta(e^Tq/n)\theta^2e+\beta\theta^2(q-(e^Tq/n)e)+(1-\beta)\theta^2q}{1-\theta+(e^Tq/n)\theta^2}$$

$$\geq \beta e - \frac{\left\|\,\beta(q-(e^Tq/n)e)+(1-\beta)q\,\right\|_2\theta^2}{1-\theta}\,e$$

$$\geq \beta e - \frac{\|\,q\,\|_2\,\theta^2}{1-\theta}\,e\,. \tag{3.15}$$

If $n \geq 5$, then $\sqrt{2\alpha/n} \leq \sqrt{2\alpha_0/5} < 1-1/\sqrt{2}$ for any values of the parameters defining our algorithm, and in this case we can easily verify that

$$\frac{\| q \|_2 \, \theta^2}{1-\theta} \leq \frac{n\theta^2}{(1-\theta)2\sqrt{2}} \leq \alpha \qquad \forall \theta \in [0, \sqrt{2\alpha/n}\,] \, .$$

It follows that $\overline{\theta} \geq \sqrt{2\alpha/n}$, and therefore

$$\mu_+ = \mu\left(\overline{\theta}\right) \leq \mu\left(\sqrt{2\alpha/n}\right) \leq \left(1 - \sqrt{2\alpha/n} + .5\alpha/n\right)\mu$$

$$\leq \left(1 - \left(1 - .25\sqrt{2\alpha_0/5}\right)\sqrt{2\alpha/n}\right)\mu < \left(1 - \sqrt{\alpha/n}\right)\mu \, .$$

Hence we have proved that

$$\mu_{j+1} \leq \left(1 - \sqrt{\alpha_j/n}\right)\mu_j \, , \qquad j = 0, 1, \ldots .$$

By repeatedly applying the above inequality, we deduce that

$$\mu_k \leq \mu_0 \prod_{j=0}^{k-1} \left(1 - \sqrt{\alpha_j/n}\right) .$$

Using the inequality $\log(1-t) < -t$, (3.14), $k \geq 8 > 2(\exp(1)+1)$, and Corollary 3.3, we obtain

$$\log\left(\frac{\mu_k}{\mu_0}\right) \leq \sum_{j=0}^{k-1} \log\left(1 - \sqrt{\alpha_j/n}\right) \leq -\sum_{j=0}^{k-1} \sqrt{\alpha_j/n} \leq -\sqrt{\frac{\nu\left(\overline{\beta} - \underline{\beta}\right)k}{n \log^{1+\nu} k}} \, ,$$

which completes the proof of our theorem. □

In the following corollary we give an explicit upper bound for the number of iterations required by Algorithm 1 to obtain a solution of the HLCP with prescribed accuracy. More precisely, given any $\epsilon > 0$, we have to find an upper bound for the number

$$K_\epsilon := \min\left\{K \ : \ x^{k\,T} s^k \leq \epsilon \ \forall k \geq K\right\} . \tag{3.16}$$

The upper bound will depend on n and

$$L_\epsilon := \log\left(\frac{x^{0\,T} s^0}{\epsilon}\right) . \tag{3.17}$$

COROLLARY 3.8. *If* $nL_\epsilon^2 > \exp(2)/\left(\nu(\overline{\beta} - \underline{\beta})\right)$, *then*

$$K_\epsilon \leq \left\lceil \frac{7^{1+\nu}}{\nu(\overline{\beta} - \underline{\beta})} \, nL_\epsilon^2 \log^{1+\nu}\left(nL_\epsilon^2\right) \right\rceil .$$

Proof. From Theorem 3.7 we deduce that $x^{k\,T} s^k \leq \epsilon$ for any K with the property

$$\frac{K^{1/2}}{\log^{(1+\nu)/2} K} \geq \sigma := \left(\nu(\overline{\beta} - \underline{\beta})\right)^{-1/2} \sqrt{n} L_\epsilon \, .$$

Since $\sigma > \exp(1)$, we can use Corollary 3.5 to show that

$$K_\epsilon \leq \overline{K}_\epsilon := \left(4\,(1+\log 2)\log\sigma\right)^{1+\nu}\sigma^2,$$

and the statement of our corollary follows by noticing that

$$4\,(1+\log 2)\log\sigma = 2\,(1+\log 2)\left(\log\left(nL_\epsilon^2\right) - \log\left(\nu(\overline{\beta}-\underline{\beta})\right)\right)$$

$$\leq 4\,(1+\log 2)\left(\log\left(nL_\epsilon^2\right)-1\right) < 4\,(1+\log 2)\log\left(nL_\epsilon^2\right) < 7\log\left(nL_\epsilon^2\right). \qquad \square$$

If ν is a positive constant independent of n and L_ϵ, then Algorithm 1 is independent of the dimension of the problem and the stopping criterion $x^{k\,T}s^k \leq \epsilon$. However, by letting ν depend on n and L_ϵ, we can slightly improve the computational complexity.

COROLLARY 3.9. *If*

$$nL_\epsilon^2 \geq \max\left\{\exp\left(\exp(1)\right),\ \frac{2.072\exp(2)}{\overline{\beta}-\underline{\beta}}\log\log\left(\frac{\exp(2)}{\overline{\beta}-\underline{\beta}}\right)\right\},$$

and we take $\nu = (\log\log(nL_\epsilon^2))^{-1}$ *in Algorithm* 1, *then*

$$K_\epsilon \leq \left\lceil \frac{134}{\overline{\beta}-\underline{\beta}}\,nL_\epsilon^2\log\left(nL_\epsilon^2\right)\log\log\left(nL_\epsilon^2\right)\right\rceil.$$

Proof. From Lemma 3.6 we deduce that $nL_\epsilon^2 \geq \exp(2)/(\nu(\overline{\beta}-\underline{\beta}))$, and Corollary 3.8 can be applied. The desired result follows immediately by noticing that $\log^\nu(nL_\epsilon^2) = \exp(1)$ and $7^{1+\nu}\exp(1) \leq 49\exp(1) < 134$. $\square$

We note that although for $\nu = (\log\log(nL_\epsilon))^{-1}$ Algorithm 1 depends on ϵ, it will produce an (infinite) sequence $z^k \in \mathcal{D}(\underline{\beta})$ with $\lim_{k\to\infty}\mu_k = 0$. As we mentioned in the introduction, this is not the case with the primal-dual affine-scaling algorithm of Monteiro, Adler, and Resende [23]. Given a starting point $z^0 \in \mathcal{C}$ on the central path and a tolerance $\epsilon > 0$, their algorithm produces a finite sequence of points $\left(z^k\right)$ by taking at each iteration a fixed stepsize $\theta_\epsilon = 1/\left(n\lceil\log L_\epsilon\rceil\right)$, i.e.,

$$z^{k+1} = z^k + \theta_\epsilon w^k, \qquad k = 0,1,\ldots,\overline{K}_\epsilon := \frac{\lceil n\log L_\epsilon\rceil}{\theta_\epsilon}.$$

It is shown that $z^k \in \mathcal{N}_\infty(\overline{\alpha}_k)$, with $\overline{\alpha}_k < n\overline{K}_\epsilon\overline{\theta}_\epsilon^{\,2} = 1\,, k = 0,1,\ldots,\overline{K}_\epsilon$, which implies the feasibility of z^k for $k \leq \overline{K}_\epsilon$. However, the positivity of z^k is no longer guaranteed for $k > \overline{K}_\epsilon$, so that the algorithm is not defined for $k > \overline{K}_\epsilon$. Since the algorithm of [23] uses a fixed (and small) steplength it cannot have superlinear convergence. Moreover since this algorithm produces only a finite sequence, we cannot even talk about its order of convergence. In contrast, in the next subsection we will show that the Q-order of convergence of Algorithm 1 is two.

3.4. Superlinear convergence. The main ingredient in the proof of superlinear convergence is provided by the following lemma, which is easily obtained by applying Theorem 2.3 with

$$\mathcal{K} = \mathcal{K}(\underline{\beta},\overline{\mu}) = \left\{(\tau,p)\ :\ 0 \leq \tau \leq \overline{\mu}\,,\ e^Tp = n\,,\ p \geq \underline{\beta}\,e\right\} \tag{3.18}$$

and (3.3).

LEMMA 3.10. *If HLCP* 2.1 *is monotone and has a strictly complementary solution, then for any* $\beta \in (0,1)$ *and any* $\overline{\mu} > 0$ *there is constant* $c = c(\beta, \overline{\mu})$ *such that the affine-scaling direction given by* (3.1) *satisfies*

$$\| u \|_2 \le c\,\mu\,, \quad \| v \|_2 \le c\,\mu \qquad \forall z \in \left\{ \lceil x,\, s \rfloor \in \mathcal{D}(\beta) \;:\; \mu = \frac{x^T s}{n} \le \overline{\mu} \right\}.$$

In the next theorem we prove that under the strict complementarity assumption the sequence (μ_k) converges to zero with Q-order 2. We recall that the Q-order of a sequence of positive numbers (η_k) that converges to zero is defined as

$$\mathcal{Q}(\eta_k) = \sup\{\omega \in (1,\infty) \;:\; \exists \Gamma \in \mathbb{R}_{++}\,,\, \forall k \in \mathbb{N}\,,\, \eta_{k+1} \le \Gamma \eta_k^{\omega}\}\,.$$

It is known (see [27]) that for $\overline{\omega} > 1$ we have $\overline{\omega} = \mathcal{Q}(\eta_k)$ if and only if

$$\overline{\omega} = \liminf \frac{\log \eta_{k+1}}{\log \eta_k}\,.$$

THEOREM 3.11. *If HLCP is monotone and has a strictly complementary solution, then the sequence* $\left(z^k\right)$ *produced by Algorithm* 1 *converges Q-superlinearly to a strictly complementary solution* $z^* \in \mathcal{F}^c$*, and the sequence of the corresponding complementarity gaps* (μ_k) *converges Q-superlinearly to zero. Moreover, the Q-orders of convergence of these sequences satisfy* $\mathcal{Q}\left(z^k\right) = \mathcal{Q}(\mu_k) \ge 2$.

Proof. From Theorem 3.7 it follows that $(\tau_k, p^k) := (\mu_k, (x^k s^k)/\mu_k) \in \mathcal{K}(\underline{\beta}, \mu_0)$ $\forall k \ge 0$, so that by using Lemma 3.10 we deduce that there is a constant $c = c(\underline{\beta}, \mu_0)$ such that

$$\| u \|_2 \le c\,\mu\,,\ \| v \|_2 \le c\,\mu \qquad \text{at each iteration } k = 0, 1, \ldots. \tag{3.19}$$

For k sufficiently large we have $\mu = \mu_k < 1$, so that $\log \mu_k < 0$, and by using Theorem 3.7, we obtain

$$0 \le \frac{\log \alpha_k}{\log \mu_k} \le \frac{\log\left(\nu\left(\overline{\beta} - \underline{\beta}\right)\right) - \log\left((k + \exp(1) + 1)\log^{1+\nu}(k + \exp(1) + 1)\right)}{\log \mu_0 - \sqrt{\frac{\nu\left(\overline{\beta} - \underline{\beta}\right)k}{n \log^{1+\nu} k}}}\,.$$

It follows that $\lim_{k\to\infty} \frac{\log \alpha_k}{\log \mu_k} = 0$, which implies $\lim_{k\to\infty} \frac{\mu_k}{\alpha_k} = 0$, so that for k sufficiently large we have $\mu = \mu_k < \alpha_k c^{-2} = \alpha c^{-2}$. By using (3.15) and the notation from the proof of Theorem 3.7,

$$0 < \theta < 1 \quad \text{and} \quad \frac{\| q \|_2\, \theta^2}{1-\theta} \le \alpha \text{ imply} \quad z(\theta) \in \mathcal{D}(\beta_+)\,.$$

Given that

$$\frac{\| q \|_2\, \theta^2}{1-\theta} \le \frac{\| u \|_2 \| v \|_2\, \theta^2}{(1-\theta)\mu} \le \frac{c^2 \mu \theta^2}{1-\theta} \le \alpha \qquad \forall \theta \in \left(0,\, 1 - \frac{c^2\mu}{\alpha}\right),$$

it follows that $\overline{\theta} \ge 1 - \frac{c^2\mu}{\alpha}$, and therefore

$$\mu_+ = \mu\left(\overline{\theta}\right) \le \mu\left(1 - \frac{c^2\mu}{\alpha}\right) \le \frac{c^2\mu^2}{\alpha} + \frac{u^T v}{n} \le \left(\frac{1}{\alpha} + \frac{1}{n}\right) c^2 \mu^2 \le \frac{2c^2}{\alpha}\mu^2\,.$$

By taking logarithms we obtain

$$\frac{\log \mu_+}{\log \mu} \geq 2 + \frac{\log\left(2c^2\right)}{\log \mu} - \frac{\log \alpha}{\log \mu}.$$

Since the right-hand side of the above inequality tends to zero as $k \to \infty$, we deduce that $\liminf \frac{\log \mu_{k+1}}{\log \mu_k} \geq 2$. Hence $\mathcal{Q}(\mu_k) \geq 2$. On the other hand, we have $\left\| z^{k+1} - z^k \right\|_2 \leq \left\| w^k \right\|_2 \leq \sqrt{2} c \mu_k$, and by applying Theorem 2 of [28], we deduce the convergence of the sequence $\left(z^k\right)$ to a strictly complementary solution $z^* \in \mathcal{F}^c$ and the fact that $\mathcal{Q}\left(z^k\right) = \mathcal{Q}(\mu_k)$. □

4. Higher order affine scaling methods.

4.1. The higher order affine scaling directions. The higher order affine scaling directions to be considered in this section are related to the higher order derivatives of the central path. As we have seen in section 2, the central path $z(\tau, p) = \lceil x(\tau, p),\, s(\tau, p) \rfloor$ passing through a positive vector $p \in \mathbb{R}^n_{++}$ is analytic in τ if HLCP has a strictly complementary solution, and in $\rho = \sqrt{\tau}$ in the general case. By repeatedly differentiating the equations of the central path,

$$\begin{aligned} x(\tau, p) s(\tau, p) &= \tau p, \\ Q x(\tau, p) + R s(\tau, p) &= b, \end{aligned} \tag{4.1}$$

with respect to τ, we obtain

$$\begin{cases} s(\tau,p)\frac{\partial}{\partial \tau} x(\tau,p) + x(\tau,p)\frac{\partial}{\partial \tau} s(\tau,p) = p, \\ Q\frac{\partial}{\partial \tau} x(\tau,p) + R\frac{\partial}{\partial \tau} s(\tau,p) = 0, \end{cases}$$

$$\begin{cases} s(\tau,p)\frac{\partial^i}{\partial \tau^i} x(\tau,p) + x(\tau,p)\frac{\partial^i}{\partial \tau^i} s(\tau,p) = -\sum_{j=1}^{i-1} \begin{pmatrix} i \\ j \end{pmatrix} \frac{\partial^j}{\partial \tau^j} x(\tau,p) \frac{\partial^{i-j}}{\partial \tau^{i-j}} s(\tau,p), \\ Q\frac{\partial^i}{\partial \tau^i} x(\tau,p) + R\frac{\partial^i}{\partial \tau^i} s(\tau,p) = 0, \end{cases}$$

$$i = 2, 3, \ldots.$$

If we reparameterize the central path equations in terms of $\rho = \sqrt{\tau}$, then the derivatives with respect to ρ of $\overline{z}(\rho, p) = \lceil \overline{x}(\rho, p),\, \overline{s}(\rho, p) \rfloor := z(\rho^2, p)$ are given by

$$\begin{cases} \overline{s}(\rho,p)\frac{\partial}{\partial \rho} \overline{x}(\rho,p) + \overline{x}(\rho,p)\frac{\partial}{\partial \rho} \overline{s}(\rho,p) = 2\rho p, \\ Q\frac{\partial}{\partial \rho} \overline{x}(\rho,p) + R\frac{\partial}{\partial \rho} \overline{s}(\rho,p) = 0, \end{cases}$$

$$\begin{cases} \overline{s}(\rho,p)\frac{\partial^2}{\partial \rho^2} \overline{x}(\rho,p) + \overline{x}(\rho,p)\frac{\partial^2}{\partial \rho^2} \overline{s}(\rho,p) = 2p - 2\frac{\partial}{\partial \rho} \overline{x}(\rho,p)\frac{\partial}{\partial \rho} \overline{s}(\rho,p), \\ Q\frac{\partial^2}{\partial \rho^2} \overline{x}(\rho,p) + R\frac{\partial^2}{\partial \rho^2} \overline{s}(\rho,p) = 0, \end{cases}$$

$$\begin{cases} \overline{s}(\rho,p)\frac{\partial^i}{\partial \rho^i} \overline{x}(\rho,p) + \overline{x}(\rho,p)\frac{\partial^i}{\partial \rho^i} \overline{s}(\rho,p) = -\sum_{j=1}^{i-1} \begin{pmatrix} i \\ j \end{pmatrix} \frac{\partial^j}{\partial \rho^j} \overline{x}(\rho,p) \frac{\partial^{i-j}}{\partial \rho^{i-j}} \overline{s}(\rho,p), \\ Q\frac{\partial^i}{\partial \rho^i} \overline{x}(\rho,p) + R\frac{\partial^i}{\partial \rho^i} \overline{s}(\rho,p) = 0, \end{cases}$$

$$i = 3, 4, \ldots.$$

If HLCP is a reformulation of an LP problem, then it is known that it has a strictly complementary solution. In general it is very difficult to establish whether HLCP has

a strictly complementary solution. However, if this is the case, the algorithm should take advantage of this information. In order to treat the two cases in a unified manner we define

$$\vartheta = \begin{cases} 0 & \text{if HLCP is known to be nondegenerate,} \\ 1 & \text{otherwise.} \end{cases} \tag{4.2}$$

At each step of our higher order affine-scaling interior point method we have a point $z = \lceil\, x,\, s \,\rfloor \in \mathcal{D}(\beta)$, and we consider the vectors

$$w^i = \lceil\, u^i,\, v^i \,\rfloor := \begin{cases} \dfrac{(-1)^i \mu^i}{i!} \left\lceil\, \dfrac{\partial^i}{\partial \tau^i} x(\tau, p),\ \dfrac{\partial^i}{\partial \tau^i} s(\tau, p) \,\right\rfloor \Bigg|_{\substack{p = \frac{xs}{\mu}, \\ \tau = \mu,}} & \text{if } \vartheta = 0, \\[2ex] \dfrac{(-1)^i \mu^{i/2}}{i!} \left\lceil\, \dfrac{\partial^i}{\partial \rho^i} \overline{x}\,(\rho, p),\ \dfrac{\partial^i}{\partial \rho^i} \overline{s}\,(\rho, p) \,\right\rfloor \Bigg|_{\substack{p = \frac{xs}{\mu}, \\ \rho = \sqrt{\mu},}} & \text{if } \vartheta = 1. \end{cases} \tag{4.3}$$

4.2. The higher order algorithm. It is easily seen that the vectors (4.3) can be obtained by solving the following m systems of linear equations:

$$\begin{cases} su^1 + xv^1 = -(1+\vartheta)xs, \\ Qu^1 + Rv^1 = 0, \end{cases}$$

$$\begin{cases} su^2 + xv^2 = \vartheta xs - u^1 v^1, \\ Qu^2 + Rv^2 = 0, \end{cases} \tag{4.4}$$

$$\begin{cases} su^i + xv^i = -\sum_{j=1}^{i-1} u^j v^{i-j}, \\ Qu^i + Rv^i = 0, \end{cases} \qquad i = 3, \ldots, m.$$

The m linear systems above have the same matrix, so that their numerical solution requires only one matrix factorization and m backsolves. This involves $O(n^3) + m\,O(n^2)$ arithmetic operations.

We note that for $\vartheta = 0$, $w^1 = \lceil\, u^1,\, v^1 \,\rfloor$ is just the affine scaling direction considered in section 3.

Given the vectors $w^i = \lceil\, u^i,\, v^i \,\rfloor$ defined by (4.4), we consider the polynomial

$$z(\theta) = z + \sum_{i=1}^{m} w^i \theta^i\,, \tag{4.5}$$

which represents the mth order Taylor expansion around $\theta = 0$ of the function $\theta \to z\,((1-\theta)\mu, xs/\mu)$ in case $\vartheta = 0$, and of the function $\theta \to \overline{z}\,(\sqrt{(1-\theta)\mu}, xs/\mu)$ in case $\vartheta = 1$.

We have $z(0) = z \in \mathcal{D}(\beta)$ and define

$$\check{\theta} = \sup\left\{ \tilde{\theta} \in (0,1] \ :\ z(\theta) \in \mathcal{D}(\beta_+)\ , \forall\, \theta \in [0,\, \tilde{\theta}] \right\}\,, \tag{4.6}$$

where β_+ is given by (3.8). From (4.4)–(4.5) we deduce that

$$x(\theta)s(\theta) = (1-\theta)^{1+\vartheta} xs + \sum_{i=m+1}^{2m} \theta^i h^i \,,$$

$$\mu(\theta) = (1-\theta)^{1+\vartheta} \mu + \sum_{i=m+1}^{2m} \theta^i (e^T h^i/n) \,,$$

$$\text{where } h^i = \sum_{j=i-m}^{m} u^j v^{i-j} \,. \tag{4.7}$$

Therefore the computation of (4.6) involves the solution of a system of polynomial inequalities of order $2m$ in θ. Good lower bounds of the exact solution can be obtained by a line search procedure. In what follows we will give simple lower bounds that will be used in the proof of global convergence.

Given $\check{\theta}$ or a suitable convenient lower bound, we compute

$$\overline{\theta} = \operatorname{argmin} \left\{ \mu(\theta) \;:\; \theta \in [\,0, \check{\theta}\,] \right\} \tag{4.8}$$

and obtain a new point

$$z^+ = z(\overline{\theta}\,) \,. \tag{4.9}$$

We have $z^+ \in \mathcal{D}(\beta_+)$ by construction, and the process can be repeated. Our higher order affine-scaling method is thus defined by the following iterative procedure.

ALGORITHM 2. Given real parameters $0 < \underline{\beta} < \overline{\beta} < 1$, $0 < \nu \le 1$, integer $m \ge 2$, Boolean variable $\vartheta \in \{0, 1\}$, and a vector $z^0 \in \mathcal{D}(\overline{\beta}\,)$:

Set $k \leftarrow 0$, $\beta_0 \leftarrow \overline{\beta}$ and consider the sequence (3.10);
repeat
  Set $z \leftarrow z^k$, $\alpha \leftarrow \alpha_k$, $\beta \leftarrow \beta_k$;
  Compute $w^1, \ldots, w^m$ by solving (4.4);
  Set $\beta_+ = \beta - \alpha$;
  Compute steplength $\overline{\theta}$ from (4.6), (4.8);
  Compute z^+ from (4.9);
  Set $\theta_k \leftarrow \overline{\theta}$, $z^{k+1} \leftarrow z^+$, $\mu_{k+1} \leftarrow \mu(z^+)$, $\beta_{k+1} \leftarrow \beta_+$;
  Set $k \leftarrow k+1$.
continue

4.3. Global convergence. The computational complexity of Algorithm 2 is the same for $\vartheta = 0$ and $\vartheta = 1$. One could eventually obtain slightly better constants if $\vartheta = 0$ and/or if HLCP is skew-symmetric, but in what follows we will obtain bounds in the monotone case that are independent of ϑ.

LEMMA 4.1. *If HLCP is monotone and if* $z = \lceil\, x, s \,\rfloor \in \mathcal{D}(\beta)$, *then the vectors* h^i *defined by* (4.4), (4.7) *satisfy*

$$\| h^i \|_2 \le \frac{2\beta\mu}{i} \left(4\sqrt{n/\beta} \right)^i \,, \quad \left| e^T h^i \right| \le \frac{\beta\mu}{i} \left(4\sqrt{n/\beta} \right)^i \,,$$

$$i = m+1, \ldots, 2m \,.$$

Proof. First let us prove that the quantities $\eta_i := \left\| Du^i + D^{-1}v^i \right\|_2$ satisfy

$$\sqrt{\| Du^i \|_2^2 + \| D^{-1}v^i \|_2^2} \le \eta_i \le 2\overline{\alpha}_i \sqrt{\beta\mu} \left(\frac{1+\vartheta}{2} \sqrt{n/\beta} \right)^i ,$$

$$i = 1, 2, \ldots, m ,$$

where the sequence

$$\overline{\alpha}_i := \frac{1}{i} \begin{pmatrix} 2i-2 \\ i-1 \end{pmatrix} \le \frac{1}{i} 4^i$$

satisfies the following recurrence scheme:

$$\overline{\alpha}_1 = 1, \quad \overline{\alpha}_i = \sum_{j=1}^{i-1} \overline{\alpha}_j \, \overline{\alpha}_{i-j} .$$

The first part of the inequality follows immediately, since by using (4.4) and the monotony of the HLCP, we deduce that $u^{i\,T} v^i \ge 0$. Hence

$$\left\| Du^i + D^{-1}v^i \right\|_2^2 = \left\| Du^i \right\|_2^2 + 2u^{i\,T} v^i + \left\| D^{-1}v^i \right\|_2^2 \ge \left\| Du^i \right\|_2^2 + \left\| D^{-1}v^i \right\|_2^2 .$$

By multiplying the first equations of (4.4) with $(xs)^{-1/2}$, we obtain

$$Du^1 + D^{-1}v^1 = -(1+\vartheta)(xs)^{1/2},$$

$$Du^2 + D^{-1}v^2 = \vartheta(xs)^{1/2} - (xs)^{-1/2} u^1 v^1,$$

$$Du^i + D^{-1}v^i = -(xs)^{-1/2} \sum_{j=1}^{i-1} Du^j D^{-1} v^{i-j} , \quad i = 3, \ldots, m .$$

Because $z \in \mathcal{D}(\beta)$ we have $(xs)^{-1/2} \le 1/\sqrt{\beta\mu}$, and, using Lemma 3.1, we deduce that

$$\eta_1 = (1+\vartheta)\sqrt{n\mu} ,$$

$$\eta_2^2 = \vartheta^2 n\mu - 2\vartheta u^{1\,T} v^1 + \left\| (xs)^{-1/2} u^1 v^1 \right\|_2^2 \le \vartheta^2 n\mu + \frac{\left\| u^1 v^1 \right\|_2^2}{\beta\mu}$$

$$\le \vartheta^2 n\mu + \frac{\eta_1^4}{8\beta\mu} \le \vartheta^2 n\mu + \frac{(1+\vartheta)^4 n^2 \mu}{8\beta} < \frac{(1+\vartheta)^4 n^2 \mu}{4\beta} ,$$

$$\eta_i \le \frac{1}{\sqrt{\beta\mu}} \sum_{j=1}^{i-1} \left\| Du^j \right\|_2 \left\| D^{-1}v^{i-j} \right\|_2 , \quad i = 3, \ldots, m .$$

Since

$$\left\| Du^j \right\|_2 \left\| D^{-1}v^{i-j} \right\|_2 + \left\| Du^{i-j} \right\|_2 \left\| D^{-1}v^j \right\|_2$$

$$\le \left(\left\| Du^j \right\|_2^2 + \left\| D^{-1}v^j \right\|_2^2 \right)^{1/2} \left(\left\| Du^{i-j} \right\|_2^2 + \left\| D^{-1}v^{i-j} \right\|_2^2 \right)^{1/2} \le \eta_j \, \eta_{i-j} ,$$

we obtain

$$\eta_i \le \frac{1}{2\sqrt{\beta\mu}} \sum_{j=1}^{i-1} \eta_j \, \eta_{i-j} \,, \quad i = 3, \ldots, m \,.$$

The required inequalities are then proved by mathematical induction (see [55, 54]).

The upper bound for $\left\| h^i \right\|_2$, $i = m+1, m+2, \ldots, 2m$, can be obtained by writing

$$\begin{aligned}
\left\| h^i \right\|_2 &\le \sum_{j=i-m}^{m} \left\| Du^j \right\|_2 \left\| D^{-1} v^{i-j} \right\|_2 \le \sum_{j=1}^{i-1} \left\| Du^j \right\|_2 \left\| D^{-1} v^{i-j} \right\|_2 \\
&= \frac{1}{2} \sum_{j=1}^{i-1} \left(\left\| Du^j \right\|_2 \left\| D^{-1} v^{i-j} \right\|_2 + \left\| Du^{i-j} \right\|_2 \left\| D^{-1} v^j \right\|_2 \right) \\
&\le \frac{1}{2} \sum_{j=1}^{i-1} \sqrt{\| Du^j \|_2^2 + \| D^{-1} v^j \|_2^2} \sqrt{\| Du^{i-j} \|_2^2 + \| D^{-1} v^{i-j} \|_2^2} \\
&\le \frac{1}{2} \sum_{j=1}^{i-1} \eta_j \, \eta_{i-j} \le 2\beta\mu \left(\frac{1+\vartheta}{2} \sqrt{n/\beta} \right)^i \sum_{j=1}^{i-1} \overline{\alpha}_j \, \overline{\alpha}_{i-j} \\
&= 2\beta\mu \left(\frac{1+\vartheta}{2} \sqrt{n/\beta} \right)^i \overline{\alpha}_i \le \frac{2\beta\mu}{i} \left(2(1+\vartheta)\sqrt{n/\beta} \right)^i \le \frac{2\beta\mu}{i} \left(4\sqrt{n/\beta} \right)^i \,.
\end{aligned}$$

Finally by using Proposition 2.1 and Lemma 3.1 (iv), we have

$$\begin{aligned}
\left| e^T h^i \right| &= \left| \sum_{j=i-m}^{m} u^{j\,T} v^{i-j} \right| \le \frac{1}{2} \sum_{j=1}^{i-1} \left| u^{j\,T} v^{i-j} + u^{i-j\,T} v^j \right| \\
&\le \sum_{j=1}^{i-1} \sqrt{u^{j\,T} v^j} \sqrt{u^{i-j\,T} v^{i-j}} \le \frac{1}{4} \sum_{j=1}^{i-1} \eta_j \, \eta_{i-j} \le \beta\mu \left(\frac{1+\vartheta}{2} \sqrt{n/\beta} \right)^i \sum_{j=1}^{i-1} \overline{\alpha}_j \, \overline{\alpha}_{i-j} \\
&= \beta\mu \left(\frac{1+\vartheta}{2} \sqrt{n/\beta} \right)^i \overline{\alpha}_i \le \frac{\beta\mu}{i} \left(2(1+\vartheta)\sqrt{n/\beta} \right)^i \le \frac{\beta\mu}{i} \left(4\sqrt{n/\beta} \right)^i \,. \qquad \square
\end{aligned}$$

In the following lemma we give a lower bound for the maximum stepsize along the higher order direction.

LEMMA 4.2. *If HLCP is monotone, $z \in \mathcal{D}(\beta)$, and $n \ge 5$, then the maximum stepsize $\check{\theta}$ defined in* (4.6) *satisfies the following inequality:*

$$\check{\theta} \ge \widehat{\theta} := \frac{\beta^{\frac{m-1}{2(m+1)}}}{4\sqrt{n}} \left(\frac{\alpha}{2} \right)^{\frac{1}{m+1}} \,,$$

and

$$\mu\left(\widehat{\theta}\right) \le \left(1 - \widehat{\theta}/2\right) \mu \,.$$

Proof. We have $z(\theta) \in \mathcal{D}(\beta_+)$ if and only if $x(\theta)s(\theta) - (\beta-\alpha)\mu(\theta)e \geq 0$. Using the fact that $z = z(0) \in \mathcal{D}(\beta)$, we can write

$$
\begin{aligned}
& x(\theta)s(\theta) - (\beta-\alpha)\mu(\theta)e \\
&= (1-\theta)^{1+\vartheta} xs + \sum_{i=m+1}^{2m} \theta^i h^i - (\beta-\alpha)\left((1-\theta)^{1+\vartheta}\mu + \frac{1}{n}\sum_{i=m+1}^{2m} \theta^i e^T h^i\right) e \\
&\geq (1-\theta)^{1+\vartheta}\beta\mu e + \sum_{i=m+1}^{2m} \theta^i h^i - (\beta-\alpha)\left((1-\theta)^{1+\vartheta}\mu + \sum_{i=m+1}^{2m} \theta^i \frac{e^T h^i}{n}\right) e \\
&= (1-\theta)^{1+\vartheta}\alpha\mu e + \sum_{i=m+1}^{2m} \theta^i h^i - (\beta-\alpha)\sum_{i=m+1}^{2m} \theta^i \frac{e^T h^i}{n} e \\
&= (1-\theta)^{1+\vartheta}\alpha\mu e + (\beta-\alpha)\sum_{i=m+1}^{2m} \theta^i \left(h^i - \frac{e^T h^i}{n} e\right) + (1-\beta+\alpha)\sum_{i=m+1}^{2m} \theta^i h^i \\
&\geq (1-\theta)^{1+\vartheta}\alpha\mu e - \sum_{i=m+1}^{2m} \theta^i \left\| h^i \right\|_2 e\,.
\end{aligned}
$$

Therefore,

$$
z \in \mathcal{D}(\beta) \;\text{ and }\; \sum_{i=m+1}^{2m} \theta^i \left\| h^i \right\|_2 \leq (1-\theta)^{1+\vartheta}\alpha\mu \quad \text{imply} \quad z(\theta) \in \mathcal{D}(\beta-\alpha)\,. \tag{4.10}
$$

From Lemma 4.1 it follows that

$$
\sum_{i=m+1}^{2m} \theta^i \left\| h^i \right\|_2 \leq 2\beta\mu \sum_{i=m+1}^{2m} \frac{1}{i}\left(4\theta\sqrt{n/\beta}\right)^i\,.
$$

For any $t \in (0,1]$ we have

$$
\sum_{i=m+1}^{2m} \frac{t^i}{i} \leq t^{m+1} \sum_{i=m+1}^{2m} \frac{1}{i} < t^{m+1} \int_m^{2m} \frac{du}{u} = t^{m+1}\log 2 < .7\, t^{m+1}\,.
$$

Hence,

$$
\theta \leq \frac{t\sqrt{\beta}}{4\sqrt{n}} \quad \text{and} \quad t \leq 1 \quad \text{imply} \quad \sum_{i=m+1}^{2m} \theta^i \left\| h^i \right\|_2 \leq 1.4\beta\mu t^{m+1}\,.
$$

On the other hand if $n \geq 5$, then

$$
\theta \leq \frac{t\sqrt{\beta}}{4\sqrt{n}} \quad \text{and} \quad t \leq 1 \quad \text{imply} \quad (1-\theta)^{1+\vartheta} \geq \left(1 - \frac{t\sqrt{\beta}}{4\sqrt{n}}\right)^{1+\vartheta} \geq \left(1 - \frac{1}{4\sqrt{5}}\right)^2 > .7.
$$

By taking $t = (\alpha/(2\beta))^{1/(m+1)}$, we deduce that if $n \geq 5$, then

$$
z(\theta) \in \mathcal{D}(\beta-\alpha) \qquad \forall \theta \in \left[0, \widehat{\theta}\right]\,,
$$

which proves that $\check{\theta} \geq \widehat{\theta}$. Taking again $t = (\alpha/(2\beta))^{1/(m+1)}$, we have $\widehat{\theta} = \frac{t\sqrt{\beta}}{4\sqrt{n}}$, so that we can write

$$\frac{\mu(\widehat{\theta})}{\mu} = \left(1-\widehat{\theta}\right)^{1+\vartheta} + \sum_{i=m+1}^{2m} \widehat{\theta}^{\,i}\, \frac{e^T h^i}{n\mu} \leq 1 - \widehat{\theta} + \frac{\beta}{n} \sum_{i=m+1}^{2m} \frac{t^i}{i}$$

$$< 1 - \frac{t\sqrt{\beta}}{4\sqrt{n}} + \frac{.7\,\beta t^{m+1}}{n} = 1 - \frac{t\sqrt{\beta}}{4\sqrt{n}} \left(1 - \frac{2.8\,\sqrt{\beta} t^m}{\sqrt{n}}\right).$$

Using the definition of t, $n \geq 5$, $0 < \nu \leq 1$, and the fact that

$$\alpha \leq \alpha_0 = \frac{\nu(\overline{\beta} - \underline{\beta})}{(\exp(1)+1)\log^{1+\nu}(\exp(1)+1)} < \frac{1}{(\exp(1)+1)\log^2(\exp(1)+1)},$$

we obtain

$$\frac{2.8\,\sqrt{\beta} t^m}{\sqrt{n}} \leq \frac{2.8\,\sqrt{\beta}}{\sqrt{5}} \left(\frac{\alpha}{2\beta}\right)^{\frac{m}{m+1}} \leq \frac{2.8\,\sqrt{\beta}}{\sqrt{5}} \left(\frac{\alpha}{2\beta}\right)^{\frac{1}{2}} \leq \frac{2\,\sqrt{\alpha}}{\sqrt{5}} \leq \frac{2\,\sqrt{\alpha_0}}{\sqrt{5}} < \frac{1}{2}.$$

It follows that $\mu(\widehat{\theta}) < (1 - \frac{t\sqrt{\beta}}{8\sqrt{n}})\mu$, which completes the proof. □

THEOREM 4.3. *If HLCP is monotone, then Algorithm 2 is well defined and produces a sequence of points* $\left(z^k\right)_{k=0}^{\infty}$, *with* $z^k \in \mathcal{D}(\beta_k) \subset \mathcal{D}(\underline{\beta})$. *If* $n \geq 5$, *then the following relations hold:*

$$\mu_{k+1} \leq \left(1 - \widehat{\theta}_k/2\right)\mu_k, \quad \widehat{\theta}_k \geq \frac{\underline{\beta}^{\frac{m-1}{2(m+1)}}}{4\sqrt{n}} \left(\frac{\alpha_k}{2}\right)^{\frac{1}{m+1}}, \qquad k = 0, 1, \ldots,$$

$$\mu_k \leq \mu_0 \exp\left(-\frac{\overline{\kappa}\nu^{\frac{1}{m+1}} k^{\frac{m}{m+1}}}{\sqrt{n}\log^{\frac{1+\nu}{m+1}} k}\right), \quad \overline{\kappa} := \frac{\underline{\beta}^{\frac{1}{6}} \left(\overline{\beta} - \underline{\beta}\right)^{\frac{1}{3}}}{2^{\frac{10}{3}}}, \qquad k = 8, 9, \ldots.$$

Proof. We have $z^k \in \mathcal{D}(\beta_k)$ by construction and $\mu_{k+1} \leq (1 - \widehat{\theta}_k/2)\mu_k$ by virtue of Lemma 4.2. It follows that

$$\mu_k \leq \mu_0 \prod_{j=0}^{k-1} \left(1 - \widehat{\theta}_j/2\right),$$

and by using the inequality $\log(1-t) < -t$, (3.14), $k \geq 8$, and Corollary 3.3, we obtain

$$\log\left(\frac{\mu_k}{\mu_0}\right) \leq \sum_{j=0}^{k-1} \log\left(1 - \widehat{\theta}_j/2\right) \leq -\frac{1}{2}\sum_{j=0}^{k-1} \widehat{\theta}_j \leq -\frac{\underline{\beta}^{\frac{m-1}{2(m+1)}}}{8\sqrt{n}} \sum_{j=0}^{k-1} \left(\frac{\alpha_j}{2}\right)^{\frac{1}{m+1}}$$

$$< -\frac{\underline{\beta}^{\frac{m-1}{2(m+1)}} \left(\nu(\overline{\beta} - \underline{\beta})\right)^{\frac{1}{m+1}} k^{\frac{m}{m+1}}}{2^{3+\frac{1}{m+1}}\sqrt{n}\log^{\frac{1+\nu}{m+1}} k} \leq -\frac{\underline{\beta}^{\frac{1}{6}} \left(\overline{\beta} - \underline{\beta}\right)^{\frac{1}{3}} \nu^{\frac{1}{m+1}} k^{\frac{m}{m+1}}}{2^{\frac{10}{3}}\sqrt{n}\log^{\frac{1+\nu}{m+1}} k},$$

which completes the proof of our theorem. □

In the next corollary we give an upper bound for the maximum number of iterations, K_ϵ, required by Algorithm 2 to obtain an approximate solution of the HLCP

with duality gap ϵ (see (3.16)). The upper bound is given in terms of the constants $\nu, \overline{\beta}, \underline{\beta}, m$ defining Algorithm 2, the dimension n, and the quantity L_ϵ from (3.17) which depends on ϵ and the starting point.

COROLLARY 4.4. *If* $\sqrt{n}L_\epsilon \geq (\overline{\kappa}\nu^{\frac{1}{m+1}})^{-1}\exp(1)$, *then*

$$K_\epsilon \leq \left\lceil \frac{6^{\frac{1+\nu}{m}}}{\overline{\kappa}^{\frac{m+1}{m}}\nu^{\frac{1}{m}}} \left(\sqrt{n}L_\epsilon\right)^{\frac{m+1}{m}} \log^{\frac{1+\nu}{m}}\left(\sqrt{n}L_\epsilon\right) \right\rceil .$$

Proof. From Theorem 4.3 we deduce that $x^{k\,T}s^k \leq \epsilon$ for any K with the property

$$\frac{K^{\frac{m}{m+1}}}{\log^{\frac{1+\nu}{m+1}} K} \geq \sigma := \left(\overline{\kappa}\nu^{\frac{1}{m+1}}\right)^{-1} \sqrt{n}L_\epsilon .$$

According to Corollary 3.5 it follows that

$$K_\epsilon \leq \overline{K}_\epsilon := \left(2\left(1+\frac{1+\nu}{m+1}\log\frac{1+\nu}{m}\right)\frac{m+1}{m}\log\sigma\right)^{\frac{1+\nu}{m}} \sigma^{\frac{m+1}{m}} .$$

By noticing that under our hypothesis $1+\nu \leq 2 \leq m$, $0 < \overline{\kappa} < 1$, we obtain

$$\sigma \geq \left(\overline{\kappa}\nu^{\frac{1}{m+1}}\right)^{-2}\exp(1) > \exp(1) ,$$

$$\log\sigma = \log\left(\sqrt{n}L_\epsilon\right) + \log\left(\left(\overline{\kappa}\nu^{\frac{1}{m+1}}\right)^{-1}\right)$$

$$< \log\left(\sqrt{n}L_\epsilon\right) + \log\left(\left(\overline{\kappa}\nu^{\frac{1}{m+1}}\right)^{-1}\exp(1)\right) \leq 2\log\left(\sqrt{n}L_\epsilon\right) ,$$

$$\overline{K}_\epsilon \leq (3\log\sigma)^{\frac{1+\nu}{m}} \sigma^{\frac{m+1}{m}} \leq \left(6\log\left(\sqrt{n}L_\epsilon\right)\right)^{\frac{1+\nu}{m}} \sigma^{\frac{m+1}{m}}$$

$$\leq \frac{6^{\frac{1+\nu}{m}}}{\overline{\kappa}^{\frac{m+1}{m}}\nu^{\frac{1}{m}}} \left(\sqrt{n}L_\epsilon\right)^{\frac{m+1}{m}} \log^{\frac{1+\nu}{m}}\left(\sqrt{n}L_\epsilon\right) . \qquad \square$$

If ν is a positive constant independent of n and L_ϵ, then Algorithm 1 is independent of the dimension of the problem and the stopping criterion $x^{k\,T}s^k \leq \epsilon$. However, by letting ν depend on n and L_ϵ, we can slightly improve the computational complexity.

COROLLARY 4.5. *If*

$$\sqrt{n}L_\epsilon \geq \max\left\{\exp\left(\exp(1)\right) , \frac{2.072\exp(1)}{\overline{\kappa}}\log\log\left(\frac{\exp(1)}{\overline{\kappa}}\right)\right\} ,$$

and if we take $\nu = (\log\log(\sqrt{n}L_\epsilon))^{-1}$ *in Algorithm* 2, *then*

$$K_\epsilon \leq \left\lceil \frac{36^{\frac{1}{m}}\exp(1/m)}{\overline{\kappa}^{\frac{m+1}{m}}} \left(\sqrt{n}L_\epsilon\right)^{\frac{m+1}{m}} \left(\log\left(\sqrt{n}L_\epsilon\right)\log\log\left(\sqrt{n}L_\epsilon\right)\right)^{\frac{1}{m}} \right\rceil .$$

Proof. Using Lemma 3.6, we deduce that

$$\frac{\sqrt{n}L_\epsilon}{\log^{\frac{1}{m}}\log(\sqrt{n}L_\epsilon)} \geq \frac{\sqrt{n}L_\epsilon}{\log\log(\sqrt{n}L_\epsilon)} \geq \frac{\exp(1)}{\overline{\kappa}}\,,$$

which shows that Corollary 4.4 can be applied for our choice of ν. By noticing that $\log^{\frac{\nu}{m}}(\sqrt{n}L_\epsilon) = \exp(1/m)$ and $6^{1+\nu} \leq 6^2 = 36$, we obtain the desired result. ∎

COROLLARY 4.6. *Assume that*

$$\sqrt{n}L_\epsilon \geq \max\left\{\exp\left(\exp(1)\right)\,,\ \frac{2.072\exp(1)}{\overline{\kappa}}\log\log\left(\frac{\exp(1)}{\overline{\kappa}}\right)\right\}\,,$$

and consider Algorithm 2 with $\nu = (\log\log(\sqrt{n}L_\epsilon))^{-1}$. *Then the following implications hold:*

$$m \geq \log\log\log(\sqrt{n}L_\epsilon) \Rightarrow K_\epsilon \leq \left\lceil \frac{36^{\frac{1}{m}}\exp(1+1/m)}{\overline{\kappa}^{\frac{m+1}{m}}}\left(\sqrt{n}L_\epsilon\right)^{\frac{m+1}{m}}\log^{\frac{1}{m}}\left(\sqrt{n}L_\epsilon\right)\right\rceil\,,$$

$$m \geq \log\log(\sqrt{n}L_\epsilon) \Rightarrow K_\epsilon \leq \left\lceil \frac{36^{\frac{1}{m}}\exp(2+1/m)}{\overline{\kappa}^{\frac{m+1}{m}}}\left(\sqrt{n}L_\epsilon\right)^{\frac{m+1}{m}}\right\rceil\,,$$

$$m \geq \log(\sqrt{n}L_\epsilon) \Rightarrow K_\epsilon \leq \left\lceil \frac{36^{\frac{1}{m}}\exp(3+1/m)}{\overline{\kappa}^{\frac{m+1}{m}}}\sqrt{n}L_\epsilon\right\rceil\,.$$

4.4. Higher order convergence. As seen in the previous subsection, the computational complexity of Algorithm 2 is basically the same for $\vartheta = 1$ and $\vartheta = 0$. By contrast its asymptotic convergence properties depend on ϑ. In what follows we show that Algorithm 2 with $\vartheta = 1$ is superlinearly convergent for general problems. However, if the problem is known to have a strictly complementary solution, it is advantageous to take $\vartheta = 0$ in order to obtain a higher order of convergence.

LEMMA 4.7. *Assume HLCP is monotone, and consider the linear systems* (4.4), *where we take* $\vartheta = 1$ *for general problems and* $\vartheta = 0$ *for problems that are known to have a strictly complementary solution. Then for any* $\beta \in (0,1)$, *any* $\overline{\mu} > 0$, *and any integer* $m \geq 2$, *there is constant* $c = c(\beta, \overline{\mu}, m)$ *such that the solution* $u^1, v^1, \ldots, u^m, v^m$ *of* (4.4) *satisfies*

$$\left\|\, u^i \,\right\|_2 \leq c\,\mu^{\frac{i}{1+\vartheta}}\,, \quad \left\|\, v^i \,\right\|_2 \leq c\,\mu^{\frac{i}{1+\vartheta}} \qquad \forall z \in \left\{ \lceil\, x,\, s \,\rfloor \in \mathcal{D}(\beta)\ :\ \mu = \frac{x^T s}{n} \leq \overline{\mu}\ \right\}.$$

Proof. Use (4.3) and apply Theorem 2.3 with $\mathcal{K}$ given by (3.18). ∎

THEOREM 4.8. *Assume that HLCP is monotone, and consider Algorithm* 2, *where we take* $\vartheta = 1$ *for general problems and* $\vartheta = 0$ *for problems that are known to have a strictly complementary solution. Then the sequence* $\left(z^k\right)$ *produced by this algorithm converges Q-superlinearly to a maximal complementary solution* $z^* \in \mathcal{F}^c$, *and the sequence of the corresponding complementarity gaps* (μ_k) *converges Q-superlinearly to zero. Moreover, the Q-orders of convergence of these sequences satisfy*

$$\mathcal{Q}\left(z^k\right) = \mathcal{Q}\left(\mu_k\right) \geq \frac{m+1}{1+\vartheta}\,.$$

Proof. Since we are analyzing asymptotic properties we may assume $\mu_k < 1$. By using Theorem 4.3, we obtain

$$0 \leq \frac{\log \alpha_k}{\log \mu_k} \leq \frac{\log\left(\nu\left(\overline{\beta} - \underline{\beta}\right)\right) - \log\left((k+\exp(1)+1)\log^{1+\nu}(k+\exp(1)+1)\right)}{\log \mu_0 - \frac{\overline{\kappa}\nu^{\frac{1}{m+1}} k^{\frac{m}{m+1}}}{\sqrt{n}\log^{\frac{1+\nu}{m+1}} k}}.$$

It follows that $\lim_{k\to\infty} \frac{\log \alpha_k}{\log \mu_k} = 0$, which implies

$$\lim_{k\to\infty} \frac{\mu_k^{\frac{m-\vartheta}{1+\vartheta}}}{\alpha_k} = 0\,.$$

From Theorem 4.3 it follows that $(\tau_k, p^k) := \left(\mu_k, \left(x^k s^k\right)/\mu_k\right) \in \mathcal{K}\left(\underline{\beta}, \mu_0\right)\ \forall k \geq 0$, so that by using Lemma 4.7, we deduce that there is a constant $c = c\left(\underline{\beta}, \mu_0, m\right)$ such that

$$\left\| u^i \right\|_2 \leq c\,\mu^{\frac{i}{1+\vartheta}}\,, \quad \left\| v^i \right\|_2 \leq c\,\mu^{\frac{i}{1+\vartheta}} \quad \text{at each iteration } k = 0, 1, \ldots. \tag{4.11}$$

From (4.7) it follows that

$$\left\| h^i \right\|_2 \leq \sum_{j=i-m}^{m} \left\| u^j \right\|_2 \left\| v^{i-j} \right\|_2 \leq mc^2\mu^{\frac{i}{1+\vartheta}}\,.$$

For k sufficiently large we have $\mu^{\frac{1}{1+\vartheta}} \leq 1/2$, so that

$$\sum_{i=m+1}^{2m} \theta^i \left\| h^i \right\|_2 \leq \sum_{i=m+1}^{2m} \left\| h^i \right\|_2 \leq mc^2\mu^{\frac{m+1}{1+\vartheta}} \sum_{i=0}^{m-1} \mu^{\frac{i}{1+\vartheta}} \leq 2mc^2\mu^{\frac{m+1}{1+\vartheta}} = \overline{c}\,\mu^{\frac{m+1}{1+\vartheta}}\,,$$

with $\overline{c} := 2mc^2$. For k sufficiently large we have $(\overline{c}/\alpha)\mu^{\frac{m-\vartheta}{1+\vartheta}} < 1$, and therefore

$$\sum_{i=m+1}^{2m} \theta^i \left\| h^i \right\|_2 \leq (1-\theta)^{1+\vartheta}\alpha\mu \qquad \forall \theta \in \left[0, \hat{\theta}\right]\,,\ \hat{\theta} = 1 - (\overline{c}/\alpha)^{\frac{1}{1+\vartheta}}\,\mu^{\frac{m-\vartheta}{(1+\vartheta)^2}}\,.$$

According to (4.6) and (4.10) we have $\check{\theta} \geq \hat{\theta}$, so that by using (4.8) we obtain

$$\begin{aligned}
\mu_+ = \mu\left(\overline{\theta}\right) &\leq \mu\left(\hat{\theta}\right) = (1-\hat{\theta})^{1+\vartheta}\mu + \sum_{i=m+1}^{2m} \hat{\theta}^{\,i}(e^T h^i/n) \\
&\leq (1-\hat{\theta})^{1+\vartheta}\mu + \frac{1}{\sqrt{n}} \sum_{i=m+1}^{2m} \hat{\theta}^{\,i} \left\| h^i \right\|_2 \leq (1-\hat{\theta})^{1+\vartheta}\left(1 + \frac{\alpha}{\sqrt{n}}\right)\mu \\
&\leq \overline{c}\left(\frac{1}{\alpha} + \frac{1}{\sqrt{n}}\right)\mu^{\frac{m+1}{1+\vartheta}} \leq \frac{2\overline{c}}{\alpha}\,\mu^{\frac{m+1}{1+\vartheta}}\,.
\end{aligned}$$

By taking logarithms we obtain

$$\frac{\log \mu_+}{\log \mu} \geq \frac{m+1}{1+\vartheta} + \frac{\log(2\overline{c})}{\log \mu} - \frac{\log \alpha}{\log \mu}\,.$$

Since the right-hand side of the above inequality tends to zero as $k \to \infty$, we deduce that

$$\mathcal{Q}(\mu_k) = \liminf \frac{\log \mu_{k+1}}{\log \mu_k} \geq \frac{m+1}{1+\vartheta}.$$

For k sufficiently large we have $(c\mu)^{\frac{1}{1+\vartheta}} < 1 - 1/\sqrt{2}$, so that by using Lemma 4.7, we obtain

$$\left\|z^+ - z\right\|_2 \leq \sum_{i=1}^{m} \left\|w^i\right\|_2 \leq \sqrt{2} \sum_{i=1}^{m} (c\mu)^{\frac{i}{1+\vartheta}} = \sqrt{2} \frac{(c\mu)^{\frac{1}{1+\vartheta}} - (c\mu)^{\frac{m+1}{1+\vartheta}}}{1 - (c\mu)^{\frac{1}{1+\vartheta}}} \leq 2(c\mu)^{\frac{1}{1+\vartheta}}.$$

Finally, by applying Theorem 2 of [28], we deduce the convergence of the sequence $\left(z^k\right)$ to a maximal complementary solution $z^* \in \mathcal{F}^*$ and the fact that $\mathcal{Q}\left(z^k\right) = \mathcal{Q}(\mu_k)$. □

Acknowledgments. The author would like to thank Josef Stoer and two anonymous referees for carefully reading the manuscript and suggesting several improvements.

REFERENCES

[1] M. Anitescu, G. Lesaja, and F. A. Potra, *An infeasible-interior-point predictor-corrector algorithm for the P_*-Geometric LCP*, Appl. Math. Optim., 36 (1997), pp. 203–228.

[2] K. M. Anstreicher and R. A. Bosch, *A new infinity-norm path following algorithm for linear programming*, SIAM J. Optim., 5 (1995), pp. 236–246.

[3] J. F. Bonnans and C. C. Gonzaga, *Convergence of interior point algorithms for the monotone linear complementarity problem*, Math. Oper. Res., 21 (1996), pp. 1–25.

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

[5] A. S. El-Bakry, R. A. Tapia, and Y. Zhang, *A study of indicators for identifying zero variables in interior-point methods*, SIAM Rev., 36 (1994), pp. 45–72.

[6] C. C. Gonzaga, *The largest step path following algorithm for monotone linear complementarity problems*, Math. Program., 76 (1997), pp. 309–332.

[7] C. C. Gonzaga, *Complexity of predictor-corrector algorithms for LCP based on a large neighborhood of the central path*, SIAM J. Optim., 10 (1999), pp. 183–194.

[8] B. Jansen, C. Roos, and T. Terlaky, *A polynomial primal-dual Dikin-type algorithm for linear programming*, Math. Oper. Res., 21 (1996), pp. 341–353.

[9] B. Jansen, C. Roos, and T. Terlaky, *A family of polynomial affine scaling algorithms for positive semidefinite linear complementarity problems*, SIAM J. Optim., 7 (1997), pp. 126–140.

[10] B. Jansen, C. Roos, T. Terlaky, and Y. Ye, *Improved complexity using higher-order correctors for primal-dual Dikin affine scaling*, Math. Program., 76 (1997), pp. 117–130.

[11] J. Ji, F. A. Potra, and S. Huang, *A predictor-corrector method for linear complementarity problems with polynomial complexity and superlinear convergence*, J. Optim. Theory Appl., 84 (1995), pp. 187–199.

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

[13] M. Kojima, N. Megiddo, T. Noma, and A. Yoshise, *A Unified Approach to Interior Point Algorithms for Linear Complementarity Problems*, Lecture Notes in Comput. Sci. 538, Springer-Verlag, New York, 1991.

[14] M. Kojima, S. Mizuno, and A. Yoshise, *A polynomial-time algorithm for a class of linear complementarity problems*, Math. Program., 44 (1989), pp. 1–26.

[15] M. Kojima, S. Mizuno, and A. Yoshise, *A primal-dual interior point algorithm for linear programming*, in Progress in Mathematical Programming (Pacific Grove, CA, 1987) Springer, New York, 1989, pp. 29–47.

[16] M. Kojima, S. Mizuno, and A. Yoshise, *An $O(\sqrt{n}\,L)$ iteration potential reduction algorithm for linear complementarity problems*, Math. Program., 50 (1991), pp. 331–342.

[17] K. McShane, *Superlinearly convergent $O(\sqrt{n}L)$-iteration interior-point algorithms for linear programming and the monotone linear complementarity problem*, SIAM J. Optim., 4 (1994), pp. 247–261.
[18] S. Mizuno, *A superlinearly convergent infeasible-interior-point algorithm for geometrical LCPs without a strictly complementary condition*, Math. Oper. Res., 21 (1996), pp. 382–400.
[19] S. Mizuno and A. Nagasawa, *A primal-dual affine-scaling potential-reduction algorithm for linear programming*, Math. Program., 62 (1993), pp. 119–131.
[20] S. Mizuno, M. J. Todd, and Y. Ye, *On adaptive-step primal-dual interior-point algorithms for linear programming*, Math. Oper. Res., 18 (1993), pp. 964–981.
[21] R. D. C. Monteiro and I. Adler, *Interior path following primal-dual algorithms. I. Linear programming*, Math. Program., 44 (1989), pp. 27–41.
[22] R. D. C. Monteiro and I. Adler, *Interior path following primal-dual algorithms. II. Convex quadratic programming*, Math. Program., 44 (1989), pp. 43–66.
[23] R. D. C. Monteiro, I. Adler, and M. G. C. Resende, *A polynomial-time primal-dual affine scaling algorithm for linear and convex quadratic programming and its power series extension*, Math. Oper. Res., 15 (1990), pp. 191–214.
[24] R. D. C. Monteiro and S. J. Wright, *Local convergence of interior-point algorithms for degenerate monotone LCP*, Comput. Optim. Appl., 3 (1994), pp. 131–155.
[25] R. D. C. Monteiro and S. J. Wright, *Superlinear primal-dual affine scaling algorithms for LCP*, Math. Program., 69 (1995), pp. 311–333.
[26] J. Peng, T. Terlaky, and Y. Zhao, *A predictor-corrector algorithm for linear optimization based on a specific self-regular proximity function*, SIAM J. Optim., 15 (2005), pp. 1105–1127.
[27] F. A. Potra, *On Q-order and R-order of convergence*, J. Optim. Theory Appl., 63 (1989), pp. 415–431.
[28] F. A. Potra, *Q-superlinear convergence of the iterates in primal-dual interior-point methods*, Math. Program., 91 (2001), pp. 99–115.
[29] F. A. Potra, *The Mizuno–Todd–Ye algorithm in a larger neighborhood of the central path*, European J. Oper. Res., 143 (2002), pp. 257–267.
[30] F. A. Potra, *A superlinearly convergent predictor-corrector method for degenerate LCP in a wide neighborhood of the central path with $O(\sqrt{n}l)$-iteration complexity*, Math. Program., 100 (2004), pp. 317–337.
[31] F. A. Potra and X. Liu, *Predictor-corrector methods for sufficient linear complementarity problems in a wide neighborhood of the central path*, Optim. Methods Softw., 20 (2005), pp. 145–168.
[32] F. A. Potra and R. Sheng, *A path following method for LCP with superlinearly convergent iteration sequence*, Ann. Oper. Res., 81 (1998), pp. 97–114.
[33] F. A. Potra and R. Sheng, *Superlinearly convergent infeasible-interior-point algorithm for degenerate LCP*, J. Optim. Theory Appl., 97 (1998), pp. 249–269.
[34] J. Renegar, *A polynomial-time algorithm, based on Newton's method, for linear programming*, Math. Program., 40 (1988), pp. 59–93.
[35] C. Roos, J.-Ph. Vial, and T. Terlaky, *Theory and Algorithms for Linear Optimization: An Interior Point Approach*, Wiley-Interscience Series in Discrete Math. Optim., John Wiley and Sons, 1997.
[36] J. Stoer, *Improved High Order Long-Step Methods for Solving Linear Complementarity Problems*, Technical Report 237, Institut für Angewandte Mathematik und Statistik Universität Würzburg, Würzburg, Germany, 1999.
[37] J. Stoer, *High order long-step methods for solving linear complementarity problems*, Ann. Oper. Res., 103 (2001), pp. 149–159.
[38] J. Stoer and M. Wechs, *The complexity of high-order predictor-corrector methods for solving sufficient linear complementarity problems*, Optim. Methods Softw., 10 (1998), pp. 393–417.
[39] J. Stoer and M. Wechs, *Infeasible-interior-point paths for sufficient linear complementarity problems and their analyticity*, Math. Program., 83 (1998), pp. 407–423.
[40] J. Stoer and M. Wechs, *On the analyticity properties of infeasible-interior-point paths for monotone linear complementarity problems*, Numer. Math., 81 (1999), pp. 631–645.
[41] J. Stoer, M. Wechs, and S. Mizuno, *High order infeasible-interior-point methods for solving sufficient linear complementarity problems*, Math. Oper. Res., 23 (1998), pp. 832–862.
[42] J. F. Sturm, *Superlinear convergence of an algorithm for monotone linear complementarity problems, when no strictly complementary solution exists*, Math. Oper. Res., 24 (1999), pp. 72–94.

[43] K. Tanabe, *Centered Newton method for mathematical programming*, in System Modelling and Optimization (Tokyo, 1987), Lecture Notes in Control and Inform. Sci. 113, Springer, Berlin, 1988, pp. 197–206.
[44] M. J. Todd and Y. Ye, *A centered projective algorithm for linear programming*, Math. Oper. Res., 15 (1990), pp. 508–529.
[45] L. Tunçel, *Constant potential primal-dual algorithms: A framework*, Math. Program., 66 (1994), pp. 145–159.
[46] L. Tunçel, *On the convergence of primal-dual interior-point methods with wide neighborhoods*, Comput. Optim. Appl., 4 (1995), pp. 139–158.
[47] R. H. Tütüncü, *Quadratic convergence of potential-reduction methods for degenerate problems*, Math. Program., 90 (2001), pp. 169–203.
[48] S. J. Wright, *Primal-Dual Interior-Point Methods*, SIAM, Philadelphia, 1997.
[49] Y. Ye, *Interior Point Algorithms: Theory and Analysis*, Wiley-Interscience Series in Discrete Math. Optim., John Wiley and Sons, New York, 1997.
[50] Y. Ye and K. Anstreicher, *On quadratic and $O(\sqrt{n}L)$ convergence of predictor-corrector algorithm for LCP*, Math. Program., 62 (1993), pp. 537–551.
[51] Y. Ye, O. Güler, R. A. Tapia, and Y. Zhang, *A quadratically convergent $O(\sqrt{n}L)$-iteration algorithm for linear programming*, Math. Program., 59 (1993), pp. 151–162.
[52] Y. Zhang, R. A. Tapia, and J. E. Dennis, Jr., *On the superlinear and quadratic convergence of primal-dual interior point linear programming algorithms*, SIAM J. Optim., 2 (1992), pp. 304–324.
[53] Y. Zhang, R. Tapia, and F. Potra, *On the superlinear convergence of interior point algorithms for a general class of problems*, SIAM J. Optim., 3 (1993), pp. 413–422.
[54] G. Zhao, *On the relationship between the curvature integral and the complexity of path-following methods in linear programming*, SIAM J. Optim., 6 (1996), pp. 57–73.
[55] G. Y. Zhao and J. S. Zhu, *The curvature integral and the complexity of linear complementarity problems*, Math. Program., 70 (1995), pp. 107–122.

SIAM J. OPTIM.
Vol. 19, No. 1, pp. 144–162

DUAL CHARACTERIZATION AND SCALARIZATION FOR BENSON PROPER EFFICIENCY*

JING-HUI QIU†

Abstract. By using dual cones and their properties, we establish a fundamental dual characterization and scalarization for Benson proper efficient points without any additional assumption on the ordering cone. From this, we obtain several scalarization theorems and Lagrange multiplier theorems for Benson proper minimizers of optimization problems with nearly cone-subconvexlike set-valued maps. The related known results are improved, and some new criteria for checking Benson proper efficiency are deduced.

Key words. locally convex space, dual cone, scalarization, nearly cone-subconvexlike set-valued map, Benson proper efficiency, Lagrange multiplier theorem

AMS subject classifications. 90C48, 90C29, 90C46, 46A20

DOI. 10.1137/060676465

1. Introduction.

One of the most important problems in vector optimization is to find efficient points of sets. However, some efficient points exhibit certain abnormal properties. In order to get rid of such anomalous efficient points, Kuhn and Tucker [12] and Geoffrion [4] introduced the concept of proper efficiency. Since then, a number of different definitions of proper efficiency have been introduced and investigated. Benson [1] introduced a definition of proper efficiency for vector optimization, which has already been shown to have many desirable properties. There have been significant studies on Benson proper efficiency in vector optimization of vector-valued maps and set-valued maps. For example, under the assumption that the ordering cone has a weakly compact base, Song [21, 22] established a number of interesting criteria for checking Benson proper efficient solutions of vector optimization problems with weakened convex and nonconvex set-valued maps. Chen and Rong [2] and Li [13] characterized the Benson proper efficiency for generalized cone-subconvexlike vector-valued maps and cone-subconvexlike set-valued maps, respectively, in terms of scalarization, Lagrange multiplier, saddle point criterion, and duality under the assumption that the ordering cone is locally compact (or, equivalently, the ordering cone has a compact base). Gasimov [3] considered nonconvex optimization problems in normed spaces ordered by convex cones with bounded bases and characterized the Benson proper efficient points as minimal points of some type of cone-monotone functions. Hernández, Jiménez, and Novo [8] introduced Benson vectorial proper efficiency and proved scalarization theorems for the Benson vectorial proper efficiency in optimization problems with some algebraic type of cone-subconvexlike set-valued maps. Recently, Yang, Li, and Wang [24], Sach [19], and Xu and Liu [23] considered a more extensive class of set-valued maps, i.e., nearly cone-subconvexlike set-valued maps, and presented, respectively, scalarization theorems, saddle point theorems, and Lagrange multiplier theorems for vector optimization problems with such set-valued maps. We observe that in [24, 19, 23] the assumption that the ordering cone has a

*Received by the editors December 1, 2006; accepted for publication (in revised form) July 31, 2007; published electronically February 8, 2008. This research was supported by the National Natural Science Foundation of China (10571035).
http://www.siam.org/journals/siopt/19-1/67646.html
†Department of Mathematics, Suzhou University, Suzhou 215006, P. R. China (qjhsd@sina.com, jhqiu@suda.edu.cn).

compact base or a weakly compact base is also exploited. In this paper, by using the theory of dual cones and the theory of polars (see [11]), we show a fundamental dual characterization and scalarization for Benson proper efficient points without any additional assumption on the ordering cone. From this, we deduce some special scalarization theorems of Benson proper efficiency under various assumptions on the ordering cone or on the underlying space. Applying these results to vector optimization, we obtain scalarization theorems and Lagrange multiplier theorems for Benson proper minimizers of optimization problems with nearly cone-subconvexlike set-valued maps. The related known theorems are improved, and some new versions for checking Benson proper efficiency are deduced. The paper is organized as follows. In section 2, we characterize the solidness of dual cones in the various polar topologies on dual spaces and give the representations of the interiors of dual cones, which are useful for the scalarization of proper efficiency. In section 3, by using the polarity of cones (see [11]), we give a fundamental dual characterization and scalarization theorem for Benson proper efficient points of a given set, where we need only to assume that the ordering cone is a closed convex cone. From this, we deduce some particular scalarization theorems of Benson proper efficiency under various additional assumptions. In particular, we show that our method is also valid for dealing with Benson vectorial proper efficiency. In section 4, applying the results in section 3, we obtain several scalarization theorems for Benson proper minimizers in vector optimization problems with nearly cone-subconvexlike set-valued maps. Finally in section 5, we deduce some Lagrange multiplier theorems.

2. The interior of a dual cone. In order to give the scalarization of Benson proper efficiency, we need to study dual cones and their properties. Dual cones have been studied in many different settings; see, e.g., [20, 10, 6, 7, 16] and their references. We shall characterize the solidness of dual cones and give the representations of the interiors of dual cones in the different polar topologies on duals. First we present some notations (see, e.g., [9, 11, 18, 20]).

In this paper, we always assume, unless stated otherwise, that Y is a real locally convex Hausdorff topological vector space (denoted by l.c.s.), $Y^{\#}$ is its algebraic dual, and Y^* is its topological dual. For any nonempty set $A \subset Y$, $\mathrm{int}A$ and $\mathrm{cl}A$ denote its topological interior and its closure in Y, respectively. Also, $\mathrm{co}(A)$ and $\Gamma(A)$ denote the *convex hull* and the *absolutely convex hull* of A, respectively. The set $\{f \in Y^* : |f(a)| \leq 1 \ \forall a \in A\}$, denoted by A°, is called the *absolute polar* of A taken in Y^*; the set $\{f \in Y^* : f(a) \leq 1 \ \forall a \in A\}$, denoted by A^r, is called the *real polar* of A taken in Y^*. Besides, A^+ and A^- denote the set $\{f \in Y^* : f(a) \geq 0 \ \forall a \in A\}$ and the set $\{f \in Y^* : f(a) \leq 0 \ \forall a \in A\}$, respectively. Mutually, for any nonempty set $A' \subset Y^*$, we may define the absolute polar and the real polar of A' taken in Y as follows: $A'^\circ = \{y \in Y : |f(y)| \leq 1 \ \forall f \in A'\}$ and $A'^r = \{y \in Y : f(y) \leq 1 \ \forall f \in A'\}$. A nonempty set $C \subset Y$ is said to be a cone if $\lambda c \in C$ for any $c \in C$ and any $\lambda \geq 0$ and a convex cone if in addition $C + C \subset C$. A cone C is said to be pointed if $C \cap (-C) = \{0\}$. Here 0 denotes the zero vector of the vector space Y. For a convex cone C in Y, the dual cone of C is defined as

$$C^+ := \{f \in Y^* : f(c) \geq 0 \ \forall c \in C\}.$$

The quasi interior of C^+ is defined as

$$C^{+i} := \{f \in Y^* : f(c) > 0 \ \forall c \in C \backslash \{0\}\}.$$

A convex cone C with $\text{int}C \neq \emptyset$ is said to be a solid cone. A convex subset B of a convex cone C is said to be a base of C if $0 \notin \text{cl}(B)$ and $C = \text{cone}(B)$, where $\text{cone}(B)$ denotes the cone generated by B, i.e., $\text{cone}(B) := [0,\infty)B = \{\lambda b : \lambda \geq 0,\ b \in B\}$. Obviously, a convex cone with a base is pointed. Moreover, a convex cone C has a base if and only if $C^{+i} \neq \emptyset$. For a base B of C, we define B^{st} to be the set

$$\{f \in Y^* : \text{there exists } \delta > 0 \text{ such that } f(b) \geq \delta > 0\ \forall b \in B\}.$$

For any l.c.s. Y, we have an abundance of possible ways of introducing a locally convex topology on the dual Y^*. If $\mathcal{M}$ is any total saturated class of bounded subsets of Y (see [9, 11, 20]), the topology of uniform convergence on the sets M of $\mathcal{M}$ is a locally convex topology on Y^*. We denote it by $\mathcal{T}_{\mathcal{M}}$. Obviously $\{M^\circ : M \in \mathcal{M}\}$ is a 0-neighborhood base in $(Y^*, \mathcal{T}_{\mathcal{M}})$. Particularly, we denote the topologies on Y^* of uniform convergence on bounded subsets, weakly compact (absolutely) convex subsets, and finite subsets of Y by $\beta(Y^*, Y)$, $\tau(Y^*, Y)$, and $\sigma(Y^*, Y)$, which are called the strong topology, Mackey topology, and weak topology, respectively.

LEMMA 2.1. *Let Y be an l.c.s. and $C \subset Y$ be a convex cone. If there exists a locally convex topology $\mathcal{T}$ on Y^* such that $\text{int}_{\mathcal{T}}C^+ \neq \emptyset$, where $\text{int}_{\mathcal{T}}C^+$ denotes the interior of C^+ in $(Y^*, \mathcal{T})$, then $\text{int}_{\mathcal{T}}C^+ \subset C^{+i}$.*

Proof. Take any $f \in \text{int}_{\mathcal{T}}C^+$, and then there exists an absolutely convex 0-neighborhood W in $(Y^*, \mathcal{T})$ such that $f + W \subset C^+$. For any $c \in C\backslash\{0\}$, there exists $g \in Y^*$ such that $g(c) < 0$. Since W is absorbing in Y^*, there exists $\epsilon > 0$ such that $\epsilon g \in W$. Thus

$$f + \epsilon g \in f + W \subset C^+.$$

Hence

$$(f + \epsilon g)(c) \geq 0 \quad \text{and} \quad f(c) \geq -\epsilon g(c) > 0.$$

That is, $f \in C^{+i}$. □

THEOREM 2.1. *Let Y be an l.c.s. and $C \subset Y$ be a convex cone. Then $\text{int}_{\mathcal{T}_{\mathcal{M}}}C^+ \neq \emptyset$ if and only if C has a base $B \in \mathcal{M}$. In this case, $\text{int}_{\mathcal{T}_{\mathcal{M}}}C^+ = B^{st}$.*

Proof. (i) Let C have a base $B \in \mathcal{M}$. Then B is convex, $0 \notin \text{cl}(B)$, and $C = \text{cone}(B)$. By the Hahn–Banach separation theorem, there exists $f \in Y^*$ and $\delta > 0$ such that

$$0 < \delta \leq f(b)\ \ \forall b \in B.$$

Take any fixed ϵ with $0 < \epsilon < \delta$. Since $B \in \mathcal{M}$, ϵB° is a 0-neighborhood in $(Y^*, \mathcal{T}_{\mathcal{M}})$. For any $g \in \epsilon B^\circ$, we have $(f + g)(b) \geq \delta - \epsilon > 0$ for all $b \in B$, which implies that $f + g \in C^+$. That is, $f + \epsilon B^\circ \subset C^+$. Thus, $f \in \text{int}_{\mathcal{T}_{\mathcal{M}}}C^+$ and $\text{int}_{\mathcal{T}_{\mathcal{M}}}C^+ \neq \emptyset$.

(ii) Conversely, assume that $\text{int}_{\mathcal{T}_{\mathcal{M}}}C^+ \neq \emptyset$. Then there exists $f \in \text{int}_{\mathcal{T}_{\mathcal{M}}}C^+$. By Lemma 2.1, $f \in C^{+i}$. Moreover, there exists $M \in \mathcal{M}$ such that $f + M^\circ = f - M^\circ \subset C^+$. Let $B := (f = 1) \cap C$, where $(f = 1)$ denotes the set $\{y \in Y : f(y) = 1\}$. Clearly, B is a base of C. For any $b \in B$ and any $g \in M^\circ$, we have $(f - g)(b) \geq 0$. Thus $g(b) \leq f(b) = 1$ for all $b \in B$, for all $g \in M^\circ$. Since M° is circled, we have $|g(b)| \leq 1$ for all $b \in B$, for all $g \in M^\circ$. From this, $B \subset M^{\circ\circ} = \text{cl}\,\Gamma(M)$. Since $\mathcal{M}$ is saturated, $\text{cl}\,\Gamma(M) \in \mathcal{M}$, and hence $B \in \mathcal{M}$. Thus we have shown that C has a base $B \in \mathcal{M}$.

(iii) Assume that $\text{int}_{\mathcal{T}_{\mathcal{M}}}C^+ \neq \emptyset$ or, equivalently, C has a base $B \in \mathcal{M}$. We are going to prove that $\text{int}_{\mathcal{T}_{\mathcal{M}}}C^+ = B^{st}$. Let $f \in B^{st}$. Then there exists $\delta > 0$ such that

$f(b) \geq \delta$ for all $b \in B$. Take ϵ such that $0 < \epsilon < \delta$. Then $f + \epsilon B^\circ \subset C^+$. Obviously ϵB° is a 0-neighborhood in $(Y^*, \mathcal{T}_\mathcal{M})$ and $f \in \text{int}_{\mathcal{T}_\mathcal{M}} C^+$. Conversely, let $f \in \text{int}_{\mathcal{T}_\mathcal{M}} C^+$. Then there exists $M \in \mathcal{M}$ such that $f - M^\circ \subset C^+$. Since B is a base of C, there exist $g \in Y^*$ and $\delta > 0$ such that $g(b) \geq \delta > 0$ for all $b \in B$. Observe that M° is absorbing in Y^*, and there exists $\epsilon > 0$ such that $\epsilon g \in M^\circ$. Thus $f - \epsilon g \in C^+$. From this, we obtain

$$(f - \epsilon g)(b) \geq 0 \quad \forall b \in B.$$

Hence

$$f(b) \geq \epsilon g(b) \geq \epsilon\delta > 0 \quad \forall b \in B.$$

That is, $f \in B^{st}$. □

In the following we denote by $\text{int}_\tau C^+$ and $\text{int}_\beta C^+$ the interior of C^+ in $(Y^*, \tau(Y^*, Y))$ and in $(Y^*, \beta(Y^*, Y))$, respectively. By Theorem 2.1, we immediately obtain the following corollaries.

COROLLARY 2.1 (refer to [10, Theorem 3.8.6] and [16, Theorem 2.3]). *Let Y be an l.c.s. and $C \subset Y$ be a convex cone. Then $\text{int}_\tau C^+ \neq \emptyset$, i.e., C^+ is a solid cone in $(Y^*, \tau(Y^*, Y))$, if and only if C has a relatively weakly compact base B. In this case, $\text{int}_\tau C^+ = B^{st}$.*

Remark 2.1. Let $C \subset Y$ be a closed convex cone. Then $\text{int}_\tau C^+ \neq \emptyset$ if and only if C has a weakly compact base. In this case, $\text{int}_\tau C^+ = B^{st} = C^{+i}$.

COROLLARY 2.2 (refer to [16, Theorem 2.2]). *Let Y be an l.c.s. and $C \subset Y$ be a convex cone. Then $\text{int}_\beta C^+ \neq \emptyset$ if and only if C has a bounded base B. In this case, $\text{int}_\beta C^+ = B^{st}$.*

3. Dual characterization and scalarization for Benson proper efficiency. Let Y be an l.c.s. and $C \subset Y$ be a convex cone, and then C can specify a partial order in Y as follows:

$$\text{for } \ x, y \in Y, \ \ x \leq_c y \ \text{ if and only if } \ y - x \in C.$$

It is clear that the partial order "$\leq_c$" satisfies the following properties:

(i) $x \leq_c x$;

(ii) if $x \leq_c y$ and $y \leq_c z$, then $x \leq_c z$.

If the ordering cone C is pointed, then the partial order also satisfies the following:

(iii) if $x \leq_c y$ and $y \leq_c x$, then $x = y$.

Moreover, if we assume that the ordering cone C is closed, then we have: if $x_\delta \to x_0$ and every $x_\delta \leq y$, then $x_0 \leq y$.

Let A be a nonempty subset of Y and a point $\bar{y} \in A$. Then $\bar{y} \in A$ is called an efficient point of A with respect to the ordering cone C, denoted by $\bar{y} \in \text{Min}[A, C]$, if $(A - \bar{y}) \cap (-C) = \{0\}$. A point $\bar{y} \in A$ is called a Benson proper efficient point (see [1]) of A with respect to the ordering cone C, denoted by $\bar{y} \in \text{PMin}[A, C]$, if

$$\text{cl cone}(A + C - \bar{y}) \cap (-C) \ = \ \{0\}.$$

Obviously, $\text{PMin}[A, C] \subset \text{Min}[A, C]$. Also, we observe that, if there exists a nonempty subset A of Y such that $\text{PMin}[A, C] \neq \emptyset$, then C must be pointed. First we give the following general dual characterization and scalarization for Benson proper efficiency.

THEOREM 3.1. *Let $C \subset Y$ be a closed convex cone, $\bar{y} \in A \subset Y$, and $\text{co}A$ denote the convex hull of A. Then the following statements are equivalent:*

(i) $\bar{y} \in \text{PMin}[\text{co}A, C]$;

(ii) $(C^+ - C^+ \cap (A-\bar{y})^+)$ *is dense in* $(Y^*, \mathcal{T})$, *where* $\mathcal{T}$ *is any locally convex topology on* Y^* *which is compatible with the dual pair* $\langle Y^*, Y\rangle$ *(i.e.,* $(Y^*, \mathcal{T})^* = Y$*);*

(iii) *for any weakly compact convex set* $K \subset C$ *and* $0 \notin K$, *there exists* $f \in C^+ \cap K^{st}$ *such that* $f(A) \geq f(\bar{y})$*;*

(iv) *for any* $c \in C,\ c \neq 0$, *there exists* $f \in C^+$ *such that* $f(c) > 0$ *and* $f(A) \geq f(\bar{y})$.

Proof. (i)⇒(ii): $\bar{y} \in \mathrm{PMin}[\mathrm{co}A, C]$ means that

$$\mathrm{cl}\,\mathrm{cone}(C + \mathrm{co}(A) - \bar{y}) \cap (-C) = \{0\}. \tag{1}$$

By the theorem of bipolars, (1) is equivalent to

$$[\mathrm{cl}\,\mathrm{cone}(C + \mathrm{co}(A) - \bar{y}) \cap (-C)]^r = Y^*. \tag{2}$$

By the theory of polars (see [11, p. 247]),

$$\begin{aligned}[\mathrm{cl}\,\mathrm{cone}(C + \mathrm{co}(A) - \bar{y}) \cap (-C)]^r &= \mathrm{cl}_{\mathcal{T}}\,\mathrm{co}((C + \mathrm{co}(A) - \bar{y})^- \cup C^+) \\ &= \mathrm{cl}_{\mathcal{T}}(-C^+ \cap (\mathrm{co}(A) - \bar{y})^+ + C^+) \\ &= \mathrm{cl}_{\mathcal{T}}(C^+ - C^+ \cap (A - \bar{y})^+).\end{aligned}$$

Combining this with (2), we have

$$\mathrm{cl}_{\mathcal{T}}(C^+ - C^+ \cap (A-\bar{y})^+) = Y^*, \text{ i.e., } C^+ - C^+ \cap (A-\bar{y})^+ \text{ is dense in } (Y^*, \mathcal{T}).$$

(ii)⇒(iii): Let K be a weakly compact convex subset of C and $0 \notin K$. By Theorem 2.1, $K^{st} = \mathrm{int}_\tau(\mathrm{cone}(K))^+ \neq \emptyset$. By (ii), we have

$$(C^+ \cap (A-\bar{y})^+ - C^+) \cap K^{st} \neq \emptyset.$$

Hence there exist $f \in C^+ \cap (A-\bar{y})^+$ and $g \in C^+$ such that $f - g \in K^{st}$; i.e., there exists $\delta > 0$ such that $(f-g)(K) \geq \delta > 0$. From this we know that $f \in K^{st}$. Obviously $f \in C^+ \cap K^{st}$ and $f(A) \geq f(\bar{y})$.

(iii)⇒(iv): It is obvious.

(iv)⇒(i): Assume that $\bar{y} \notin \mathrm{PMin}[\mathrm{co}A, C]$; then there exists

$$z \in \mathrm{cl}\,\mathrm{cone}(C + \mathrm{co}A - \bar{y}) \cap (-C) \quad \text{and} \quad z \neq 0.$$

Clearly $-z \in C$ and $-z \neq 0$. By (iv), there exists $f \in C^+ \cap (A-\bar{y})^+$ such that $f(-z) > 0$, i.e., $f(z) < 0$. On the other hand, since $z \in \mathrm{cl}\,\mathrm{cone}(C + \mathrm{co}A - \bar{y})$ and $f \in C^+ \cap (A-\bar{y})^+ = C^+ \cap (\mathrm{co}A - \bar{y})^+ \subset [\mathrm{cone}(C + \mathrm{co}A - \bar{y})]^+$, we have $f(z) \geq 0$, a contradiction! □

From Theorem 3.1 we can deduce some particular scalarization theorem for Benson proper efficiency.

THEOREM 3.2. *Let* $C \subset Y$ *be a closed convex cone and* $\bar{y} \in A \subset Y$. *If there exists a locally convex topology* $\mathcal{T}$ *on* Y^* *such that* $(Y^*, \mathcal{T})^* = Y$ *and* $\mathrm{int}_{\mathcal{T}} C^+ \neq \emptyset$, *then the following statements are equivalent:*

(i) $\bar{y} \in \mathrm{PMin}[\mathrm{co}A, C]$*;*

(ii) *there exists* $f \in \mathrm{int}_{\mathcal{T}} C^+$ *such that* $f(A) \geq f(\bar{y})$*;*

(iii) *there exists* $f \in C^{+i}$ *such that* $f(A) \geq f(\bar{y})$.

Proof. (i)⇒(ii): Since $\bar{y} \in \mathrm{PMin}[\mathrm{co}A, C]$, by Theorem 3.1,

$$C^+ - C^+ \cap (A-\bar{y})^+ \text{ is dense in } (Y^*, \mathcal{T}).$$

By the assumption that $\text{int}_{\mathcal{T}}C^+ \neq \emptyset$, we have

$$(C^+ - C^+ \cap (A-\bar{y})^+) \cap (-\text{int}_{\mathcal{T}}C^+) \neq \emptyset.$$

That is, there exists $g \in C^+$, $f \in C^+ \cap (A-\bar{y})^+$, and $h \in \text{int}_{\mathcal{T}}C^+$ such that $g-f=-h$. Thus

$$f = g + h \in C^+ + \text{int}_{\mathcal{T}}C^+ = \text{int}_{\mathcal{T}}C^+ \quad \text{and} \quad f(A) \geq f(\bar{y}).$$

(ii)⇒(iii): By Lemma 2.1, $\text{int}_{\mathcal{T}}C^+ \subset C^{+i}$. Hence the implication is obvious.

(iii)⇒(i): Assume that there exists $f \in C^{+i}$ such that $f(A) \geq f(\bar{y})$. Since $f \in C^{+i} \subset C^+$ and $f(A-\bar{y}) \geq 0$, we have $f(C + \text{co}A - \bar{y}) \geq 0$ and hence

$$f(\text{cl cone}(C + \text{co}A - \bar{y})) \geq 0. \tag{3}$$

On the other hand, $f(C) \geq 0$ and hence

$$f(-C) \leq 0. \tag{4}$$

Let $z \in \text{cl cone}(C + \text{co}A - \bar{y}) \cap (-C)$, and then by (3) and (4) we have $f(z) \geq 0$ and $f(z) \leq 0$, which leads to $f(z) = 0$. Since $-z \in C$, $f(-z) = 0$, and $f \in C^{+i}$, we conclude that $-z = 0$ and hence $z = 0$. Thus we have shown that

$$\text{cl cone}(C + \text{co}A - \bar{y}) \cap (-C) = \{0\}. \qquad \square$$

COROLLARY 3.1. *Let $C \subset Y$ be a closed convex cone with a weakly compact base B and $\bar{y} \in A \subset Y$. Then the following statements are equivalent:*

(i) $\bar{y} \in \text{PMin}[\text{co}A, C]$;

(ii) *there exists $f \in B^{st}$ such that $f(A) \geq f(\bar{y})$;*

(iii) *there exists $f \in C^{+i}$ such that $f(A) \geq f(\bar{y})$.*

Proof. The result follows immediately from Corollary 2.1, Remark 2.1, and Theorem 3.2. $\square$

COROLLARY 3.2 (refer to [24, Theorem 6.2]). *Let $C \subset Y$ be a closed convex cone with a compact base B and $\bar{y} \in A \subset Y$. Then the statements* (i), (ii), *and* (iii) *in Corollary* 3.1 *are equivalent.*

As we know, every compact set is weakly compact, but a weakly compact set may be noncompact. In fact, the closed unit ball (or its translation) in an infinite-dimensional reflexive Banach space is weakly compact and not compact. Hence Corollary 3.1 contains and improves Corollary 3.2. In the following we construct a closed convex pointed cone C having a base B with the following properties: $\text{int}C = \emptyset$, B is weakly compact, B is not compact, and $\text{int}(C^+) \neq \emptyset$.

Example 3.1. As in [14, p. 326], a matrix $A = (a_{j.k})_{j,\,k\in N}$ of nonnegative numbers is called a Köthe matrix if A satisfies the following conditions:

(i) for each $j \in N$, there exists a $k \in N$ such that $a_{j,k} > 0$;

(ii) $a_{j,k} \leq a_{j,k+1}$ for all $j, k \in N$.

Now let $A = (a_{j,k})_{j,k\in N}$ be a given Köthe matrix such that

(C1) $a_{2j,\,k} < \frac{1}{j}\, a_{2j,\,k+1}$, for all $j, k \in N$;

(C2) $a_{2j-1,\,k} = a_{2j-1,\,k+1}$.

For $1 < p < \infty$, we define

$$\lambda^p(A) := \left\{ x = (x_j)_{j\in N} \in R^N : |x\|_k := \left(\sum_{j=1}^{\infty} |x_j a_{j,k}|^p \right)^{\frac{1}{p}} < \infty \ \forall k \in N \right\}.$$

Then $\lambda^p(A)$ is a reflexive Fréchet space (see [14, Lemma 27.1 and Proposition 27.3]). By condition (C1) we see that the topology induced by $\|\ \|_{k+1}$ is strictly finer than one induced by $\|\ \|_k$, and hence we conclude that $\lambda^p(A)$ is not normable. By condition (C2) and [14, Theorem 27.9], we know that $\lambda^p(A)$ is not a Montel space. Here a reflexive l.c.s. in which every closed bounded set is compact is called a Montel space (see, e.g., [20, p. 141]). Thus there exists a closed convex bounded set B in $\lambda^p(A)$ which is not compact. Without loss of generality, we may assume that $0 \notin B$. Since $\lambda^p(A)$ is reflexive, we know that B is weakly compact. Let $C := \text{cone}(B)$. Then C is a closed convex pointed cone, and C has a weakly compact base B which is noncompact. Moreover, we assert that $\text{int}C = \emptyset$. If not, then $\text{int}C \neq \emptyset$, with C having a bounded base B, will imply that $\lambda^p(A)$ is normable, contradicting condition (C1). Finally, since B is weakly compact, by Corollary 2.1 we know that $\text{int}_\tau(C^+) \neq \emptyset$ (certainly, we also have $\text{int}_\beta C^+ \neq \emptyset$).

When Y is a separable normed space, we can obtain a scalarization result for Benson proper efficient points without any restriction on the ordering cone.

COROLLARY 3.3 (see [5, Theorem 3.2]). *Let $(Y, \|\ \|)$ be a separable normed space, $C \subset Y$ be a closed convex pointed cone, and $\bar{y} \in A \subset Y$. Then $\bar{y} \in \text{PMin}[\text{co}A, C]$ if and only if there exists $f \in C^{+i}$ such that $f(A) \geq f(\bar{y})$.*

Proof. Assume that $\bar{y} \in \text{PMin}[\text{co}A, C]$. Then by Theorem 3.1, for any $c \in C,\ c \neq 0$, there exists $f \in C^+ \cap (A - \bar{y})^+$ such that $f(c) > 0$. Put $\bar{f} = f/\|f\|$. Then $\bar{f} \in U^\circ \cap C^+ \cap (A - \bar{y})^+$ and $\bar{f}(c) > 0$, where U denotes the closed unit ball in $(Y, \|\ \|)$. Since $(Y, \|\ \|)$ is separable, $U^\circ \cap C^+ \cap (A - \bar{y})^+$ with the topology induced by $\sigma(Y^*, Y)$ is a compact metric space and hence is separable. Let $\{f_1, f_2, \ldots\}$ be a countable dense subset of $U^\circ \cap C^+ \cap (A - \bar{y})^+$. Then for any $c \in C,\ c \neq 0$, there exists $n \in N$ such that $f_n(c) > 0$. Since $\{f_1, f_2, \ldots\} \subset U^\circ$ and U is absorbing, we may define

$$f(x) = \sum_{n=1}^{\infty} \frac{1}{2^n} f_n(x), \quad x \in Y.$$

Then $f \in U^\circ \cap C^+ \cap (A - \bar{y})^+$ and for all $c \in C \backslash \{0\}$, $f(c) > 0$. Clearly $f \in C^{+i}$ and $f(A) \geq f(\bar{y})$.

Conversely, assume that there exists $f \in C^{+i}$ such that $f(A) \geq f(\bar{y})$. Then as shown in the proof of (iii)⇒(i) in Theorem 3.1, we can deduce that $\bar{y} \in \text{PMin}[\text{co}A, C]$. □

Remark 3.1. In fact, by using the theory of strictly extreme points and strictly exposed points (see [17]), we see that every closed convex pointed cone in separable normed spaces automatically has a base.

Next we try to give another class of scalarization results for Benson proper efficiency. A closed convex cone C in Y is said to have a *countable weakly compact base* if there exists a base B of C such that $B = \cup_{n=1}^{\infty} K_n$, where every K_n is a weakly compact convex set. As in [15, p. 249], an l.c.s. Y is said to have the *countable neighborhood property* (denoted by c.n.p.) if, given any sequence $\{U_n\}_{n \in N}$ of 0-neighborhoods in Y, there are $\alpha_n > 0$ such that $U := \cap_{n=1}^{\infty} \alpha_n U_n$ is a 0-neighborhood in Y. We know that every (DF)-space has the c.n.p. (see [15, p. 249]). The class of (DF)-space was introduced by Grothendieck (see [20, p. 154]). It comprises all of the strong duals of metrizable locally convex spaces and all normed spaces; for details, see [9, 11, 14, 20]. Now we introduce a weaker property, the *weak countable neighborhood property*, as follows: an l.c.s. Y is said to have the weak countable neighborhood property (denoted by w.c.n.p.) if, given any sequence $\{f_n\}_{n \in N}$ in Y^*, there are $\alpha_n > 0$ such that $U := \cap_{n=1}^{\infty} \alpha_n f_n^\circ$ is a 0-neighborhood in $(Y, \tau(Y, Y^*))$. Here f_n° denotes the polar of

the singleton $\{f_n\}$, i.e., $f_n^\circ = \{y \in Y : |f(y)| \leq 1\}$. Obviously, the c.n.p. implies the w.c.n.p., and the w.c.n.p. is duality invariant. Also, we observe that an l.c.s. Y has the w.c.n.p. if and only if for every sequence $\{f_n\}_{n\in N}$ in Y^* there are $\lambda_n > 0$ such that the closed (absolutely) convex hull of $\{\lambda_n f_n\}_{n\in N}$ in $(Y^*, \sigma(Y^*, Y))$ is $\sigma(Y^*, Y)$-compact.

Example 3.2 (a closed convex pointed cone with a countable weakly compact base and without any bounded base). Let $X := (l^1, \|\ \|_1)$ and $Y = X^* = l^\infty$ with the Mackey topology $\tau(l^\infty, l^1)$. Let $C = \{y = (\eta_n) \in l^\infty : \eta_n \geq 0\ \forall n \in N\}$. Then clearly C is a convex pointed cone. Let $e_n \in l^1$ denote the sequence whose only nonzero term is a 1 in the nth place. Then $C = \cap_{n=1}^\infty \{y \in l^\infty : \langle y, e_n\rangle \geq 0\}$, as the intersection of the closed sets $\{y \in l^\infty : \langle y, e_n\rangle \geq 0\}$, is closed in Y. Thus, C is a closed convex pointed cone in Y. We shall see that C has no bounded base, but it has a countable weakly compact base. If C has a bounded base, then by Corollary 2.2, $\mathrm{int}_\beta C^+ \neq \emptyset$. Clearly, $C^+ = \{x = (\xi_n) \in l^1 : \xi_n \geq 0\ \forall n \in N\}$. Let $x = (\xi_n) \in \mathrm{int}_\beta C^+$. Then there exists $\epsilon > 0$ such that $x + \epsilon U \subset C^+$, where U denotes the closed unit ball in $(l^1, \|\ \|_1)$, i.e., $U = \{x \in l^1 : \|x\|_1 \leq 1\}$. Since $\xi_n \to 0$, we may take n large enough such that $0 \leq \xi_n < \epsilon$. Thus, $\xi_n - \epsilon < 0$, and hence $x - \epsilon e_n \notin C^+$, which contradicts that $x + \epsilon U \subset C^+$. This shows that C has not any bounded base. Take fixed $x' = (\xi_n') \in l^1$, with every $\xi_n' > 0$, and let $B := \{y = (\eta_n) \in C : \langle y, x'\rangle = \sum_{n=1}^\infty \xi_n' \eta_n = 1\}$. Then B is a base of C. We observe that $B = \cup_{n=1}^\infty (B \cap nU^\circ)$, where every $B \cap nU^\circ$ is $\sigma(l^\infty, l^1)$-compact, i.e., $\sigma(Y, Y^*)$-compact. That is, B is a countable weakly compact base of C.

Example 3.3 (an l.c.s. has the w.c.n.p. but does not have the c.n.p.). Let $(Y, \mathcal{T})$ be an infinite dimensional l.c.s. such that the strong dual $(Y^*, \beta(Y^*, Y))$ is a complete metrizable l.c.s. For any sequence $\{f_n\}_{n\in N}$ in Y^*, there are $\lambda_n > 0$ such that $\lambda_n f_n \to 0$ in $(Y^*, \beta(Y^*, Y))$. Since $(Y^*, \beta(Y^*, Y))$ is complete and metrizable, the closed convex hull of $\{\lambda_n f_n\}_{n\in N}$ is compact in $(Y^*, \beta(Y^*, Y))$ and hence compact in $(Y^*, \sigma(Y^*, Y))$. Thus $(Y, \mathcal{T})$ has the w.c.n.p. From this we know that $(Y, \sigma(Y, Y^*))$ has the w.c.n.p. However, we shall see that $(Y, \sigma(Y, Y^*))$ does not have the c.n.p. Assume the contrary. Let $\{f_n\}_{n\in N} \subset Y^*$ be a countable infinite set whose members are linearly independent. Then there exist $\alpha_n > 0$ such that $\cap_{n=1}^\infty \alpha_n f_n^\circ$ is a 0-neighborhood in $(Y, \sigma(Y, Y^*))$, where f_n° denotes the polar of the singleton $\{f_n\}$. Hence there exists $\epsilon > 0$ such that

$$\epsilon \bigcap_{i=1}^m g_i^\circ = \bigcap_{i=1}^n (|g_i| \leq \epsilon) \subset \bigcap_{n=1}^\infty \alpha_n f_n^\circ.$$

Taking polars in the two sides of the above containment, we obtain

$$\left(\epsilon \bigcap_{i=1}^m g_i^\circ\right)^\circ \supset \left(\bigcap_{n=1}^\infty \alpha_n f_n^\circ\right)^\circ.$$

By using the theory of polars (see [11, p. 247]), we have

$$\frac{1}{\epsilon} \mathrm{cl}\,\Gamma\left(\bigcup_{i=1}^m g_i^{\circ\circ}\right) \supset \mathrm{cl}\,\Gamma\left(\bigcup_{n=1}^\infty \frac{1}{\alpha_n} f_n^{\circ\circ}\right).$$

From this,

$$\bigcup_{n=1}^\infty \frac{1}{\alpha_n} f_n \subset \frac{1}{\epsilon}\,\Gamma(\{g_1, g_2, \ldots, g_m\}).$$

This contradicts the assumption that the set $\{f_n : n \in N\}$ is linearly independent.

COROLLARY 3.4. *Let Y be an l.c.s. with the w.c.n.p. and $C \subset Y$ be a closed convex cone with a countable weakly compact base B. Then $\bar{y} \in \mathrm{PMin}[\mathrm{co}A, C]$ if and only if there exists $f \in C^{+i}$ such that $f(A) \geq f(\bar{y})$.*

Proof. We need only to give the proof of one direction. We assume that $B = \cup_{n=1}^{\infty} K_n$, where every K_n is a weakly compact set. Let $\bar{y} \in \mathrm{PMin}[\mathrm{co}A, C]$. By Theorem 3.1, for every $K_n \subset C$, there exists $f_n \in C^+$ and $\delta_n > 0$ such that $f_n(K_n) \geq \delta_n$ and $f_n(A) \geq f_n(\bar{y})$. Since Y has the w.c.n.p., there are $\lambda_n > 0$ such that the closed convex hull of $\{\lambda_n f_n\}_{n\in N}$ in $(Y^*, \sigma(Y^*, Y))$ is $\sigma(Y^*, Y)$-compact. From this, we can deduce that

$$\sum_{n=1}^{\infty} \frac{1}{2^n} \lambda_n f_n \ \text{ is convergent in } \ (Y^*, \sigma(Y^*, Y)).$$

Let $f := \sum_{n=1}^{\infty} \frac{1}{2^n} \lambda_n f_n$. Then $f \in Y^*$ and $f(B) > 0$ for all $b \in B$. Clearly, $f \in C^{+i}$ and $f(A) \geq f(\bar{y})$. □

COROLLARY 3.5. *Let (X, d) be a Fréchet space (i.e., a complete metrizable l.c.s.) and $Y = X^*$ with a locally convex topology $\mathcal{T}$ such that $(Y, \mathcal{T})^* = X$. Let $C \subset Y$ be a closed convex cone with a base B. Then $\bar{y} \in \mathrm{Pmin}[\mathrm{co}A, C]$ if and only if there exists $f \in C^{+i}$ such that $f(A) \geq f(\bar{y})$.*

Proof. Since $(Y^*, \beta(Y^*, Y)) = (X, d)$ is a complete metrizable l.c.s., as shown in Example 3.3, we know that Y has the w.c.n.p. On the other hand, let $U_1 \supset U_2 \supset \cdots$ be a base of 0-neighborhoods in (X, d), and then $Y = \cup_{n=1}^{\infty} n U_n^{\circ}$. Thus $B = \cup_{n=1}^{\infty} (n U_n^{\circ} \cap B) = \cup_{n=1}^{\infty} K_n$, where $K_n := n U_n^{\circ} \cap B$ is $\sigma(X^*, X)$-compact, i.e., $\sigma(Y, Y^*)$-compact. Applying Corollary 3.4, we conclude that $\bar{y} \in \mathrm{PMin}[\mathrm{co}A, C]$ if and only if there exists $f \in C^{+i}$ such that $f(A) \geq f(\bar{y})$. □

By using Corollary 3.5, we immediately obtain the following.

COROLLARY 3.6. *Let Y be a semireflexive (DF)-space (particularly, let Y be a reflexive Banach space). Let $C \subset Y$ be a closed convex cone with a base. Then $\bar{y} \in \mathrm{PMin}[\mathrm{co}A, C]$ if and only if there exists $f \in C^{+i}$ such that $f(A) \geq f(\bar{y})$.*

Remark 3.2. Even for reflexive Banach spaces, we cannot expect to obtain the result as for separable normed spaces (see Corollary 3.3), since in such spaces there might exist a closed convex pointed cone without any base.

Example 3.4 (see [11, p. 137]). Let A be an uncountable index set of cardinality d, and let $x = (\xi_\alpha)_{\alpha\in A}$ be a vector with d coordinates, of which at most countably many are nonzero. For $1 < p < \infty$, define

$$\|x\|_p = \left(\sum_{\alpha\in A} |\xi_\alpha|^p \right)^{\frac{1}{p}}.$$

Let l_d^p denote the space $\{x = (\xi_\alpha)_{\alpha\in A} : \|x\|_p < \infty\}$ with the norm $\| \ \|_p$. Then l_d^p is a Banach space with the dual l_d^q, where $\frac{1}{p} + \frac{1}{q} = 1$. Obviously l_d^p is a reflexive Banach space. Put $C := \{x = (\xi_\alpha) \in l_d^p : \xi_\alpha \geq 0 \ \forall \alpha \in A\}$. Then C is a closed convex pointed cone in l_d^p. However, C has no bases. If not, there exists $f = (\eta_\alpha)_{\alpha\in A} \in (l_d^p)^* \cong l_d^q$ such that $f \in C^{+i}$. Particularly, for every $\alpha \in A$, $f(e_\alpha) > 0$, i.e., $\eta_\alpha > 0$. But this is impossible since at most countably many η_α are nonzero.

Remark 3.3. Yang, Li, and Wang [24] introduced *nearly C-subconvexlike set-valued maps*, and Sach [19] introduced *nearly C-subconvexlike sets*. Let Y be an l.c.s. and $\bar{y} \in A \subset Y$. The set A is said to be *nearly C-subconvexlike* at $\bar{y}$ if $\mathrm{cl\,cone}(A - \bar{y} + C)$

is convex. In this case, $\operatorname{cl}\operatorname{cone}(A-\bar{y}+C) = \operatorname{cl}\operatorname{cone}(\operatorname{co}A-\bar{y}+C)$, and hence the statement $\bar{y}\in \operatorname{PMin}[\operatorname{co}A, C]$ in Theorems 3.1 and 3.2 and Corollaries 3.1, 3.2, 3.3, 3.4, 3.5, and 3.6 can be replaced by $\bar{y}\in \operatorname{PMin}[A, C]$.

In [8], Hernández, Jiménez, and Novo introduced and studied the Benson-vectorial proper efficiency in real linear spaces. We shall see that our method (for example, Theorem 3.2) also can be applied to the scalarization of Benson-vectorial proper efficiency. As in [8], let Y be a real linear space and $A\subset Y$ be nonempty. The *core* (*algebraic interior*) and the *intrinsic core* (*relatively algebraic interior*) of A are defined, respectively, as follows:

$$\operatorname{cor}(A)=\{y\in A:\ \forall v\in Y,\ \exists t>0,\ \forall \alpha\in[0,t],\ y+\alpha v\in A\},$$

$$\operatorname{icr}(A)=\{y\in A:\ \forall v\in \operatorname{span}[A-A],\ \exists t>0,\ \forall \alpha\in[0,t],\ y+\alpha v\in A\}.$$

The *vectorial closure* of A is defined as follows:

$$\operatorname{vcl}(A)\ =\ \{y\in Y;\ \exists v\in Y,\ \forall t>0,\ \exists \alpha\in(0,t],\ y+\alpha v\in A\}.$$

If A is convex, then $y\in\operatorname{vcl}(A)$ if and only if there exists $a\in A$ such that $[a,y)\subset A$. A set A is said to be *vectorially closed* if $A=\operatorname{vcl}(A)$. If Y is endowed with the finest locally convex topology γ (which is also called *convex core topology*), then $(Y,\gamma)^*=Y^{\#}$; i.e., every linear functional on (Y,γ) is continuous (see [18, pp. 22 and 43]). Let C be a convex cone in Y. Then in this case we have

$$C^+=\{f\in Y^{\#}:\ f(c)\geq 0\ \forall c\in C\};$$

$$C^{+i}=\{f\in Y^{\#}:\ f(c)>0\ \forall c\in C\backslash\{0\}\}.$$

We need the following two lemmas.

LEMMA 3.1. *If $B\subset Y$ is convex and $\operatorname{icr}(B)\neq\emptyset$, then $\operatorname{vcl}(B)=\operatorname{cl}_\gamma(B)$.*

Proof. Obviously, $\operatorname{vcl}(B)\subset\operatorname{cl}_\gamma(B)$. Conversely, assume that $x\notin\operatorname{vcl}(B)$, and we shall see that $x\notin\operatorname{cl}_\gamma(B)$. For convenience, we assume that $x=0$. Take any fixed $x_0\in\operatorname{icr}(B)$. Then $[x_0,0)\not\subset B$. Hence there exists $\lambda,\ 0<\lambda<1$, such that $\lambda x_0\notin B$ and certainly $\lambda x_0\notin\operatorname{icr}(B)$. By the separation theorem in real linear spaces, there exists $f\in Y^{\#}$ such that $f(\operatorname{icr}(B))<f(\lambda x_0)$. Particularly, we have

$$f(x_0)\ <\ f(\lambda x_0)\ =\ \lambda f(x_0).$$

From this, we conclude that $f(x_0)<0$. Thus

$$f(\operatorname{icr}(B))\ <\ \lambda f(x_0)\ <\ 0,$$

which implies that $f(B)\leq\lambda f(x_0)<0$ and hence $0\notin\operatorname{cl}_\gamma(B)$. □

LEMMA 3.2. *If $\operatorname{cor}(C^+)\neq\emptyset$, then $\operatorname{cor}(C^+)=\operatorname{int}_\sigma(C^+)=\operatorname{int}_\tau(C^+)=\operatorname{int}_\beta(C^+)$. Here $\operatorname{int}_\sigma(C^+)$, $\operatorname{int}_\tau(C^+)$, and $\operatorname{int}_\beta(C^+)$ denote, respectively, the interiors of C^+ in $(Y^{\#},\sigma(Y^{\#},Y))$, $(Y^{\#},\tau(Y^{\#},Y))$, and $(Y^{\#},\beta(Y^{\#},Y))$.*

Proof. Let $f\in\operatorname{cor}(C^+)$. Then there exists an absolutely convex absorbing set W in $Y^{\#}$ such that $f+W\subset C^+$. Since C^+ is $\sigma(Y^{\#},Y)$-closed, we have $f+\operatorname{cl}_\sigma(W)\subset C^+$, where $\operatorname{cl}_\sigma(W)$ denotes the closure of W in $(Y^{\#},\sigma(Y^{\#},Y))$. In fact, on $Y^{\#}$, the topologies $\beta(Y^{\#},Y)$, $\tau(Y^{\#},Y)$, and $\sigma(Y^{\#},Y)$ coincide, and they are barrelled spaces (see [18, pp. 64 and 75]). Thus the barrel $\operatorname{cl}_\sigma(W)$ is a 0-neighborhood in $(Y^{\#},\sigma(Y^{\#},\sigma(Y^{\#},Y))$ and $f\in\operatorname{int}_\sigma C^+$. The converse containment $\operatorname{int}_\sigma(C^+)\subset\operatorname{cor}(C^+)$ is obvious. □

As in [8], a point $\bar{y} \in A$ is called a Benson-vectorial proper efficient point of A with respect to the convex cone C, denoted by $\bar{y} \in \mathrm{PVMin}[A, C]$, if $\mathrm{vcl}[\mathrm{cone}(A - \bar{y} + C)] \cap (-C) = \{0\}$. Now we can give a scalarization result for Benson-vectorial proper efficient points.

COROLLARY 3.7. *Let Y be a real linear space, $C \subset Y$ be a vectorially closed cone with $\mathrm{cor}(C^+) \neq \emptyset$, and $\bar{y} \in A \subset Y$. Moreover, assume that $\mathrm{icr}[\mathrm{cone}(A-\bar{y})+\mathrm{icr}(C)] \neq \emptyset$. Then the following statements are equivalent:*

(i) $\bar{y} \in \mathrm{PVMin}[\mathrm{co}A, C]$;

(ii) *there exists $f \in \mathrm{cor}(C^+)$ such that $f(A) \geq f(\bar{y})$;*

(iii) *there exists $f \in C^{+i}$ such that $f(A) \geq f(\bar{y})$.*

In particular, if $\mathrm{cone}(A-\bar{y}) + \mathrm{icr}(C)$ is convex, then statement (i) *can be replaced by the following:*

(i)′ $\bar{y} \in \mathrm{PVMin}[A, C]$.

Proof. (i)⇒(ii): Since $\mathrm{icr}[\mathrm{cone}(A-\bar{y})+\mathrm{icr}(C)] \neq \emptyset$, we know that $\mathrm{icr}[\mathrm{cone}(\mathrm{co}A - \bar{y}) + \mathrm{icr}(C)] \neq \emptyset$. Clearly, $\mathrm{cone}(\mathrm{co}A - \bar{y}) + \mathrm{icr}(C)$ is convex. By Lemma 3.1 we have

$$\mathrm{vcl}[\mathrm{cone}(\mathrm{co}A - \bar{y}) + \mathrm{icr}(C)] = \mathrm{cl}_\gamma[\mathrm{cone}(\mathrm{co}A - \bar{y}) + \mathrm{icr}(C)] = \mathrm{cl}_\gamma[\mathrm{cone}(\mathrm{co}A - \bar{y}) + C].$$

On the other hand, $\mathrm{cor}(C^+) \neq \emptyset$ implies that $\mathrm{icr}(C) \neq \emptyset$. Furthermore, C is convex and $C = \mathrm{vcl}(C)$. By Lemma 3.1 we have $\mathrm{cl}_\gamma(C) = \mathrm{vcl}(C) = C$ and C is γ-closed. Thus statement (i) becomes

$$\mathrm{cl}_\gamma[\mathrm{cone}(\mathrm{co}A - \bar{y} + C)] \cap (-C) = \{0\},$$

where C is a closed convex pointed cone in (Y, γ). That is, $\bar{y} \in A$ is a Benson proper efficient point of $\mathrm{co}A$ with respect to C in (Y, γ). By Lemma 3.2, we have $\mathrm{int}_\sigma(C^+) = \mathrm{cor}(C^+) \neq \emptyset$. Observing that the topological dual of $(Y^\#, \sigma(Y^\#, Y))$ is Y, we can apply Theorem 3.2 and conclude that there exists $f \in \mathrm{cor}(C^+)$ such that $f(A) \geq f(\bar{y})$.

(ii)⇒ (iii): It is obvious.

(iii)⇒ (i): If not, there exists $c_0 \in C,\ c_0 \neq 0$, such that

$$-c_0 \in \mathrm{vcl}[\mathrm{cone}(\mathrm{co}A - \bar{y} + C)].$$

Since $f \in C^{+i}$, we have $f(c_0) > 0$, and so $f(-c_0) < 0$. On the other hand, $f \in C^+$ and $f(A) \geq f(\bar{y})$ imply that

$$f(\mathrm{vcl}[\mathrm{cone}(\mathrm{co}A - \bar{y} + C)]) \geq 0,$$

so $f(-c_0) \geq 0$, a contradiction!

Moreover, if $\mathrm{cone}(A - \bar{y}) + \mathrm{icr}(C)$ is convex, then

$$\mathrm{cone}(A - \bar{y}) + \mathrm{cir}(C) = \mathrm{cone}(\mathrm{co}A - \bar{y}) + \mathrm{icr}(C).$$

Hence we have $\mathrm{PVMin}[\mathrm{co}A, C] = \mathrm{PVMin}[A, C]$, and statement (i) can be replaced by (i)′. □

By Corollary 3.7, we can deduce the scalarization results in [8]. As in [8], let X be a nonempty set, Y be a real linear space, $C \subset Y$ be an ordering cone, and $F : X \to 2^Y$ be a set-valued map. Consider the following unconstrained optimization problem:

(P) $C - \mathrm{Min}F(x)$, subject to $x \in X$.

A pair $(\bar{x}, \bar{y}) \in X \times Y$ is said to be a Benson-vectorial proper minimizer of (P) if $\bar{y} \in F(\bar{x})$ and $\bar{y} \in \mathrm{PVMin}[F(X), C]$, i.e.,

$$\mathrm{vcl}[\mathrm{cone}(F(X) - \bar{y} + C)] \cap (-C) = \{0\}.$$

If $\operatorname{icr}(C) \neq \emptyset$, then a set-valued map $F : X \to 2^Y$ is said to be relatively solid C-subconvexlike on X (see [8]) if the following conditions are satisfied:

(i) $F(X) + \operatorname{icr}(C)$ is convex,

(ii) $\operatorname{icr}[F(X) + \operatorname{icr}(C)] \neq \emptyset$.

Substituting A in Corollary 3.7 by $F(X)$, we immediately obtain the following scalarization result (see [8, Theorems 4 and 5 and Corollary 2]).

COROLLARY 3.8. *Let C be a vectorially closed convex cone and $\operatorname{cor}(C^+) \neq \emptyset$. Let F be a relatively solid C-subconvexlike on X. Then $(\bar{x}, \bar{y})$ is a Benson-vectorial proper minimizer of* (P) *if and only if there exists $f \in C^{+i}$ such that $f(F(X)) \geq f(\bar{y})$.*

4. Benson proper minimizers in vector optimization problems with set-valued maps. In the following, we assume that X is a nonempty set, Y and Z are l.c.s., and $C \subset Y$ and $D \subset Z$ are closed convex cones. Let $F : X \to 2^Y$ and $G : X \to 2^Z$ be set-valued maps with nonempty values.

Consider the following constrained vector optimization problem with set-valued maps:

$$\text{(VP)} \quad C - \operatorname{Min} F(x) \quad \text{s.t.} \quad G(x) \cap (-D) \neq \emptyset, \quad x \in X.$$

Denote the *feasible set* of (VP) by

$$\Gamma = \{x \in X : G(x) \cap (-D) \neq \emptyset\}$$

and the image of Γ under F by

$$F(\Gamma) = \bigcup_{x \in \Gamma} F(x).$$

Li [13] introduced the concept of *Benson proper efficiency* for set-valued vector optimization problems as follows.

DEFINITION 4.1. *A point $\bar{x} \in \Gamma$ is called a* Benson properly efficient solution *of* (VP) *if*

$$F(\bar{x}) \cap \operatorname{PMin}[F(\Gamma), C] \neq \emptyset.$$

A point $(\bar{x}, \bar{y})$ is called a Benson proper minimizer *of* (VP) *if*

$$\bar{y} \in F(\bar{x}) \cap \operatorname{PMin}[F(\Gamma), C].$$

In recent years, much attention has been paid to generalized convexity of set-valued maps. Let us recall some notions (see [24]).

A set-valued map F is said to be *C-convexlike* on Γ if $F(\Gamma) + C$ is convex, to be *C-subconvexlike* on Γ if $F(\Gamma) + \operatorname{int} C$ is convex, and to be *nearly C-convexlike* on Γ if $\operatorname{cl}(F(\Gamma) + C)$ is convex. Recently, a new generalized convexity for set-valued maps, called *near C-subconvexlikeness*, was introduced in [24]. A set-valued map F is said to be *nearly C-subconvexlike* on Γ if $\operatorname{cl}\operatorname{cone}(F(\Gamma) + C)$ is convex. We know that C-convexlikeness $\Rightarrow$ C-subconvexlikeness (when $\operatorname{int} C \neq \emptyset$) $\Rightarrow$ near C-convexlikeness $\Rightarrow$ near C-subconvexlikeness, and none of the converses is true (see [24]). Hence near C-subconvexlikeness is the weakest convexity among the above four kinds of generalized convexity. Also, we observe that the former three kinds of generalized convexity are invariant by translations. Using the results in section 3, we can easily obtain the following scalarization theorems for Benson proper efficiency in optimization problems with nearly C-subconvexlike objectives.

THEOREM 4.1. *Let $\bar{x} \in \Gamma$, $\bar{y} \in F(\bar{x})$, and $F - \bar{y}$ be nearly C-subconvexlike on Γ. The point $(\bar{x}, \bar{y})$ is a Benson proper minimizer of* (VP) *if and only if, for any weakly compact convex set $K \subset C$ and $0 \notin K$, there exists $f \in C^+ \cap K^{st}$ such that $f(F(\Gamma)) \geq f(\bar{y})$.*

Proof. Clearly, $F - \bar{y}$ being nearly C-subconvexlike on Γ means that $\operatorname{cl}\operatorname{cone}(F(\Gamma) - \bar{y} + C)$ is convex. By taking $A = F(\Gamma)$ in Theorem 3.1 (and in Remark 3.3) we immediately obtain the result. □

Similarly, from Theorem 3.2 and Corollaries 3.1, 3.2, 3.3, 3.4, 3.5, and 3.6 we can obtain the following scalarization results. In the following, we always assume that $\bar{x} \in \Gamma$, $\bar{y} \in F(\bar{x})$, and $F - \bar{y}$ is nearly C-subconvexlike on Γ.

THEOREM 4.2. *Let the ordering cone $C \subset Y$ satisfy $\operatorname{int}_{\mathcal{T}} C^+ \neq \emptyset$, where $\mathcal{T}$ is any locally convex topology on Y^* compatible with the dual pair $\langle Y, Y^* \rangle$. Then $(\bar{x}, \bar{y})$ is a Benson proper minimizer of* (VP) *if and only if there exists $f \in C^{+i}$ (or $f \in \operatorname{int}_{\mathcal{T}} C^+$) such that $f(F(\Gamma)) \geq f(\bar{y})$.*

COROLLARY 4.1. *Let the ordering cone $C \subset Y$ have a weakly compact base B. Then $(\bar{x}, \bar{y})$ is a Benson proper minimizer of* (VP) *if and only if there exists $f \in C^{+i}$ (or $f \in B^{st}$) such that $f(F(\Gamma)) \geq f(\bar{y})$.*

Remark 4.1. Under the assumption that the ordering cone C has a compact base (equivalently, C is locally compact), Li [13, Theorem 4.2] and Cheng and Rong [2, Theorem 4.1], respectively, established the scalarization theorems on the Benson proper minimizer of (VP) with C-subconvexlike set-valued maps and with generalized C-subconvexlike vector-valued maps. Under the assumption that the ordering cone C is locally compact and under the assumption that $F - \bar{y}$ is nearly C-subconvexlike, Yang, Li, and Wang [24, Theorem 6.2] gave the scalarization theorem on the Benson proper minimizer of (VP). Since C-subconvexlikeness for set-valued maps is invariant by translations and it implies near C-subconvexlikeness, and since a vector-valued map may be regarded as a special set-valued map, Theorems 4.1 and 4.2 and Corollary 4.1 generalize and improve [13, Theorem 4.2] and [2, Theorem 4.1]. Moreover, it is easy to see that our results also generalize and improve [24, Theorem 6.2] (please refer to Example 3.1).

COROLLARY 4.2 (see [5, Theorem 3.2]). *Let $(Y, \|\ \|)$ be a separable normed space and $C \subset Y$ be a closed convex pointed cone. Then $(\bar{x}, \bar{y})$ is a Benson proper minimizer of* (VP) *if and only if there exists $f \in C^{+i}$ such that $f(F(\Gamma)) \geq f(\bar{y})$.*

COROLLARY 4.3. *Let Y be an l.c.s. with the w.c.n.p. and $C \subset Y$ be a closed convex cone with a countable weakly compact base B. Then $(\bar{x}, \bar{y})$ is a Benson proper minimizer of* (VP) *if and only if there exists $f \in C^{+i}$ such that $f(F(\Gamma)) \geq f(\bar{y})$.*

COROLLARY 4.4. *Let Y be the dual of a Fréchet space (E, d) and Y have a locally convex topology compatible with the dual pair $\langle E, Y \rangle$. Let $C \subset Y$ be a closed convex cone with a base. Then $(\bar{x}, \bar{y})$ is a Benson proper minimizer of* (VP) *if and only if there exists $f \in C^{+i}$ such that $f(F(\Gamma)) \geq f(\bar{y})$.*

COROLLARY 4.5. *Let Y be a semireflexive (DF)-space (particularly, a reflexive Banach space) and $C \subset Y$ be a closed convex cone with a base. Then $(\bar{x}, \bar{y})$ is a Benson proper minimizer of* (VP) *if and only if there exists $f \in C^{+i}$ such that $f(F(\Gamma)) \geq f(\bar{y})$.*

5. Lagrange multipliers. In this section, under the weaker assumption we present two Lagrange multiplier theorems which show that a Benson proper minimizer of the constrained set-valued vector optimization problem (VP) is exactly a Benson proper minimizer for an appropriate unconstrained set-valued vector optimization problem. Let $L(Z, Y)$ (respectively, $\mathcal{L}(Z, Y)$) be the set of all linear maps (respectively,

continuous linear maps) from Z to Y. Also, $T \in L(Z,Y)$ (respectively, $T \in \mathcal{L}(Z,Y)$) is said to be nonnegative with respect to the cones C and D, denoted by $T \in L_+(Z,Y)$ (respectively, $T \in \mathcal{L}_+(Z,Y)$), if $T(C) \subset D$. We say that (VP) satisfies the *generalized Slater constraint qualification* if there exists $x' \in X$ such that $G(x') \cap (-\mathrm{int}D) \neq \emptyset$ (see, e.g., [13, Definition 2.2]). Here we introduce its extension as follows.

DEFINITION 5.1. *We say that* (VP) *satisfies the conelike generalized Slater constraint qualification if*

$$\mathrm{cl}[\Lambda(G(X))] \cap (-\mathrm{int}D) \neq \emptyset,$$

where $\Lambda(G(X))$ *denotes the set* $\{\sum_{i=1}^{n} \lambda_i z_i : \ n \in N,\ \lambda_i \geq 0,\ z_i \in G(X),\ i = 1,2,\ldots,n\}$.

Remark 5.1. Since

$$\mathrm{cl}[\Lambda(G(X))] \cap (-\mathrm{int}D) \ \subset \ \mathrm{cl}[\Lambda(G(X)) \cap (-\mathrm{int}D)],$$

the condition that $\mathrm{cl}[\Lambda(G(X))] \cap (-\mathrm{int}D) \neq \emptyset$ in Definition 5.1 is equivalent to $\Lambda(G(X)) \cap (-\mathrm{int}D) \neq \emptyset$.

We easily see that the above condition is strictly weaker than the generalized Slater constraint qualification.

Example 5.1. Let $X = \{(x_1,x_2) \in R^2 : \ x_1 \leq 0,\ x_2 \leq 0,\ 1 \leq x_1^2 + x_2^2 \leq 4\}$, $Z = R^2$, and $D \subset Z$ be the cone

$$\left\{(x_1,x_2) \in R^2 : \ x_1 \geq 0,\ x_2 \geq 0,\ \frac{\sqrt{3}}{3}x_1 \leq x_2 \leq \sqrt{3}x_1\right\}.$$

Define a set-valued map $G : X \to 2^Z$ as follows:

$$(x_1,x_2) \to G((x_1,x_2))$$
$$= \begin{cases} \{(\rho x_1, 0) : \rho \geq 0\} & \text{if } (x_1,x_2) \in X \cap \{(x_1,x_2) \in R^2 : \sqrt{3}x_1 < x_2 < \frac{\sqrt{3}}{3}x_1\}; \\ \{(x_1,x_2)\} & \text{if } (x_1,x_2) \in X \backslash \{(x_1,x_2) \in R^2 : \sqrt{3}x_1 < x_2 < \frac{\sqrt{3}}{3}x_1\}. \end{cases}$$

Obviously, for any $(x_1,x_2) \in X$,

$$G((x_1,x_2)) \cap (-\mathrm{int}D) = \emptyset.$$

But

$$-\Lambda(G(X)) \cap (-\mathrm{int}D) \neq \emptyset.$$

This shows that the set-valued map G does not satisfy the generalized Slater constraint qualification, but it satisfies the conelike generalized Slater constraint qualification.

The following Theorems 5.1, 5.2, and 5.3 improve and extend [13, Theorems 5.1 and 5.2], [2, Theorem 5.1], and [23, Theorem 3.1].

THEOREM 5.1. *Let* Y *and* Z *be l.c.s.,* $C \subset Y$ *be a closed convex pointed cone,* $D \subset Z$ *be a convex cone with nonempty interior,* $\bar{x} \in \Gamma$, *and* $\bar{y} \in F(\bar{x})$. *Furthermore, let* $(F - \bar{y}, G)$ *be nearly* $C \times D$*-subconvexlike on* X *and* (VP) *satisfy the conelike generalized Slater constraint qualification. If there exists* $f \in C^{+i}$ *such that* $f(F(\Gamma)) \geq f(\bar{y})$, *then there exists* $T \in \mathcal{L}_+(Z,Y)$ *such that*

$$T(G(\bar{x}) \cap (-D)) = \{0\} \ \text{ and } \ \bar{y} \in \mathrm{PMin}[(F+TG)(X), C].$$

Proof. Set $B := (f = 1) \cap C$. Then B is a base of C. By the hypothesis, $f \in C^{+i}$ and $f(F(\Gamma) - \bar{y}) \geq 0$, so $f(\mathrm{cl}\,\mathrm{cone}(F(\Gamma) - \bar{y} + C)) \geq 0$. Also, clearly $f(-B) = \{-1\}$. Let V be an absolutely convex 0-neighborhood in Y, with $V \subset \{y \in Y : |f(y)| < \frac{1}{4}\}$. Then

$$(\mathrm{cl}\,\mathrm{cone}(F(\Gamma) - \bar{y} + C) + V) \cap -(B + V) = \emptyset. \tag{5}$$

From (5), $V \cap (-B - V) = \emptyset$, and hence $0 \notin \mathrm{cl}(B + V)$. Thus $B + V$ is a base of $\mathrm{cone}(B+V)$, and clearly $\mathrm{cone}(B+V)$, and has nonempty interior. Denote $\mathrm{cone}(B+V)$ by $C_V(B)$. We assert that

$$\mathrm{cone}((F - \bar{y}, G)(X)) \cap (-\mathrm{int}C_V(B), -\mathrm{int}D) = \emptyset. \tag{6}$$

First we observe that any element of $\mathrm{cone}((F - \bar{y}, G)(X))$ may be written in the following form:

$$\alpha(y - \bar{y}, z) \in Y \times Z, \quad \text{where } \alpha \geq 0,\ y \in F(x),\ z \in G(x),\ x \in X.$$

We show (6) according to the following two cases.

Case 1. If $x \in \Gamma$, then $\alpha(y - \bar{y}) \in \mathrm{cone}(F(\Gamma) - \bar{y})$. Combining this with (5), we can deduce that

$$\alpha(y - \bar{y}) \notin -\mathrm{int}C_V(B).$$

Hence

$$\alpha(y - \bar{y}, z) \notin (-\mathrm{int}C_V(B), -\mathrm{int}D).$$

Case 2. If $x \notin \Gamma$, then $G(x) \cap (-D) = \emptyset$. Since $z \in G(x)$, we have $z \notin -D$ and $\alpha z \notin -\mathrm{int}D$. Hence we also have

$$\alpha(y - \bar{y}, z) \notin (-\mathrm{int}C_V(B), -\mathrm{int}D).$$

Thus we have shown (6). From this, we have

$$(\mathrm{cone}((F - \bar{y}, G)(X)) + (C, D)) \cap (-\mathrm{int}C_V(B), -\mathrm{int}D) = \emptyset.$$

Noting that $(-\mathrm{int}C_V(B), -\mathrm{int}D)$ is open in $Y \times Z$, we conclude that

$$\mathrm{cl}[\mathrm{cone}((F - \bar{y}, G)(X)) + (C, D)] \cap (-\mathrm{int}C_V(B), -\mathrm{int}D) = \emptyset. \tag{7}$$

By the assumption that $(F - \bar{y}, G)$ is nearly $C \times D$-subconvexlike on X, we know that $\mathrm{cl}[\mathrm{cone}((F - \bar{y}, G)(X)) + (C, D)]$ is a closed convex set, which does not intersect the open convex set $(-\mathrm{int}C_V(B), -\mathrm{int}D)$ in $Y \times Z$. By the Hahn–Banach separation theorem, there exists $(\varphi, \psi) \in (Y \times Z)^* = Y^* \times Z^*$, $(\varphi, \psi) \neq (0, 0)$, such that

$$\begin{aligned} (\varphi, \psi)(\mathrm{cone}((F - \bar{y}, G)(X)) + (C, D)) \geq 0 &> (\varphi, \psi)(-\mathrm{int}C_V(B), -\mathrm{int}D) \\ &= \varphi(-\mathrm{int}C_V(B)) + \psi(-\mathrm{int}D). \end{aligned}$$

From this, we have

$$\varphi(F(x) - \bar{y}) + \psi(G(x)) \geq 0 \quad \forall x \in X, \tag{8}$$

$$\varphi \in (C_V(B))^+, \tag{9}$$

and

$$\psi \in D^+. \tag{10}$$

Since $\bar{x} \in \Gamma$, we have $G(\bar{x}) \cap (-D) \neq \emptyset$. Take any $\bar{z} \in G(\bar{x}) \cap (-D)$. By (10) we have

$$\psi(\bar{z}) \leq 0. \tag{11}$$

By taking $x = \bar{x}$ in (8), we obtain

$$\varphi(F(\bar{x}) - \bar{y}) + \psi(G(\bar{x})) \geq 0.$$

Since $\bar{y} \in F(\bar{x})$ and $\bar{z} \in G(\bar{x})$, we have

$$\psi(\bar{z}) \geq 0. \tag{12}$$

Combining (11) and (12), we have $\psi(\bar{z}) = 0$, and hence

$$\psi(G(\bar{x}) \cap (-D)) = \{0\}. \tag{13}$$

Next we show that $\varphi \neq 0$. If not, we assume that $\varphi = 0$. Then $\psi \in D^+ \backslash \{0\}$. From (8) and $\varphi = 0$, we have

$$\psi(G(X)) \geq 0. \tag{14}$$

On the other hand, (VP) satisfies the conelike generalized Slater constraint qualification, so

$$\Lambda(G(X)) \cap (-\mathrm{int}D) \neq \emptyset.$$

That is, there exist $n \in N$, $\lambda_i \geq 0$, and $z_i \in G(X)$ for $i = 1, 2, \ldots, n$ such that $\sum_{i=1}^{n} \lambda_i z_i \in -\mathrm{int}D$. Since $\psi \in D^+ \backslash \{0\}$, we have

$$\sum_{i=1}^{n} \lambda_i \psi(z_i) = \psi\left(\sum_{i=1}^{n} \lambda_i z_i\right) < 0.$$

Thus there exists some i, $1 \leq i \leq n$, such that $\psi(z_i) < 0$, where $z_i \in G(X)$. This contradicts (14). Therefore, $\varphi \neq 0$ and $\varphi \in (C_V(B))^+ \backslash \{0\}$. From this, we know that $\varphi(B + V) \geq 0$, and hence there exists $\delta > 0$ such that $\varphi(B) \geq \delta > 0$, so $\varphi \in B^{st}$. Obviously, there exists $\bar{c} \in C = \mathrm{cone}(B)$ such that $\varphi(\bar{c}) = 1$. Define $T \in \mathcal{L}(Z, Y)$ as follows:

$$T(z) = \psi(z)\,\bar{c} \quad \forall z \in Z. \tag{15}$$

Since $\psi(D) \geq 0$ and $\bar{c} \in C$, we have

$$T(D) \subset C, \quad \text{i.e.,} \quad T \in \mathcal{L}_+(Z, Y).$$

By (13) and the definition of T, we have

$$T(G(\bar{x}) \cap (-D)) = \psi(G(\bar{x}) \cap (-D))\,\bar{c} = \{0\}.$$

For any $x \in X$, any $y \in F(x)$, and any $z \in G(x)$, we have

$$\begin{aligned} \varphi(y - \bar{y} + Tz) &= \varphi(y - \bar{y}) + \varphi(Tz) \\ &= \varphi(y - \bar{y}) + \varphi(\psi(z)\bar{c}) \\ &= \varphi(y - \bar{y}) + \psi(z) \geq 0, \end{aligned} \tag{16}$$

where the final inequality is due to (8). By (16) and $\varphi \in B^{st} \subset C^{+i}$, we have

$$\varphi(\mathrm{cl\,cone}((F+TG)(X)-\bar{y}+C)) \geq 0. \tag{17}$$

On the other hand,

$$\varphi(-C\backslash\{0\}) < 0. \tag{18}$$

Combining (17) and (18), we conclude that

$$\mathrm{cl\,cone}((F+TG)(X)-\bar{y}+C)\cap(-C) = \{0\}.$$

That is,

$$\bar{y} \in \mathrm{PMin}[(F+TG)(X), C].$$

Thus we complete the proof. □

THEOREM 5.2. *Let $F: X \to 2^Y$ and $G: X \to 2^Z$ be set-valued maps and $C \subset Y$ and $D \subset Z$ be convex cones. Let $\Gamma = \{x \in X : G(x)\cap(-D) \neq \emptyset\}$, $\bar{x} \in \Gamma$, and $\bar{y} \in F(\bar{x})$. If there exists $T \in L_+(Z,Y)$ such that $\bar{y} \in \mathrm{PMin}[(F+TG)(X), C]$, then $\bar{y} \in \mathrm{PMin}[F(\Gamma), C]$.*

Proof. For any $x \in \Gamma$, there exists $z_x \in G(x)\cap(-D)$. Since $T(-D) \subset -C$, we have $Tz_x \in -C$. Thus

$$\{0\} \subset Tz_x + C \subset TG(x) + C.$$

From this,

$$F(x)-\bar{y} \subset F(x)-\bar{y}+TG(x)+C \quad \forall x \in \Gamma.$$

That is,

$$\begin{aligned} F(\Gamma)-\bar{y} &\subset (F+TG)(\Gamma)+C-\bar{y} \\ &\subset (F+TG)(X)+C-\bar{y}. \end{aligned}$$

Since C is a convex cone, it is easy to see that

$$F(\Gamma)-\bar{y}+C \subset (F+TG)(X)+C-\bar{y}$$

and certainly

$$\mathrm{cl\,cone}(F(\Gamma)-\bar{y}+C) \subset \mathrm{cl\,cone}((F+TG)(X)+C-\bar{y}). \tag{19}$$

By the assumption that $\bar{y} \in \mathrm{PMin}[(F+TG)(X), C]$, we have

$$\mathrm{cl\,cone}((F+TG)(X)+C-\bar{y})\cap(-C) = \{0\}.$$

Combining this with (19), we conclude that

$$\mathrm{cl\,cone}(F(\Gamma)-\bar{y}+C)\cap(-C) = \{0\}, \quad \text{that is,} \quad \bar{y} \in \mathrm{PMin}[F(\Gamma), C]. \qquad \square$$

Here it is not necessary that T be continuous. Also, we need not assume that $0 \in TG(\bar{x})$ in advance. Please compare this result with [13, Theorem 5.2]. Combining Theorems 5.1 and 5.2 and the results in section 4, we obtain the following.

THEOREM 5.3. *Let $\bar{x} \in \Gamma$, $\bar{y} \in F(\bar{x})$, $F - \bar{y}$ be nearly C-subconvexlike on Γ, $(F - \bar{y}, G)$ be nearly $C \times D$-subconvexlike on X, and* (VP) *satisfy the conelike generalized Slater constraint qualification. Assume that one of the following conditions is satisfied:*

(i) *the ordering cone $C \subset Y$ has a weakly compact base (or a compact base);*

(ii) *Y is a separable normed space;*

(iii) *Y is an l.c.s. with the w.c.n.p., and C has a countable weakly compact base;*

(iv) *Y is the dual of a Fréchet space E, endowed with a locally convex topology compatible with the dual pair $\langle E, Y\rangle$, and the ordering cone C has a base;*

(v) *Y is a semireflexive (DF)-space (particularly, Y is a reflexive Banach space), and the ordering cone C has a base.*

Then $(\bar{x}, \bar{y})$ is a Benson proper minimizer of (VP) *if and only if there exists $T \in \mathcal{L}_+(Z, Y)$ such that*

$$T(G(\bar{x}) \cap (-D)) = \{0\} \quad \text{and} \quad \bar{y} \in \mathrm{PMin}[(F + TG)(X), C].$$

Acknowledgments. The author thanks the anonymous referees for many helpful suggestions and for [3, 6, 7, 8].

REFERENCES

[1] H. P. BENSON, *An improved definition of proper efficiency for vector maximization with respect to cones*, J. Math. Anal. Appl., 71 (1979), pp. 232–241.

[2] G. Y. CHEN AND W. D. RONG, *Characterizations of the Benson proper efficiency for nonconvex vector optimization*, J. Optim. Theory Appl., 98 (1998), pp. 365–384.

[3] R. N. GASIMOV, *Characterization of the Benson proper efficiency and scalarization in nonconvex vector optimization*, in Multiple Criteria Decision Making in the New Millennium, Ankara, 2000, Lecture Notes in Econom. and Math. Systems 507, Springer-Verlag, Berlin, 2001, pp. 189–198.

[4] A. M. GEOFFRION, *Proper efficiency and the theory of vector maximization*, J. Math. Anal. Appl., 22 (1968), pp. 616–630.

[5] A. GUERRAGGIO, E. MOLHO, AND A. ZAFFARONT, *On the notion of proper efficiency in vector optimization*, J. Optim. Theory Appl., 82 (1994), pp. 1–21.

[6] Z. Q. HAN, *Remarks on angle property and solid cones*, J. Optim. Theory Appl., 82 (1994), pp. 149–157.

[7] Z. Q. HAN, *Relationship between solid cones and cones with bases*, J. Optim. Theory Appl., 90 (1996), pp. 457–463.

[8] E. HERNÁNDEZ, B. JIMÉNEZ, AND V. NOVO, *Benson proper efficiency in set-valued optimization on real linear spaces*, in Recent Advances in Optimization, Lecture Notes in Econom. and Math. Systems 563, Springer-Verlag, Berlin, 2006, pp. 45–59.

[9] J. HORVÁTH, *Topological Vector Spaces and Distributions*, Vol. 1, Addison-Wesley, Reading, MA, 1966.

[10] G. JAMESON, *Ordered Linear Spaces*, Springer-Verlag, Berlin, 1970.

[11] G. KÖTHE, *Topological Vector Spaces* I, Springer-Verlag, Berlin, 1983.

[12] H. W. KUHN AND A. W. TUCKER, *Nonlinear programming*, in Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, CA, 1951, pp. 481–492.

[13] Z. F. LI, *Benson proper efficiency in the vector optimization of set-valued maps*, J. Optim. Theory Appl., 98 (1998), pp. 623–649.

[14] R. MEISE AND D. VOGT, *Introduction to Functional Analysis*, Oxford University Press, New York, 1997.

[15] P. PEREZ CARRERAS AND J. BONET, *Barrelled Locally Convex Spaces*, North–Holland, Amsterdam, 1987.

[16] J. H. QIU, *On solidness of polar cones*, J. Optim. Theory Appl., 109 (2001), pp. 199–214.

[17] J. H. QIU AND K. MCKENNON, *Strictly extreme and strictly exposed points*, Int. J. Math. Math. Sci., 17 (1994), pp. 451–456.

[18] A. P. ROBERTSON AND W. J. ROBERTSON, *Topological Vector Spaces*, Cambridge University Press, Cambridge, UK, 1964.

[19] P. H. Sach, *Nearly subconvexlike set-valued maps and vector optimization problems*, J. Optim. Theory Appl., 119 (2003), pp. 335–356.
[20] H. H. Schaefer, *Topological Vector Spaces*, Springer-Verlag, New York, 1971.
[21] W. Song, *Duality for vector optimization of set-valued functions*, J. Math. Anal. Appl., 201 (1996), pp. 212–225.
[22] W. Song, *Lagrangian duality for minimization of nonconvex multifunctions*, J. Optim. Theory Appl., 93 (1997), pp. 167–182.
[23] Y. Xu and S. Liu, *Benson proper efficiency in the nearly cone-subconvexlike vector optimization with set-valued functions*, Appl. Math. J. Chinese Univ. Ser. B, 18 (2003), pp. 95–102.
[24] X. M. Yang, D. Li, and S. Y. Wang, *Near-subconvexlikeness in vector optimization with set-valued functions*, J. Optim. Theory Appl., 110 (2001), pp. 413–427.

SIAM J. OPTIM.
Vol. 19, No. 1, pp. 163–187

CONSTRAINT QUALIFICATIONS FOR CONVEX INEQUALITY SYSTEMS WITH APPLICATIONS IN CONSTRAINED OPTIMIZATION*

CHONG LI†, K. F. NG‡, AND T. K. PONG‡

Abstract. For an inequality system defined by an infinite family of proper convex functions, we introduce some new notions of constraint qualifications in terms of the epigraphs of the conjugates of these functions and study relationships between these new constraint qualifications and other well-known constraint qualifications including the basic constraint qualification studied by Hiriart-Urrutty and Lemarechal and by Li, Nahak, and Singer. Extensions of known results to more general settings are presented, and applications to particular important problems, such as conic programming and approximation theory, are also studied.

Key words. convex inequality system, basic constraint qualification, strong conical hull intersection property, best constrained approximation, conic programming

AMS subject classifications. Primary, 90C34, 90C25; Secondary, 52A07, 41A29, 90C46

DOI. 10.1137/060676982

1. Introduction.

Many problems in optimization and approximation theory can be recast into one of the following two types: one is a system of convex inequalities

$$g_i(x) \le 0 \quad \text{for each } i \in I, \tag{1.1}$$

and the other is a minimization problem

$$\begin{aligned} &\text{Minimize} && f(x), \\ &\text{s.t.} && x \in C,\ g_i(x) \le 0,\ i \in I, \end{aligned} \tag{1.2}$$

where C is a convex set, not necessarily closed. Many authors have studied these two problems with various degrees of generality imposed on the index set I, the family of functions $\{g_i : i \in I\}$, or on the underlying space; see, for example, [4, 5, 6, 13, 14, 15, 16, 17, 18, 19, 21, 23, 29, 30, 31, 32, 33, 34, 35, 36] and references therein.

A special case of (1.1) occurs when each g_i is the indicator function of a closed convex set C_i; that is, one considers a family of closed convex sets $\{C_i : i \in I\}$. In [13], Deutsch, Li, and Ward introduced the notion of the strong conical hull intersection property (the strong CHIP) for a family of finitely many closed convex sets in a Hilbert space in connection with the reformulation of some best approximation problems. Their work was recently extended in [32, 34] to the setting of a normed linear space with I being an infinite set.

For the case when $\{g_i : i \in I\}$ is a finite family of continuous convex functions on a finite dimensional vector space, the notion of basic constraint qualification (BCQ) was

*Received by the editors December 6, 2006; accepted for publication (in revised form) August 27, 2007; published electronically February 8, 2008.
http://www.siam.org/journals/siopt/19-1/67698.html

†Department of Mathematics, Zhejiang University, Hangzhou 310027, P. R. China (cli@zju.edu.cn). This author was supported in part by the National Natural Science Foundation of China (grants 10671175 and 10731060) and the Program for New Century Excellent Talents in University.

‡Department of Mathematics, Chinese University of Hong Kong, Hong Kong, P. R. China (kfng@math.cuhk.edu.hk, tkpong@gmail.com). The second author was supported by a direct grant (CUHK) and an Earmarked Grant from the Research Grant Council of Hong Kong.

introduced by Hiriart-Urrutty and Lemarechal (see [19]). The notion was extended to cover the case of an infinite family of continuous convex functions with a continuous sup-function $\sup_{i\in I} g_i$ by Li, Nahak, and Singer (see [36]), who also studied many aspects of the BCQ in relation to other constraint qualifications. Recall from [32,33,36] that the inequality system (1.1) is said to satisfy the BCQ at $x \in S := \{x \in X : g_i(x) \le 0,\ i \in I\}$ if

$$N_S(x) = \text{cone} \bigcup_{i\in I(x)} \partial\, g_i(x)$$

(see the next section for notation and definitions). The concept of a BCQ relative to C was introduced in [31, 32, 33] in order to take care of the abstract constraint set C. In these papers, under some continuity assumption such as the one used in [36], the system (1.1) with the family $\{\delta_C;\, g_i,\ i \in I\}$ was considered in place of $\{g_i :\ i \in I\}$.

Constraint qualifications involving epigraphs (first introduced in [8,9]) have been extensively used by many authors (see, for example, [4, 5, 6, 7, 8, 9, 10, 15, 16, 24, 25, 26, 27, 28, 29, 35]). Particularly in connection with the study of a conic programming problem (see Example 2.1 below), Jeyakumar and coworkers [25, 26, 29] and Boţ and Wanka [7] studied several new constraint qualifications (such as what they called the condition (**C***) and the CCCQ; see [26, 29] for their definitions). Inspired by these works as well as that of Dinh, Goberna, and López in [15] (especially with regard to the new optimality conditions for (1.2)), we define the following concept: the inequality system (1.1) is said to have the conical epigraph hull property (conical EHP) if

$$\text{epi}\, \sigma_S = \text{cone} \bigcup_{i\in I} \text{epi}\, g_i^*, \tag{1.3}$$

where $S = \{x : g_i(x) \le 0\ \ \forall i \in I\}$. In particular, (1.3) reduces to the sum of epigraph constraint qualification (SECQ) introduced in [35] if $g_i = \delta_{C_i}$ for some family of closed convex sets $\{C_i :\ i \in I\}$. We show that by suitably choosing the family $\{g_i : i \in I\}$, the conical EHP reduces to the closed cone constraint qualification (CCCQ) defined in [7, 26]. In section 4, we derive some relationships between the EHP, the BCQ, and the Pshenichyni–Levin–Valadier property (PLV property). We also give some applications involving the strong CHIP and the convex Farkas–Minkowski systems (studied by Li, Nahak, and Singer in [36]).

In this paper, we consider (1.2) under minimal assumptions: f is a proper convex lower semicontinuous function and $\{g_i : i \in I\}$ is a family of proper convex functions (not necessarily lower semicontinuous) defined on a locally convex Hausdorff topological vector space X with proper sup-function, where I is an arbitrary index set. The last three sections of this paper are on applications of results obtained in section 4. An optimality condition (of Lagrange type) for (1.2) is established in section 5, and as a consequence we provide an improved version of [16, Theorem 3] on a characterization of minimizers for the problem (1.2); our argument differs from [16] and allows us to treat the case when each g_i is not necessarily lower semicontinuous. In particular, our results here cover the interesting conic programming case in which the feasible solution set is not necessarily closed (as the involved functions are not necessarily lower semicontinuous). Several known results in the conic programming problem (see [24,26,29]) are extended/improved in section 6. Finally, we study a best approximation problem in section 7.

2. Notation and preliminary results. The notation used in the present paper is standard (cf. [11, 19, 40]). In particular, we assume throughout the whole paper

(unless otherwise specified) that X is a real locally convex Hausdorff topological vector space, and we let X^* denote the dual space of X, whereas $\langle x^*, x\rangle$ denotes the value of a functional x^* in X^* at $x \in X$, i.e., $\langle x^*, x\rangle = x^*(x)$. Let A be a set in X. The interior (resp., closure, convex hull, convex cone hull, linear hull, affine hull, boundary) of A is denoted by $\operatorname{int} A$ (resp., $\overline{A}$, $\operatorname{co} A$, $\operatorname{cone} A$, $\operatorname{span} A$, $\operatorname{aff} A$, $\operatorname{bd} A$). The positive polar cone $A^{\oplus}$ and the negative polar cone $A^{\ominus}$ are defined respectively by

$$A^{\oplus} := \{x^* \in X^* \ : \ \langle x^*, z\rangle \geq 0 \ \forall z \in A\}$$

and

$$A^{\ominus} := \{x^* \in X^* \ : \ \langle x^*, z\rangle \leq 0 \ \forall z \in A\}.$$

The normal cone of A at $z_0 \in A$ is denoted by $N_A(z_0)$ and is defined by $N_A(z_0) = (A - z_0)^{\ominus}$. The indicator function δ_A and the support function σ_A of A are respectively defined by

$$\delta_A(x) := \begin{cases} 0, & x \in A, \\ \infty, & \text{otherwise}, \end{cases}$$

and

$$\sigma_A(x^*) := \sup_{x \in A} \langle x^*, x\rangle \quad \text{for each } x^* \in X^*.$$

Let f and g be proper functions respectively defined on X and X^*. Let f^*, g^* denote their conjugate functions, that is,

$$f^*(x^*) := \sup\{\langle x^*, x\rangle - f(x) : \ x \in X\} \quad \text{for each } x^* \in X^*,$$

$$g^*(x) := \sup\{\langle x^*, x\rangle - g(x^*) : \ x^* \in X^*\} \quad \text{for each } x \in X.$$

The epigraph of a function f on X is denoted by $\operatorname{epi} f$ and defined by

$$\operatorname{epi} f := \{(x, r) \in X \times \mathbb{R} : \ f(x) \leq r\}.$$

For a proper convex function f, the subdifferential of f at $x \in X$, denoted by $\partial f(x)$, is defined by

$$\partial f(x) := \{x^* \in X^* : \ \ f(x) + \langle x^*, y - x\rangle \leq f(y) \quad \text{for each } y \in X\}.$$

Moreover, the Young's equality holds (cf. [40, Theorem 2.4.2(iii)]):

$$f(x) + f^*(x^*) = \langle x^*, x\rangle \quad \text{if and only if } x^* \in \partial f(x). \tag{2.1}$$

In particular,

$$(x^*, \langle x^*, x\rangle - f(x)) \in \operatorname{epi} f^* \quad \text{for each } x^* \in \partial\, f(x). \tag{2.2}$$

We also define

$$\operatorname{im} \partial f := \{y^* \in X^* : \ y^* \in \partial f(x) \text{ for some } x \in X\}$$

and

$$\operatorname{dom} \partial f := \{x \in X : \ \partial f(x) \neq \emptyset\}.$$

For a convex subset A of X, the following statements are standard and easily verified:

$$\sigma_A = \delta_A^*, \quad N_A(x) = \partial \delta_A(x) \quad \text{for each } x \in A, \tag{2.3}$$

$$\sigma_A(x^*) = \langle x^*, x\rangle \Leftrightarrow x^* \in N_A(x) \Longleftrightarrow (x^*, \langle x^*, x\rangle) \in \operatorname{epi} \sigma_A \text{ for each } (x, x^*) \in A \times X^*. \tag{2.4}$$

Moreover, for each $(x^*, \alpha) \in X^* \times \mathbb{R}$,

$$(x^*, \alpha) \in \operatorname{epi} \sigma_A \Longleftrightarrow \langle x^*, x\rangle \le \alpha \quad \text{for each } x \in A. \tag{2.5}$$

Let $\{A_i : i \in J\}$ be a family of subsets of X containing the origin. The set $\sum_{i\in J} A_i$ is defined by

$$\sum_{i\in J} A_i = \begin{cases} \left\{\sum_{i\in J_0} a_i : a_i \in A_i, \emptyset \ne J_0 \subseteq J \text{ being finite}\right\} & \text{if } J \ne \emptyset, \\ \{0\} & \text{if } J = \emptyset. \end{cases}$$

In the remainder of this paper, let $\{g_i : i \in I\}$ denote a family of proper convex functions on X, where I is an index set. Let G denote the sup-function of $\{g_i : i \in I\}$, that is,

$$G(x) := \sup\{g_i(x) : i \in I\} \quad \text{for each } x \in X.$$

We always assume that the sup-function is proper. Let S denote the solution set of the inequality system (1.1) defined by $\{g_i : i \in I\}$, that is,

$$S := \{x : g_i(x) \le 0 \ \forall\, i \in I\} = \{x : G(x) \le 0\}.$$

For each $x \in X$, we define

$$I(x) = \{i \in I : g_i(x) = G(x) = 0\}$$

and

$$\tilde{I}(x) := \{i \in I : g_i(x) = G(x)\}.$$

The consideration of optimization problem (1.2) abounds in the literature. We end this section with one such example (which will be discussed in detail in section 6). Consider the following conic programming problem that has been studied in [6] and has also been studied in [2, 7, 24, 25, 26, 27, 28, 29] for the special case when X, Z are Banach spaces and $g : X \to Z$ is K-convex continuous.

Example 2.1. Suppose that X, Z are locally convex Hausdorff topological vector spaces, $C \subseteq X$ is a convex set, and $K \subseteq Z$ is a closed convex cone. Define an order on Z by saying that $y \le_K x$ if $y - x \in -K$. We attach a greatest element ∞ with respect to $\le_K$ and denote $Z^\bullet := Z \cup \{+\infty\}$. The following operations are defined on $Z^\bullet$: for any $z \in Z$, $z + \infty = \infty + z = \infty$ and $t\infty = \infty$ for all $t \ge 0$.

Consider the following conic programming problem:

$$\begin{array}{ll} \text{Minimize} & f(x), \\ \text{s.t.} & x \in C,\ g(x) \in -K, \end{array} \tag{2.6}$$

where $f : X \to \mathbb{R} \cup \{+\infty\}$ is a proper convex lower semicontinuous function and $g : X \to Z^\bullet$ is K-convex in the sense that for every $u, v \in X$ and every $t \in [0, 1]$,

$$g(tu + (1-t)v) \leq_K tg(u) + (1-t)g(v)$$

(see [2, 5, 6, 22, 25]). As in [5], we define for each $\lambda \in K^\oplus$

$$(\lambda g)(x) := \begin{cases} \langle \lambda, g(x) \rangle & \text{if } x \in \operatorname{dom} g, \\ +\infty & \text{otherwise,} \end{cases} \tag{2.7}$$

where $\operatorname{dom} g := \{x \in X : g(x) \in Z\}$. It is easy to see that g is K-convex if and only if $(\lambda g)(\cdot) : X \to \mathbb{R} \cup \{+\infty\}$ is a convex function for each $\lambda \in K^\oplus$. The problem (2.6) can be equivalently stated as

$$\begin{aligned} &\text{Minimize} && f(x), \\ &\text{s. t.} && \delta_C(x) \leq 0,\ (\lambda g)(x) \leq 0 \quad \text{for each } \lambda \in K^\oplus. \end{aligned}$$

Thus (2.6) can be viewed as an example of (1.2) by letting $I = K^\oplus \cup \{i_0\}$ with $i_0 \notin I$ and

$$g_{i_0} = \delta_C,\ g_\lambda = \lambda g \quad \text{for each } \lambda \in K^\oplus. \tag{2.8}$$

3. The BCQ. We begin with the following definitions adopted from [32, 36]. In the remainder, we shall adopt the convention that $\operatorname{cone} A = \{0\}$ when A is an empty set.

DEFINITION 3.1. *Let C be a convex set in X. The family $\{g_i : i \in I\}$ is said to satisfy*

(i) *the PLV property at $x \in X$ if*

$$\partial G(x) = \operatorname{co} \bigcup_{i \in \bar{I}(x)} \partial g_i(x); \tag{3.1}$$

(ii) *the BCQ at $x \in S$ if*

$$N_S(x) = \operatorname{cone} \bigcup_{i \in I(x)} \partial g_i(x); \tag{3.2}$$

(ii′) *the BCQ relative to C at $x \in C \cap S$ if*

$$N_{C \cap S}(x) = N_C(x) + \operatorname{cone} \bigcup_{i \in I(x)} \partial g_i(x); \tag{3.3}$$

(iii) *the PLV property (resp., the BCQ, the BCQ relative to C) if* (3.1) *(resp.,* (3.2)*,* (3.3)*) holds for each $x \in X$ (resp., $x \in S$, $x \in C \cap S$).*

Remark 3.1. In [16, Definition 2] (under the assumption that each g_i is lower semicontinuous), the property that the family $\{g_i : i \in I\}$ satisfies the BCQ relative to C was also described as the system $\{\delta_C;\ g_i,\ i \in I\}$ being locally Farkas–Minkowski(FM).

A relationship between the notions (ii) and (ii)′ in Definition 3.1 is shown in the following proposition.

PROPOSITION 3.1. *Consider a convex set C and $x \in C \cap S$. Then the family $\{g_i : i \in I\}$ satisfies the BCQ relative to C at x if and only if the family $\{\delta_C;\ g_i,\ i \in I\}$*

satisfies the BCQ at x. Consequently, the family $\{g_i : i \in I\}$ satisfies the BCQ relative to C if and only if the family $\{\delta_C; g_i, i \in I\}$ satisfies the BCQ.

Proof. Take $j \notin I$ and set $g_j := \delta_C$. Writing $J := I \cup \{j\}$, the family $\{\delta_C; g_i, i \in I\}$ becomes $\{g_i : i \in J\}$ such that $C \cap S = \{y \in X : g_i(y) \leq 0 \ \forall i \in J\}$ and $J(x) = \{j\} \cup I(x)$, where for $x \in C \cap S$,

$$J(x) := \left\{ i \in J : \ g_i(x) = \max \left\{ \sup_{i \in I} g_i(x), \delta_C(x) \right\} = 0 \right\}.$$

Then by (2.3),

$$N_C(x) + \operatorname{cone} \bigcup_{i \in I(x)} \partial\, g_i(x) = \operatorname{cone} \bigcup_{i \in J(x)} \partial\, g_i(x).$$

Thus the first assertion follows. The second follows immediately from the first. ∎

Remark 3.2.

(i) We have

$$\partial\, G(x) \supseteq \operatorname{co} \bigcup_{i \in \tilde{I}(x)} \partial\, g_i(x) \quad \text{for each } x \in X. \tag{3.4}$$

Indeed, let $i \in \tilde{I}(x)$ and $y^* \in \partial\, g_i(x)$. Then $g_i(x) = G(x)$. Since $y^* \in \partial g_i(x)$ and g_i is proper, $g_i(x) \neq +\infty$. Now it follows that

$$\langle y^*, y - x \rangle \leq g_i(y) - g_i(x) \leq G(y) - G(x) \quad \text{for each } y \in X. \tag{3.5}$$

This shows that $y^* \in \partial\, G(x)$, and so (3.4) is proved. Thus, the family $\{g_i : i \in I\}$ has the PLV property at $x \in X$ if and only if

$$\partial\, G(x) \subseteq \operatorname{co} \bigcup_{i \in \tilde{I}(x)} \partial\, g_i(x). \tag{3.6}$$

Hence, the family $\{g_i : i \in I\}$ has the PLV property if and only if (3.6) holds for each $x \in \operatorname{dom} \partial\, G$.

(ii) If $i \in I(x)$ and $y^* \in \partial\, g(x)$, then $G(x) = 0$. It follows from (3.5) that $y^* \in N_S(x)$. Thus

$$N_S(x) \supseteq \operatorname{cone} \bigcup_{i \in I(x)} \partial\, g_i(x) \quad \text{for each } x \in S. \tag{3.7}$$

Therefore, the family $\{g_i : i \in I\}$ satisfies the BCQ at $x \in S$ if and only if

$$N_S(x) \subseteq \operatorname{cone} \bigcup_{i \in I(x)} \partial\, g_i(x). \tag{3.8}$$

(iii) Note that if $x \in \operatorname{int} S$, then $N_S(x) = \{0\}$. Recalling our convention $\operatorname{cone} \emptyset = \{0\}$ (see Definition 3.1(ii)), it follows from (3.7) that

$$\{0\} = N_S(x) = \operatorname{cone} \bigcup_{i \in I(x)} \partial\, g_i(x) \quad \text{for each } x \in \operatorname{int} S.$$

Hence, the family $\{g_i : i \in I\}$ satisfies the BCQ if and only if (3.8) holds for each $x \in S \backslash \operatorname{int} S$.

(iv) Applying parts (ii) and (iii) to the family of functions $\{\delta_C;\ g_i,\ i \in I\}$ in place of $\{g_i :\ i \in I\}$ and invoking Proposition 3.1, we obtain that the family $\{g_i :\ i \in I\}$ satisfies the BCQ relative to C at x if and only if we have

$$N_{C\cap S}(x) \subseteq N_C(x) + \operatorname{cone} \bigcup_{i\in I(x)} \partial\, g_i(x), \tag{3.9}$$

and that the family $\{g_i :\ i \in I\}$ satisfies the BCQ relative to C if and only if (3.9) holds for each $x \in (C \cap S)\backslash \operatorname{int}(C \cap S)$.

Recall from [13, 32, 34] that a family of convex sets $\{C_i :\ i \in I\}$ is said to have the strong conical hull intersection property (the strong CHIP) at $x \in \bigcap_{i\in I} C_i$ if

$$N_{\bigcap_{i\in I} C_i}(x) = \sum_{i\in I} N_{C_i}(x). \tag{3.10}$$

If (3.10) holds for every $x \in \bigcap_{i\in I} C_i$, then we say that the family has the strong CHIP.

PROPOSITION 3.2. *Let $x \in C \cap S$, and suppose that the family $\{g_i : i \in I\}$ satisfies the BCQ at x. Then $\{C, S\}$ has the strong CHIP at x if and only if the family $\{g_i :\ i \in I\}$ satisfies the BCQ relative to C at x.*

Proof. By the given assumption, (3.2) holds. Hence we have the following equivalences:

$$\begin{aligned}
&\{C, S\} \text{ has the strong CHIP at } x \\
&\quad\Leftrightarrow N_{C\cap S}(x) = N_C(x) + N_S(x) \\
&\quad\Leftrightarrow N_{C\cap S}(x) = N_C(x) + \operatorname{cone} \bigcup_{i\in I(x)} \partial\, g_i(x) \\
&\quad\Leftrightarrow \{g_i :\ i \in I\} \text{ satisfies the BCQ relative to } C \text{ at } x. \qquad \square
\end{aligned}$$

Recall from [21] that the inequality system $f \le 0$ satisfies the weak BCQ at $x \in S_f\backslash \operatorname{int} S_f$ if

$$N_{S_f}(x) \subseteq \operatorname{cone} \partial\, f(x) + N_{\operatorname{dom} f}(x), \tag{3.11}$$

where $S_f := \{x \in X :\ f(x) \le 0\}$. The following proposition describes a relationship between the BCQ and the weak BCQ .

PROPOSITION 3.3. *Let f be a proper convex function on X and $x \in S_f\backslash \operatorname{int} S_f$. Then the family $\{f, \delta_{\operatorname{dom} f}\}$ satisfies the BCQ at x if and only if the inequality system $f \le 0$ satisfies the weak BCQ at x.*

Proof. Write $g_1 = f$ and $g_2 = \delta_{\operatorname{dom} f}$. If x satisfies, in addition, that $f(x) < 0$, then both the necessary condition and the sufficient condition in the statement of the proposition are satisfied. In fact, since $x \notin \operatorname{int} S_f$, [21, Lemma 2.2] states that

$$N_{S_f}(x) = N_{\operatorname{dom} f}(x).$$

Consequently (3.11) holds, and the family $\{f, \delta_{\operatorname{dom} f}\}$ satisfies the BCQ at x because of (2.3) and $I(x) = \{2\}$ (as $f(x) < 0$).

Therefore, to complete our proof we need only consider the case when $f(x) = 0$. For this case note that $I(x) = \{1, 2\}$. Thus, by (2.3), the family $\{f, \delta_{\mathrm{dom}\, f}\}$ satisfies the BCQ at x if and only if

$$N_{S_f}(x) = \mathrm{cone}\, \partial f(x) + N_{\mathrm{dom}\, f}(x). \tag{3.12}$$

Since the inclusion $N_{S_f}(x) \supseteq \mathrm{cone}\, \partial f(x) + N_{\mathrm{dom}\, f}(x)$ holds trivially (thanks to $f(x) = 0$ and Remark 3.2(ii) as applied to $\{f, \delta_{\mathrm{dom}\, f}\}$), (3.11) and (3.12) are equivalent. This completes the proof. □

4. The epigraph hull property. Recall that the meaning of $\{g_i : i \in I\}, G$, X, S, and I has been specified in section 2. The sup-function G is sometimes denoted by $\sup_{i\in I} g_i$. Recall also that we always assume that G is proper.

DEFINITION 4.1. *The family $\{g_i : i \in I\}$ is said to have the following:*

(i) *the convex EHP if*

$$\mathrm{epi} \left(\sup_{i\in I} g_i\right)^* = \mathrm{co} \bigcup_{i\in I} \mathrm{epi}\, g_i^*; \tag{4.1}$$

(ii) *the conical EHP if*

$$\mathrm{epi}\, \sigma_S = \mathrm{cone} \bigcup_{i\in I} \mathrm{epi}\, g_i^*. \tag{4.2}$$

Remark 4.1. It is routine to show that

$$\mathrm{epi} \left(\sup_{i\in I} g_i\right)^* \supseteq \mathrm{co} \bigcup_{i\in I} \mathrm{epi}\, g_i^* \quad \text{and} \quad \mathrm{epi}\, \sigma_S \supseteq \mathrm{cone} \bigcup_{i\in I} \mathrm{epi}\, g_i^*.$$

Thus the family has the convex EHP (resp., conical EHP) if and only if

$$\mathrm{epi} \left(\sup_{i\in I} g_i\right)^* \subseteq \mathrm{co} \bigcup_{i\in I} \mathrm{epi}\, g_i^*, \quad (\text{resp.}, \quad \mathrm{epi}\, \sigma_S \subseteq \mathrm{cone} \bigcup_{i\in I} \mathrm{epi}\, g_i^*.)$$

Results in the following proposition are known: for (i) see [35] and for (ii) see [16, 23, 30]. Recall that we have assumed that $\sup_{i\in I} g_i$ is proper.

PROPOSITION 4.1. *Suppose in addition that each g_i is lower semicontinuous. Then the following assertions regarding epigraphs hold:*

(i)

$$\mathrm{epi} \left(\sup_{i\in I} g_i\right)^* = \overline{\mathrm{co} \bigcup_{i\in I} \mathrm{epi}\, g_i^*}^{w^*}. \tag{4.3}$$

(ii)

$$\mathrm{epi}\, \sigma_S = \mathrm{epi}\, \delta_S^* = \overline{\mathrm{cone} \bigcup_{i\in I} \mathrm{epi}\, g_i^*}^{w^*} \quad \textit{if } S \textit{ is nonempty.} \tag{4.4}$$

COROLLARY 4.1. *Let $\{g_i : i \in I\}$ be as in the preceding proposition. Then the following assertions are valid:*

(i) *$\{g_i : i \in I\}$ has the convex EHP if and only if*

$$\mathrm{co} \bigcup_{i\in I} \mathrm{epi}\, g_i^* \quad \textit{is } w^*\textit{-closed;}$$

(ii) $\{g_i : i \in I\}$ *has the conical EHP if and only if*

$$\operatorname{cone} \bigcup_{i \in I} \operatorname{epi} g_i^* \quad \textit{is } w^*\textit{-closed,}$$

provided that $S \neq \emptyset$.

Proof. Since $G = \sup_{i \in I} g_i$ is proper, (4.3) holds. Thus (i) is seen to hold. Similarly, (ii) holds by (4.4), provided that $S \neq \emptyset$. □

Remark 4.2. In [16, Definition 1] and [17, Definition 3.1] (under the assumption that each g_i is lower semicontinuous), the system $\{\delta_C;\ g_i,\ i \in I\}$ is said to be FM if

$$\operatorname{epi} \sigma_C + \operatorname{cone} \bigcup_{i \in I} \operatorname{epi} g_i^* \quad \text{is } w^*\text{-closed.}$$

Letting $i_0 \notin I$ and writing $g_{i_0} := \delta_C$, one sees that $\{\delta_C; g_i,\ i \in I\}$ being FM is equivalent to $\{g_i,\ i \in I \cup \{i_0\}\}$ having conical EHP.

The following example shows that the lower semicontinuity assumption for $\{g_i : i \in I\}$ cannot be dropped in Proposition 4.1. In other words, Corollary 4.1 fails without a lower semicontinuity assumption on $\{g_i : i \in I\}$.

Example 4.1. Consider the real Hilbert space l^2 of square-summable series, and let Ω_+ be the convex subset defined by

$$\Omega_+ := \{x \in l^2 : x_i \geq 0\ \forall i \in \mathbb{N},\ x_i \neq 0 \text{ for at most finitely many } i\},$$

where x_i denotes the ith coordinate of x. Let $I = \{t \in \mathbb{R} : t > 0\}$, and define a family $\{g_t : t \in I\}$ of proper convex functions by

$$g_t(x) := \begin{cases} -t \sum_{i=1}^{\infty} i x_i & \text{if } x \in \Omega_+, \\ +\infty & \text{otherwise,} \end{cases} \quad \text{for each } t \in I.$$

Note in particular that $g_t(\cdot) \leq 0$ on $\operatorname{dom} g_t$, $\operatorname{dom} g_t = \Omega_+$, and $\{x : g_t(x) \leq 0\} = \Omega_+$ for each $t \in I$. Thus $S := \bigcap_{t \in I} \{x : g_t(x) \leq 0\} = \Omega_+$.

Let $y := (1, \frac{1}{2}, \dots, \frac{1}{n}, \dots)$ and $y_n := (1, \frac{1}{2}, \dots, \frac{1}{n}, 0, \dots)$ for each natural number n. Then $y \in l^2 \backslash \Omega_+$ and $y_n \in \Omega_+$ for each n. Furthermore, one has that $y_n \to y$ and

$$\lim g_t(y_n) = \lim_{n \to \infty} -nt = -\infty < g_t(y) = +\infty \quad \text{for each } t \in I$$

(so each g_t is not lower semicontinuous); consequently, $\overline{g_t}(y) = -\infty$, where $\overline{g_t}$ denotes the closure of the function g_t (cf. [40, p. 62]). Since $g_t^* = \overline{g_t}^*$ by [40, Theorem 2.3.1(iv)], we see that

$$g_t^*(x^*) = \overline{g_t}^*(x^*) \geq \langle x^*, y \rangle - \overline{g}_t(y) = +\infty \quad \text{for each } x^* \in l^2 \text{ and } t \in I.$$

Thus

$$\operatorname{epi} g_t^* = \emptyset \quad \text{for each } t \in I.$$

On the other hand, $(\sup_{t \in I} g_t)^* = \sigma_{\Omega_+}$, which is proper since $\Omega_+ \neq \emptyset$. Hence the lower semicontinuity assumption in Proposition 4.1 cannot be dropped.

PROPOSITION 4.2. *The family* $\{g_i, \delta_{\operatorname{dom} g_i} : i \in I\}$ *has the conical EHP if and only if the family* $\{tg_i, \delta_{\operatorname{dom} g_i} : i \in I, t > 0\}$ *has the convex EHP.*

Proof. By definition, the family $\{g_i, \delta_{\operatorname{dom} g_i} : i \in I\}$ has the conical EHP if and only if

$$\operatorname{epi} \sigma_S = \operatorname{cone} \bigcup_{i\in I} (\operatorname{epi} g_i^* \cup \operatorname{epi} \sigma_{\operatorname{dom} g_i}), \tag{4.5}$$

while the family $\{tg_i, \delta_{\operatorname{dom} g_i} : i \in I, t > 0\}$ has the convex EHP if and only if

$$\operatorname{epi} \sigma_S = \operatorname{co} \bigcup_{i\in I} \left[\bigcup_{t>0} \operatorname{epi} (tg_i)^* \cup \operatorname{epi} \sigma_{\operatorname{dom} g_i} \right], \tag{4.6}$$

thanks to the easily checked equality $\delta_S = \sup_{t>0, i\in I} tg_i$. It suffices to prove that the sets on the right-hand side of (4.5) and of (4.6) are equal. To do this, recall from [40, Theorem 2.3.1(v)] that $(tG)^*(x^*) = tG^*(\frac{x^*}{t})$ for each $t > 0$ and $x^* \in X^*$. It follows that

$$\operatorname{epi} (tg_i)^* = t \operatorname{epi} g_i^* \quad \text{for each } i \in I \text{ and } t > 0.$$

Since $\operatorname{epi} \sigma_{\operatorname{dom} g_i}$ is a cone, it follows that

$$\begin{aligned}
\operatorname{co} \bigcup_{i\in I} \left[\bigcup_{t>0} \operatorname{epi} (tg_i)^* \cup \operatorname{epi} \sigma_{\operatorname{dom} g_i} \right] &= \operatorname{co} \bigcup_{i\in I} \left[\bigcup_{t>0} t \operatorname{epi} g_i^* \cup \operatorname{epi} \sigma_{\operatorname{dom} g_i} \right] \\
&= \operatorname{co} \bigcup_{i\in I} \bigcup_{t>0} t (\operatorname{epi} g_i^* \cup \operatorname{epi} \sigma_{\operatorname{dom} g_i}) \\
&= \operatorname{co} \bigcup_{i\in I} \bigcup_{t\geq 0} t (\operatorname{epi} g_i^* \cup \operatorname{epi} \sigma_{\operatorname{dom} g_i}) \\
&= \operatorname{cone} \bigcup_{i\in I} (\operatorname{epi} g_i^* \cup \operatorname{epi} \sigma_{\operatorname{dom} g_i}),
\end{aligned}$$

where the third equality holds because $(0, 0) \in \operatorname{epi} \sigma_{\operatorname{dom} g_i}$ for each $i \in I$. This completes the proof. □

Remark 4.3. The first part of the second conclusion of the following theorem was also independently obtained in [16, Corollary 2] for the special case when g_i were assumed to be lower semicontinuous.

THEOREM 4.1. *The following assertions are valid:*

(i) *If the family $\{g_i : i \in I\}$ has the convex EHP, then it has the PLV property. The converse implication also holds if* $\operatorname{dom} G^* \subseteq \operatorname{im} \partial G$.

(ii) *Suppose $S \neq \emptyset$. If the family $\{g_i : i \in I\}$ has the conical EHP, then it satisfies the BCQ. The converse implication also holds if* $\operatorname{dom} \sigma_S \subseteq \operatorname{im} \partial \delta_S$.

Proof. (i) Suppose that the family $\{g_i : i \in I\}$ has the convex EHP. By Remark 3.2(i), it suffices to show that (3.6) holds for each $x \in \operatorname{dom} \partial G$. Take $x^* \in \partial G(x)$. By (2.2), $(x^*, \langle x^*, x\rangle - G(x)) \in \operatorname{epi} G^*$. Now (4.1) implies that $(x^*, \langle x^*, x\rangle - G(x))$ can be represented as

$$(x^*, \langle x^*, x\rangle - G(x)) = \sum_{i\in J} \lambda_i (x_i^*, \alpha_i),$$

for some finite subset $J \subseteq I$, $(x_i^*, \alpha_i) \in \operatorname{epi} g_i^*$, $i \in J$, and $0 < \lambda_i \leq 1$ with $\sum_{i\in J} \lambda_i = 1$. This implies

$$\langle x^*, x\rangle - G(x) = \sum_{i\in J} \lambda_i \alpha_i \geq \sum_{i\in J} \lambda_i g_i^*(x_i^*) \geq \sum_{i\in J} \lambda_i (\langle x_i^*, x\rangle - g_i(x)). \tag{4.7}$$

The equalities hold throughout (4.7) because $\langle x^*, x\rangle = \sum_{i\in J} \lambda_i \langle x_i^*, x\rangle$ and $g_i(x) \leq G(x)$ for each i. As $\lambda_i \neq 0$ for each $i \in J$, it follows that $g_i(x) = G(x)$ and $g_i^*(x_i^*) = \langle x_i^*, x\rangle - g_i(x)$ for each $i \in J$. Thus $J \subseteq \tilde{I}(x)$ and $x_i^* \in \partial\, g_i(x)$ for $i \in J$, thanks to (2.1). Hence

$$x^* = \sum_{i\in \tilde{I}(x)} \lambda_i x_i^* \in \operatorname{co} \bigcup_{i\in \tilde{I}(x)} \partial\, g_i(x);$$

i.e., the family $\{g_i : i \in I\}$ has the PLV property. This proves the first part of (i).

Now we assume $\operatorname{dom} G^* \subseteq \operatorname{im} \partial\, G$ and prove the converse implication. In view of Remark 4.1, we need to show only that

$$\operatorname{epi} G^* \subseteq \operatorname{co} \bigcup_{i\in I} \operatorname{epi} g_i^*. \tag{4.8}$$

Take $(y^*, \alpha) \in \operatorname{epi} G^*$. Then $y^* \in \operatorname{dom} G^*$, and by assumption there exists $x \in X$ such that $y^* \in \partial\, G(x)$. Now (3.1) implies that y^* can be represented as

$$y^* = \sum_{i\in J} \lambda_i y_i^*$$

for some finite subset $J \subseteq \tilde{I}(x)$, $y_i^* \in \partial\, g_i(x)$ for each $i \in J$, and $0 < \lambda_i \leq 1$ with $\sum_{i\in J} \lambda_i = 1$. Note that, for each $i \in J$, $\langle y_i^*, x\rangle - G(x) = g_i^*(y_i^*)$ because $y_i^* \in \partial\, g_i(x)$ and $G(x) = g_i(x)$. Since

$$\alpha \geq \langle y^*, x\rangle - G(x) = \sum_{i\in J} \lambda_i (\langle y_i^*, x\rangle - g_i(x)),$$

there exists a set $\{\alpha_i : i \in J\}$ of real numbers such that

$$\alpha = \sum_{i\in J} \lambda_i \alpha_i \quad \text{and} \quad g_i^*(y_i^*) = \langle y_i^*, x\rangle - g_i(x) \leq \alpha_i \quad \text{for each } j \in J.$$

This implies that $(y_i^*, \alpha_i) \in \operatorname{epi} g_i^*$ for each i and thus $(y^*, \alpha) \in \operatorname{co} \bigcup_{i\in I} \operatorname{epi} g_i^*$. Hence (4.8) is proved.

(ii) Suppose that the family has the conical EHP. We wish to show that it satisfies the BCQ, that is, to show that (3.8) holds for each $x \in S \backslash \operatorname{int} S$ (see Remark 3.2(iii)). Take $x \in S \backslash \operatorname{int} S$ and $x^* \in N_S(x)$. Since the set on the right-hand side of (3.8) contains the origin, we assume without loss of generality that $x^* \neq 0$. By (2.4), $(x^*, \langle x^*, x\rangle) \in \operatorname{epi} \sigma_S$. Now (4.2) implies that $(x^*, \langle x^*, x\rangle)$ can be represented as

$$(x^*, \langle x^*, x\rangle) = \sum_{i\in J} \lambda_i (x_i^*, \alpha_i),$$

for some finite subset $J \subseteq I$, $(x_i^*, \alpha_i) \in \operatorname{epi} g_i^*$, $\lambda_i > 0$, $i \in J$. Then we have

$$\langle x^*, x\rangle = \sum_{i\in J} \lambda_i \alpha_i \geq \sum_{i\in J} \lambda_i g_i^*(x_i^*) \geq \sum_{i\in J} \lambda_i (\langle x_i^*, x\rangle - g_i(x)). \tag{4.9}$$

Since $\langle x^*, x\rangle = \sum_{i\in J} \lambda_i \langle x_i^*, x\rangle$ and $g_i(x) \leq 0$ for each $i \in J$, the equalities in (4.9) hold throughout. Since $\lambda_i \neq 0$ for each $i \in J$, we obtain that for each $i \in J$

$$g_i(x) = 0 \tag{4.10}$$

and

$$g_i^*(x_i^*) = \langle x_i^*, x\rangle. \tag{4.11}$$

It follows from (4.10) that $J \subseteq I(x)$. Also, summing up (4.10) and (4.11), we obtain for each $i \in J$ that

$$g_i^*(x_i^*) + g_i(x) = \langle x_i^*, x\rangle,$$

which, by (2.1), is equivalent to $x_i^* \in \partial g_i(x)$. Thus we have

$$x^* = \sum_{i\in J} \lambda_i x_i^* \in \operatorname{cone} \bigcup_{i\in I(x)} \partial\, g_i(x).$$

Therefore the family $\{g_i : i \in I\}$ satisfies the BCQ.

We now turn to the converse implication. Assume $\operatorname{dom}\sigma_S \subseteq \operatorname{im}\partial\,\delta_S$. In view of Remark 4.1, we need to show only that

$$\operatorname{epi}\sigma_S \subseteq \operatorname{cone}\bigcup_{i\in I} \operatorname{epi} g_i^*. \tag{4.12}$$

Take $(y^*, \alpha) \in \operatorname{epi}\sigma_S$. Since $(0,0)$ clearly belongs to the right-hand side of (4.12), we assume without loss of generality that $(y^*, \alpha) \neq (0,0)$. Now, since $y^* \in \operatorname{dom}\sigma_S \subseteq \operatorname{im}\partial\,\delta_S$, there exists $x_0 \in S$ such that $y^* \in \partial\,\delta_S(x_0) = N_S(x_0)$ by (2.3). The definition of BCQ implies that y^* can be expressed as

$$y^* = \sum_{i\in J} \lambda_i y_i^*$$

for some finite subset $J \subseteq I(x_0)$, $y_i^* \in \partial\, g_i(x_0)$, and $\lambda_i \geq 0$ for each $i \in J$. Note that, for each $i \in J$, $\langle y_i^*, x_0\rangle = g_i^*(y_i^*)$ because $y_i^* \in \partial\, g_i(x_0)$ and $g_i(x_0) = G(x_0) = 0$. On the other hand, since $\alpha \geq \langle y^*, x_0\rangle = \sum_{i\in J} \lambda_i \langle y_i^*, x_0\rangle$, there exists a set $\{\alpha_i : i \in J\}$ of real numbers such that

$$\alpha = \sum_{i\in J} \lambda_i \alpha_i \quad \text{and} \quad g_i^*(y_i^*) = \langle y_i^*, x_0\rangle \leq \alpha_i \quad \text{for each } i \in J.$$

This implies that $(y_i^*, \alpha_i) \in \operatorname{epi} g_i^*$ for each i and thus $(y^*, \alpha) \in \operatorname{cone}\bigcup_{i\in I} \operatorname{epi} g_i^*$. Hence (4.12) is proved. □

Recall from [35] that a family of convex sets $\{C_i : i \in I\}$ in X with nonempty intersection satisfies the sum of epigraphs constraint qualification (SECQ) if

$$\operatorname{epi}\sigma_{\bigcap_{i\in I} C_i} = \sum_{i\in I} \operatorname{epi}\sigma_{C_i}. \tag{4.13}$$

The following proposition is on the relationships between the strong CHIP, the SECQ for a family of convex sets, and the conical EHP for the family consisting of the corresponding indicator functions.

PROPOSITION 4.3. *Let $\{C_i : i \in I\}$ be a family of convex sets in X with nonempty intersection. Then the following assertions are valid:*

(i) *The family $\{C_i : i \in I\}$ has the strong CHIP if and only if the family of functions $\{\delta_{C_i} : i \in I\}$ has the BCQ.*

(ii) *The family $\{C_i : i \in I\}$ satisfies the SECQ if and only if the family of functions $\{\delta_{C_i} : i \in I\}$ has the conical EHP.*

Proof. Consider the family of functions $\{\delta_{C_i} : i \in I\}$, i.e., $g_i := \delta_{C_i}$ for each $i \in I$. Then,

$$S := \left\{x : \sup_{i\in I} \delta_{C_i}(x) \le 0\right\} = \{x : \delta_{\bigcap_{i\in I} C_i}(x) \le 0\} = \bigcap_{i\in I} C_i \neq \emptyset.$$

Also, $G(x) := \sup_{i\in I} \delta_{C_i}(x) = \delta_{\bigcap_{i\in I} C_i}(x)$ and $G(x) = 0$ for each $x \in S$. Moreover, since $\delta_{C_i}(x) = 0$ for each $x \in S$ and each $i \in I$, it follows that $I(x) = I$ for each $x \in S$.

(i) For each $x \in S$, we have from (2.3) that

$$\operatorname{cone} \bigcup_{i\in I(x)} \partial\, \delta_{C_i}(x) = \operatorname{cone} \bigcup_{i\in I} \partial\, \delta_{C_i}(x) = \operatorname{cone} \bigcup_{i\in I} N_{C_i}(x)$$

$$= \sum_{i\in I} \operatorname{cone} N_{C_i}(x) = \sum_{i\in I} N_{C_i}(x).$$

Therefore, when $\{\delta_{C_i} : i \in I\}$ replaces $\{g_i : i \in I\}$, we see that (3.2) and (3.10) are equivalent, and so (i) is proved.

(ii) Note that each $\operatorname{epi} \sigma_{C_i}$ is a cone, and so

$$\operatorname{cone} \bigcup_{i\in I} \operatorname{epi} \sigma_{C_i} = \sum_{i\in I} \operatorname{cone} (\operatorname{epi} \sigma_{C_i}) = \sum_{i\in I} \operatorname{epi} \sigma_{C_i}.$$

Therefore, when $\{\delta_{C_i} : i \in I\}$ replaces $\{g_i : i \in I\}$, we see that (4.2) and (4.13) are equivalent, and so (ii) is proved. □

COROLLARY 4.2 (see [35]). *Let $\{C_i : i \in I\}$ be a family of convex sets in X with nonempty intersection. If the family satisfies the SECQ, then it has the strong CHIP. The converse implication also holds if $\operatorname{dom} \sigma_S \subseteq \operatorname{im} \partial\, \delta_S$, where $S = \bigcap_{i\in I} C_i$.*

Proof. The corollary follows from Proposition 4.3 and Theorem 4.1(ii) (applied to the family of functions $\{\delta_{C_i} : i \in I\}$ in place of $\{g_i : i \in I\}$). □

Adopting a definition given in [36, Definition 5.3] originally in a more restrictive case, we say that a linear inequality

$$\langle a^*, x\rangle \le b \tag{4.14}$$

(where $a^* \in X^*$ and $b \in \mathbb{R}$) is a consequence relation of (1.1) if every $x \in S$ satisfies (4.14). Moreover, the system (1.1) is said to be a convex FM system if every linear consequence relation of the system (1.1) is also a consequence relation of some finite subsystem of it. The following result was independently obtained in [16, Proposition 1] under the additional assumption that each g_i is lower semicontinuous.

PROPOSITION 4.4. *Suppose that $S \neq \emptyset$ and that the family $\{g_i : i \in I\}$ has the conical EHP. Then the system* (1.1) *is a convex FM system.*

Proof. Let $a^* \in X^*$ and $b \in \mathbb{R}$ be such that $\langle a^*, x\rangle \le b$ for each $x \in S$. This means that

$$(a^*, b) \in \operatorname{epi} \sigma_S,$$

thanks to (2.5). Since the family $\{g_i : i \in I\}$ has the conical EHP, it follows from (4.2) that

$$(a^*, b) \in \operatorname{cone} \bigcup_{i\in I} \operatorname{epi} g_i^*.$$

Thus, there exists a finite subset $J \subseteq I$ such that

$$(a^*, b) \in \operatorname{cone} \bigcup_{i\in J} \operatorname{epi} g_i^* \subseteq \operatorname{epi} \sigma_{S_J},$$

where $S_J := \{x \in X : g_i(x) \le 0 \ \forall i \in J\}$, and the inclusion follows from Remark 4.1. Again by (2.5), one has $\langle a^*, x\rangle \le b$ for each $x \in S_J$. This completes the proof. □

The following corollary was proved by Li, Nahak, and Singer in [36, Proposition 5.4] under the additional assumptions that $X = \mathbb{R}^n$, S is compact, and each g_i is continuous. (Recall that if S is a weakly compact convex set and X is a normed linear space, then $\operatorname{dom} \sigma_S \subseteq \operatorname{im} \partial\delta_S$; see [35, Proposition 3.1].) A similar result was obtained in [16, Proposition 3], in which they assumed the family of lower semicontinuous functions to have BCQ at a point z and deduced that every linear consequence relation (4.14) of the system (1.1) with $b = \langle a^* z\rangle$ is a consequence relation of some finite subsystem of it.

COROLLARY 4.3. *Suppose that $S \ne \emptyset$ and that the family $\{g_i : i \in I\}$ satisfies the BCQ. Suppose further that* $\operatorname{dom} \sigma_S \subseteq \operatorname{im} \partial\, \delta_S$. *Then* (1.1) *is a convex FM system.*

Proof. By Theorem 4.1(ii), the assumptions imply that the family $\{g_i : i \in I\}$ has the conical EHP. Hence the conclusion follows from Proposition 4.4. □

5. Optimality conditions. Let X be a locally convex Hausdorff topological vector space as before. We use $\Gamma(X)$ to denote the class of all proper convex lower semicontinuous functions on X as in [40]. For a subset of X, we define

$$\mathcal{F}_A := \{f \in \Gamma(X) : \operatorname{dom} f \cap A \ne \emptyset,\ \operatorname{epi} \sigma_A + \operatorname{epi} f^* \text{ is } w^*\text{-closed}\}.$$

Since $\operatorname{epi} \sigma_A = \operatorname{epi} \sigma_{\overline{A}}$ for any convex set A,

$$f \in \mathcal{F}_A \Leftrightarrow f \in \mathcal{F}_{\overline{A}}. \tag{5.1}$$

It is known from [4, Theorem 3.2] that if $f \in \mathcal{F}_A$ and closed convex set A are such that $\operatorname{epi} \sigma_A + \operatorname{epi} f^*$ is w^*-closed, then the subdifferential sum formula holds:

$$f \in \mathcal{F}_A \Rightarrow \partial\,(f + \delta_A)(x) = \partial\, f(x) + \partial\, \delta_A(x) \quad \text{for each } x \in A \cap \operatorname{dom} f.$$

Thus, (5.1) entails that

$$f \in \mathcal{F}_A \Rightarrow \partial\,(f + \delta_{\overline{A}})(x) = \partial\, f(x) + \partial\, \delta_{\overline{A}}(x) \quad \text{for each } x \in \overline{A} \cap \operatorname{dom} f. \tag{5.2}$$

As in [40], let $\Lambda(X)$ denote the class of all proper convex functions on X. Let $f \in \Lambda(X)$, and recall that the meanings of $\{g_i : i \in I\}, G,\ X,\ S$, and I have been specified in section 2 and that we always assume that G is proper. We consider the following minimization problem:

$$\begin{aligned} &\text{Minimize} && f(x), \\ &\text{s.t.} && g_i(x) \le 0,\ i \in I. \end{aligned} \tag{5.3}$$

Clearly, $\overline{x} \in S$ is a minimizer of (5.3) if and only if it is a minimizer of (5.4) defined as follows:

$$\begin{array}{ll} \text{Minimize} & f(x), \\ \text{s.t.} & x \in S. \end{array} \tag{5.4}$$

The following theorem gives a characterization for a feasible point $\overline{x}$ to be a minimizer. Note in particular that it improves a result in [16, Theorem 4] as far as the lower semicontinuity of the functions g_i is relaxed. See also [6] for other related results. For $h \in \Lambda(X)$, let $\operatorname{cont} h$ denote the set of all points at each of which h is continuous, that is,

$$\operatorname{cont} h := \{x \in X : \ h \text{ is continuous at } x\}.$$

THEOREM 5.1. *Let $\overline{x}$ be a feasible point of* (5.3). *Then the following statements are equivalent:*

(i) *The family $\{g_i : i \in I\}$ satisfies the BCQ at $\overline{x}$.*

(ii) *For each $f \in \mathcal{F}_S$, $\overline{x}$ is a minimizer of* (5.4) *with $\overline{S}$ in place of S if and only if there exist a finite subset $J \subseteq I(\overline{x})$ and $\lambda_i \geq 0$, $i \in J$, such that*

$$0 \in \partial f(\overline{x}) + \sum_{i \in J} \lambda_i \partial g_i(\overline{x}). \tag{5.5}$$

(iii) *For any $f \in \Lambda(X)$ such that $\operatorname{cont} f \cap S \neq \emptyset$, $\overline{x}$ is a minimizer of* (5.4) *if and only if there exist a finite subset $J \subseteq I(\overline{x})$ and $\lambda_i \geq 0$, $i \in J$, such that* (5.5) *holds.*

(iv) *For each continuous linear functional f, $\overline{x}$ is a minimizer of* (5.4) *if and only if there exist a finite subset $J \subseteq I(\overline{x})$ and $\lambda_i \geq 0$, $i \in J$, such that* (5.5) *holds.*

Proof. Recall a well-known result in convex analysis (cf. [40, Theorem 2.5.7]) that if $f \in \Lambda(X)$ and A is a convex subset, then

$$\overline{x} \text{ minimizes } f \text{ on } A \Longleftrightarrow \overline{x} \text{ minimizes } (f + \delta_A) \text{ on } X \Longleftrightarrow 0 \in \partial(f + \delta_A)(\overline{x}). \tag{5.6}$$

We now first prove (i)⇒(ii). Fix $f \in \mathcal{F}_S$. By (5.2), we have

$$\partial\,(f + \delta_{\overline{S}})(x) = \partial\, f(x) + \partial\, \delta_{\overline{S}}(x) \quad \text{for each } x \in \overline{S} \cap \operatorname{dom} f. \tag{5.7}$$

By (2.3), we know further that

$$\partial\, \delta_{\overline{S}}(x) = N_{\overline{S}}(x) = N_S(x) \quad \text{for each } x \in S. \tag{5.8}$$

Thus by (5.6), (5.7), and (5.8), the assumption (i) implies the following equivalences:

$$\overline{x} \text{ minimizes } f \text{ on } \overline{S} \Longleftrightarrow 0 \in \partial\, f(\overline{x}) + N_S(\overline{x}) \Longleftrightarrow 0 \in \partial\, f(\overline{x}) + \operatorname{cone} \bigcup_{i \in I(\overline{x})} \partial\, g_i(\overline{x}).$$

The implication (i)⇒(ii) is now clear.

Next, we prove (i)⇒(iii). Let $f \in \Lambda(X)$ be such that $\operatorname{cont} f \cap S \neq \emptyset$. Then [40, Theorem 2.8.7(iii)] states that

$$\partial\,(f + \delta_S)(x) = \partial\, f(x) + \partial\, \delta_S(x) \quad \text{for each } x \in S \cap \operatorname{dom} f. \tag{5.9}$$

Thus the implication (i)⇒(iii) is seen to hold by (5.6) and (5.9).

To show the implication (ii)⇒(iv), let f be a continuous linear functional on X. Then $\inf_{x\in S} f(x) = \inf_{x\in\overline{S}} f(x)$ and $f \in \mathcal{F}_{\overline{S}}$ by [15, Remark 5.6]. The latter condition is equivalent to $f \in \mathcal{F}_S$, thanks to (5.1). Since $\overline{x} \in S$, the implication (ii)⇒(iv) is clear.

The implication (iii)⇒(iv) is immediate.

Finally, we turn to the proof of (iv)⇒(i). We need to show that (3.8) holds for $x = \overline{x}$. Let $y^* \in N_S(\overline{x})$. Then $\overline{x}$ is a minimizer of the following optimization problem:

$$\begin{aligned} &\text{Minimize} && -\langle y^*, x\rangle,\\ &\text{s.t.} && x \in S. \end{aligned}$$

By (iv), there exist a finite subset $J \subseteq I(\overline{x})$ and $\lambda_i \geq 0$, $i \in J$, such that

$$0 \in -y^* + \sum_{i\in J} \lambda_i \partial\, g_i(\overline{x}).$$

Thus

$$y^* \in \sum_{i\in J} \lambda_i \partial\, g_i(\overline{x}) \subseteq \operatorname{cone} \bigcup_{i\in I(\overline{x})} \partial\, g_i(\overline{x}).$$

Therefore $\overline{x}$ satisfies (3.8), as required. This completes the proof. □

Let A be a convex set in X and $a \in A$. In the literature there are several unequivalent conditions on "the relative interior" of A such as

$$a \in U \cap \overline{\operatorname{aff} A} \subseteq A \tag{5.10}$$

and

$$a \in U \cap \operatorname{aff} A \subseteq A, \tag{5.11}$$

where U is a neighborhood of a. For example, (5.10) was considered in [3, 20] and (5.11) in [2, 40]. As (5.11) is not needed for our present study, we will follow the terminology of [3, 20] to say that a is in the relative interior of A and denoted by $a \in \operatorname{ri} A$ if there exists a neighborhood of a such that (5.10) holds.

Remark 5.1. If $0 \in \operatorname{ri} A$ and $x_0 \in \overline{A}$, then $tx_0 \in \operatorname{ri} A$ for each $t \in [0, 1)$. To see this, consider the Minkowski functional $p_A(x) := \inf\{\lambda :\ \lambda^{-1}x \in A\}$ for each $x \in \overline{\operatorname{span} A}$. By considering A and $\overline{A}$ as subsets of $\overline{\operatorname{span} A}$, we have $\operatorname{ri} A = \{x \in \overline{\operatorname{span} A} :\ p_A(x) < 1\}$ and $\overline{A} = \{x \in \overline{\operatorname{span} A} :\ p_A(x) \leq 1\}$ (cf. [40, Proposition 1.1.1(ii)]). Since $p_A(x_0) \leq 1$, we have $p_A(tx_0) = tp_A(x_0) < 1$ for each $t \in [0, 1)$. Thus $tx_0 \in \operatorname{ri} A$, as required.

For $f \in \Lambda(X)$, define

$$\operatorname{cont}_A f = \{x \in \operatorname{dom} f \cap \overline{\operatorname{aff} A} :\ \ f|_{\overline{\operatorname{aff} A}} \text{ is continuous at } x\}.$$

Remark 5.2. If $0 \in \operatorname{cont}_A f$ and $x_0 \in \operatorname{dom} f \cap \overline{\operatorname{span} A}$, then $tx_0 \in \operatorname{cont}_A f$ for all $t \in [0, 1)$. To see this, consider $\operatorname{dom} f \cap \overline{\operatorname{span} A}$ as a subset of $\overline{\operatorname{span} A}$; in particular, $0 \in \operatorname{cont}_A f \subseteq \operatorname{int}(\operatorname{dom} f \cap \overline{\operatorname{span}} A)$. Since $x_0 \in \operatorname{dom} f \cap \overline{\operatorname{span} A}$, it follows that $tx_0 \in \operatorname{int}(\operatorname{dom} f \cap \overline{\operatorname{span}} A)$ for all $t \in [0, 1)$. Consequently, by the convexity of f, $tx_0 \in \operatorname{cont}_A f$ for all $t \in [0, 1)$ thanks to [40, Theorem 2.2.9] because $f|_{\overline{\operatorname{span} A}}$ is continuous at 0.

LEMMA 5.1. *Consider the problem* (5.4)*, and let* $f \in \Lambda(X)$*. Suppose that*

$$(\operatorname{dom} f \cap \operatorname{ri} S) \cup (\overline{S} \cap \operatorname{cont}_S f) \neq \emptyset. \tag{5.12}$$

Then

$$\inf_{x\in S} f(x) = \inf_{x\in \overline{S}} f(x). \tag{5.13}$$

Proof. By (5.12), we assume without loss of generality that

$$0 \in (\operatorname{dom} f \cap \operatorname{ri} S) \cup (\overline{S} \cap \operatorname{cont}_S f).$$

Let $\lambda > \inf_{x\in\overline{S}} f(x)$ and take $x_0 \in \overline{S}$ such that $\lambda > f(x_0)$. To show (5.13) it suffices to show that $\lambda > \inf_{x\in S} f(x)$. By the convexity we have

$$f(tx_0) \le tf(x_0) + (1-t)f(0) \quad \text{for each } t \in [0,1]. \tag{5.14}$$

Letting $t \uparrow 1$ in (5.14), we obtain

$$\limsup_{t\to 1^-} f(tx_0) \le \lim_{t\to 1^-} [tf(x_0) + (1-t)f(0)] = f(x_0) < \lambda. \tag{5.15}$$

This and Remark 5.1 imply that $\inf_{x\in S} f(x) \le f(x_0)$ if $0 \in \operatorname{ri} S$ (so $tx_0 \in \operatorname{ri} S$ for each $t \in [0,1)$). It remains to consider the case when $0 \in \overline{S} \cap \operatorname{cont}_S f$. But then Remark 5.2 entails that tx_0 is a continuity point of $f|_{\overline{\operatorname{span} S}}$ if $t \in [0,1)$. Noting $tx_0 \in \overline{S}$, it follows from (5.15) that $f(x_t) < \lambda$ for $x_t \in S$ close enough to tx_0, provided that $t < 1$ is sufficiently near to 1. Therefore $\inf_{x\in\overline{S}} f(x) < \lambda$ in any case. This completes the proof. □

Remark 5.3. For two convex sets A, C in a Banach space X, recall from [34] that an $a \in A$ belongs to $\operatorname{rint}_{\overline{\operatorname{aff} C}} A$ if $a \in B(a,\epsilon) \cap \overline{\operatorname{aff} C} \subseteq A$ for some $\epsilon > 0$. Note that if X is a Banach space and $f \in \Gamma(X)$, then

$$\operatorname{rint}_{\overline{\operatorname{aff} S}} \operatorname{dom} f \subseteq \operatorname{cont}_S f. \tag{5.16}$$

To see this we assume without loss of generality that $0 \in S$. Then $\overline{\operatorname{aff} S} = \overline{\operatorname{span} S}$ is a Banach space. Since $f|_{\overline{\operatorname{aff} S}} \in \Gamma(\overline{\operatorname{aff} S})$, (5.16) follows from [40, Theorem 2.2.20].

COROLLARY 5.1. *Under the assumption of Theorem* 5.1*, for any* $\overline{x} \in S$*, the following statements are equivalent:*

(i) *The family* $\{g_i : i \in I\}$ *satisfies the BCQ at* $\overline{x}$*.*

(ii′) *For each* $f \in \mathcal{F}_S$ *satisfying* (5.12)*,* $\overline{x}$ *is a minimizer of* (5.4) *if and only if there exist a finite subset* $J \subseteq I(\overline{x})$ *and* $\lambda_i \ge 0$*,* $i \in J$*, such that* (5.5) *holds.*

Proof. Suppose that (i) holds. Then Theorem 5.1(ii) holds. Let $f \in \mathcal{F}_S$ satisfy (5.12). Then $\inf_{x\in S} f(x) = \inf_{x\in\overline{S}} f(x)$ by Lemma 5.1. Since $\overline{x} \in S$, applying Theorem 5.1(ii) to this f, (ii′) is seen to hold. Conversely, suppose that (ii′) holds. Then part (iv) (and so part (i)) of Theorem 5.1 holds because any continuous linear functional f on X belongs to $\mathcal{F}_S$ (by [15, Remark 5.6] and (5.1)) and satisfies (5.12) (since $\operatorname{dom} f = X$). The proof is complete. □

The following result was proved in [15, Theorem 5.5] under the additional assumption that each g_i is continuous, which was recently extended in [16, Theorem 3] to the setting that some g_i are allowed to be merely lower semicontinuous.

COROLLARY 5.2. *Suppose that* $f \in \mathcal{F}_S$ *and that the family* $\{g_i : i \in I\}$ *has the conical EHP. Assume that either* S *is closed or the condition* (5.12) *is satisfied. Let* $\overline{x} \in S$*. Then* $\overline{x}$ *is a minimizer of* (5.4) *if and only if there exist a finite subset* $J \subseteq I(\overline{x})$ *and* $\lambda_i \ge 0$*,* $i \in J$*, such that*

$$0 \in \partial f(\overline{x}) + \sum_{i\in J} \lambda_i \partial g_i(\overline{x}).$$

Proof. Since the family $\{g_i : i \in I\}$ has the conical EHP, it has the BCQ at $\overline{x}$ by Theorem 4.1(ii). Moreover, (5.13) holds by the assumptions and Lemma 5.1. Since $\overline{x} \in S$ it follows that $\overline{x}$ is a minimizer of (5.4) if and only if $\overline{x}$ is a minimizer of (5.4) but with $\overline{S}$ in place of S. Thus the corollary follows from the implication (i)⇒(ii) in Theorem 5.1. □

The following example shows that (5.13) and the related corollaries may fail if the assumption (5.12) is dropped.

Example 5.1. Define $S := \{(x,y) \in \mathbb{R}^2 : x \geq 0\} \backslash \{(0,y) \in \mathbb{R}^2 : y < 1\}$ and

$$f(x,y) := \begin{cases} \frac{y^2}{2} & \text{if } x = 0, \\ +\infty & \text{otherwise.} \end{cases}$$

Then f is proper convex lower semicontinuous, S is convex, and $\overline{S} = \{(x,y) \in \mathbb{R}^2 : x \geq 0\}$. Note that in this case $\operatorname{dom} f = \{(0,y) \in \mathbb{R}^2 : y \in \mathbb{R}\}$, which is disjoint from the set $\operatorname{ri} S = \operatorname{int} S = \{(x,y) \in \mathbb{R}^2 : x > 0\}$. It is easy to see that $\inf_{(x,y)\in S} f(x,y) = f(0,1) = \frac{1}{2}$ but $\inf_{(x,y)\in \overline{S}} f(x,y) = f(0,0) = 0$. Thus (5.13) fails.

Next we wish to show that $f \in \mathcal{F}_S$. To do this, note first that for each $(x,y) \in \mathbb{R}^2$

$$f^*(x,y) = \sup\{\langle (u,v),(x,y)\rangle - f(u,v) : (u,v) \in \operatorname{dom} f\} = \frac{y^2}{2},$$

and that

$$\sigma_S(x,y) = \sup_{u\geq 0, v\in\mathbb{R}} \langle (u,v),(x,y)\rangle = \begin{cases} 0 & \text{if } x \leq 0 \text{ and } y = 0, \\ +\infty & \text{otherwise.} \end{cases}$$

It is easy to see by the definition that

$$\operatorname{epi} f^* = \left\{(x,y,r) \in \mathbb{R}^3 : \frac{y^2}{2} \leq r\right\} \quad \text{and} \quad \operatorname{epi} \sigma_S = \{(x,0,r) \in \mathbb{R}^3 : x \leq 0,\ r \geq 0\}.$$

Therefore, $\operatorname{epi} f^* + \operatorname{epi} \sigma_S = \operatorname{epi} f^*$. This implies that $\operatorname{epi} f^* + \operatorname{epi} \sigma_S$ is weak*-closed. Since f is proper and $\operatorname{dom} f \cap S \neq \emptyset$, we see that $f \in \mathcal{F}_S$.

Now, note that $(0,1)$ is the minimizer of the following problem:

$$\begin{aligned} &\text{Minimize} && f(x,y), \\ &\text{s.t.} && \delta_S(x,y) \leq 0. \end{aligned}$$

Note also that $\partial f(0,1) = \{(x,1) : x \in \mathbb{R}\}$ and that $N_S(0,1) = \{\lambda(-1,0) : \lambda \geq 0\}$. Hence the optimality condition $(0,0) \in \partial f(0,1) + N_S(0,1)$ fails even though $f \in \mathcal{F}_S$ and the family $\{\delta_S\}$ has the BCQ property.

6. Applications to conic programming. We continue our study of the conic programming problem with notation as explained in Example 2.1. It can be checked in a straightforward manner that the following facts are true. They are known when g is continuous on X; see [24, 30]. Some related results can be founded in [17].

Fact 6.1. $\operatorname{cone} \bigcup_{\lambda\in K^\oplus} \operatorname{epi} (\lambda g)^* = \bigcup_{\lambda\in K^\oplus} \operatorname{epi} (\lambda g)^*$.

Fact 6.2. $\operatorname{cone} \bigcup_{\lambda\in K^\oplus} \partial(\lambda g)(x) = \bigcup_{\lambda\in K^\oplus} \partial(\lambda g)(x)$ for each $x \in X$.

Fact 6.3. $\operatorname{cone} \bigcup_{\substack{\lambda\in K^\oplus \\ \lambda g(x)=0}} \partial(\lambda g)(x) = \bigcup_{\substack{\lambda\in K^\oplus \\ \lambda g(x)=0}} \partial(\lambda g)(x) = Ng(x)_0$ for each $x \in X$,

where $Ng(x)_0$ is defined by

$$Ng(x)_0 := \left\{u^* \in X^* : (u^*, u^*(x)) \in \bigcup_{\lambda\in K^\oplus} \operatorname{epi} (\lambda g)^*\right\}.$$

Generalizing the corresponding notions in [24, 26, 29] to suit our present non-continuous situation, we make the following definitions (it is routine to see that the notions in the following definition coincide with corresponding ones in [24, 26, 29] in the case when $g : X \to Z$ is continuous).

DEFINITION 6.1. *For g, C, K as in Example* 2.1, *we say that*

(i) *the condition* (**C**$*$) *holds if the family* $\{\lambda g : \lambda \in K^{\oplus}\}$ *has the conical EHP;*

(ii) *the closed cone constraint qualification (CCCQ) holds if the family* $\{\delta_C, \lambda g : \lambda \in K^{\oplus}\}$ *has the conical EHP;*

(iii) *the pair* $\{C, g^{-1}(-K)\}$ *has the sharpened strong CHIP at* $x \in C \cap g^{-1}(-K)$ *if* $N_{C\cap g^{-1}(-K)}(x) = N_C(x) + Ng(x)_0$.

The following notion of K-lower semicontinuity was introduced in [37] and extended in [1, 12] for functions $g : X \to Z^{\bullet}$. It was also considered in [5].

DEFINITION 6.2. *For g, K as in Example* 2.1, *the function g is said to be K-lower semicontinuous at $x_0 \in X$ if for each neighborhood V of zero in Z and any $b \in Z$ with $b \leq_K g(x_0)$ there exists a neighborhood U of zero in X such that*

$$g(x_0 + U) \subseteq b + V + K \cup \{\infty\}. \tag{6.1}$$

Clearly, if $g : X \to Z$ is continuous, then g is K-convex lower semicontinuous. Below we give an example of a function that is K-convex lower semicontinuous but not continuous.

Example 6.1. Let $X = l^1$ and $Z = l^1$ respectively under the l^∞-norm $\|\cdot\|_\infty$ and the l^1-norm $\|\cdot\|_1$, and let g denote the identity map from X into Z. Then g has the desired properties (that g is not continuous is well known). To see this, let us fix a nonzero element $c \in X$, and let K denote its kernel in Z, that is, $K := \{z \in Z : \langle c, z\rangle = 0\}$. By a well-known result (cf. [39, p. 24]), the distance to each $z \in Z$ from the closed subspace K satisfies the so-called Ascoli formula,

$$d(z, K) = \frac{|\langle c, z\rangle|}{\|c\|_\infty} \quad \text{for each } z \in Z,$$

and it follows that

$$d(z, K) \leq \alpha\|z\|_\infty \quad \text{for each } z \in Z,$$

where $\alpha := \frac{\|c\|_1}{\|c\|_\infty}$; in particular we have

$$d(g(x), K) \leq \alpha\|g(x)\|_\infty \quad \text{for each } x \in X.$$

This implies that g is K-lower semicontinuous at $x_0 := 0$ (for $\epsilon > 0$ and $V = \{z \in Z : \|z\|_1 < \epsilon\}$, and (6.1) holds with $U = \{x \in X : \|x\|_\infty < \frac{\epsilon}{\alpha}\}$). By the linearity of g, we conclude that g is K-lower semicontinuous on the whole X.

PROPOSITION 6.1. *Let g, K be as in Example* 2.1. *Suppose that g is K-lower semicontinuous and that* $\operatorname{dom} g$ *is closed. Then for each $\lambda \in K^{\oplus}$, λg is lower semicontinuous.*

Proof. Let $\lambda \in K^{\oplus}\backslash\{0\}$, and let $x_0 \in X$. To show the lower semicontinuity of λg at x_0, we assume without loss of generality that $x_0 \in \operatorname{dom} g$ (thanks to the assumption that $\operatorname{dom} g$ is closed). Let $\epsilon > 0$. By the continuity of λ, take a neighborhood V of zero in Z such that $|\lambda(v)| < \epsilon$ for each $v \in V$. By definition of K-lower semicontinuity, there exists a neighborhood U of x_0 in X such that (6.1) holds. Let $u \in x_0 + U$.

By (6.1), there exist $v_1 \in V$ and $k \in K \cup \{\infty\}$ such that $g(u) = g(x_0) + v_1 + k$. For the case when $k \in K$, we have

$$\lambda g(x_0) = \langle \lambda, g(x_0)\rangle \leq \langle \lambda, g(u) - v_1\rangle < \lambda g(u) + \epsilon.$$

For the case when $k = \infty$, one has $g(u) = \infty$, i.e., $u \notin \operatorname{dom} g$. Thus, it follows from definition that $\lambda g(x_0) - \epsilon < \lambda g(u) = \infty$. Therefore, λg is lower semicontinuous at x_0. □

Example 6.2 below shows that the converse of Proposition 6.1 is not true.

Example 6.2. Let $X = L^2[0,1]$ and $Z = L^2[0,1]$ respectively under the $\|\cdot\|_1$-norm and $\|\cdot\|_2$-norm. Let $K = \{0\}$, and let $Z^\bullet := Z \cup \{+\infty\}$ as in Example 2.1. Let $D = \{x \in X : \ \|x\|_2 \leq 1\}$, and let $g(x) = x + \delta_D(x)$. Then g is not continuous on D (as there exists a sequence $\{x_n\}$ in X such that each $\|x_n\|_2 = 1$ but $\|x_n\|_1 = \frac{1}{n}$; for example, let $x_n = n\chi_{[0,1/n^2]}$ be the characteristic function of the interval $[0, 1/n^2]$), and so g is not $\{0\}$-lower semicontinuous.

Let $\lambda \in Z^* = \{0\}^\oplus$. We claim that λg is lower semicontinuous. Let $r \in \mathbb{R}$ and let

$$A_r := \{x \in X : \ \lambda g(x) \leq r\}.$$

It suffices to show that A_r is closed in X. To do this, let $x \in X$ and $\{x_n\}$ be a sequence in A_r such that $\|x_n - x\|_1 \to 0$. By an elementary result in Lebesgue theory (see [38, Proposition 18, p. 95]), there exists a subsequence $\{x_{n_k}\}$ convergent to x almost everywhere. By Fatou's lemma and the fact that $A_r \subseteq D$ (by (2.7)), it follows that

$$\int_0^1 |x|^2 dt \leq \liminf_k \int_0^1 |x_{n_k}|^2 dt \leq 1,$$

and thus $x \in D$. Let $\epsilon > 0$. Since $\lambda \in L^2[0,1]$, there exists a simple function h such that $\|\lambda - h\|_2 < \epsilon$. Noting $\|x - x_n\|_2 \leq 2$ (as $x, x_n \in D$), it follows from the Hölder inequality that

$$\begin{aligned}\lambda g(x) - \lambda g(x_n) = \langle \lambda, x - x_n\rangle &= \langle \lambda - h, x - x_n\rangle + \langle h, x - x_n\rangle \\ &\leq 2\epsilon + \|h\|_\infty \|x - x_n\|_1 \to 2\epsilon.\end{aligned}$$

This implies that

$$\lambda g(x) \leq \liminf_n \lambda g(x_n) + 2\epsilon \leq r + 2\epsilon,$$

and hence that $\lambda g(x) \leq r$ as $\epsilon > 0$ is arbitrary. Therefore $x \in A_r$ and A_r is closed, as we wished to show.

By Fact 6.1, Corollary 4.1, and Proposition 6.1, the following remark is obvious.

Remark 6.1. Let g, C, K be as in Example 2.1. Suppose in addition that g is K-lower semicontinuous and that $\operatorname{dom} g$ is closed. Then the condition (**C***) holds if and only if

$$\textstyle\bigcup_{\lambda \in K^\oplus} \operatorname{epi}(\lambda g)^* \quad \text{is } w^*\text{-closed};$$

more generally, the CCCQ holds if and only if

$$\operatorname{epi}\sigma_C + \textstyle\bigcup_{\lambda \in K^\oplus} \operatorname{epi}(\lambda g)^* \quad \text{is } w^*\text{-closed}.$$

(Thus, conditions (**C***) and CCCQ defined in Definition 6.1 agree with the ones defined in [26, 29] for the continuous case.)

COROLLARY 6.1. *The following equivalence holds for any* $x \in C \cap g^{-1}(-K)$*:*

$$\begin{aligned}&\text{The family } \{\delta_C, \lambda g :\ \lambda \in K^{\oplus}\} \text{ satisfies the BCQ at } x\\ &\quad\Leftrightarrow \{C, g^{-1}(-K)\} \text{ has the sharpened strong CHIP at } x.\end{aligned}$$

Proof. Note first that $\{x : \delta_C(x) \le 0, \lambda g(x) \le 0, \lambda \in K^{\oplus}\} = C \cap g^{-1}(-K)$. Hence the family $\{\delta_C, \lambda g :\ \lambda \in K^{\oplus}\}$ satisfies the BCQ at x if and only if

$$N_{C\cap g^{-1}(-K)}(x) = N_C(x) + \operatorname{cone} \bigcup_{\substack{\lambda\in K^{\oplus}\\ \lambda g(x)=0}} \partial(\lambda g)(x),$$

which is equivalent to saying that $\{C, g^{-1}(-K)\}$ has the sharpened strong CHIP at x, by Fact 6.3. □

The next two corollaries were respectively proved in [24, Propositions 3.3 and 3.4] under the additional assumption that g is continuous.

COROLLARY 6.2. *If* g, C, K *are as in Example* 2.1 *and the CCCQ holds, then*

$$N_{C\cap g^{-1}(-K)}(x) = N_C(x) + \bigcup_{\substack{\lambda\in K^{\oplus}\\ \lambda g(x)=0}} \partial\,(\lambda g)(x) \quad \text{for each } x \in C \cap g^{-1}(-K); \tag{6.2}$$

that is, $\{C, g^{-1}(-K)\}$ *has the sharpened strong CHIP at each point in* $C \cap g^{-1}(-K)$*. In particular, if the condition* (**C**$*$) *holds, then*

$$N_{g^{-1}(-K)}(x) = \bigcup_{\substack{\lambda\in K^{\oplus}\\ \lambda g(x)=0}} \partial\,(\lambda g)(x) \quad \text{for each } x \in g^{-1}(-K). \tag{6.3}$$

Proof. We need prove only the first assertion. Define $I := K^{\oplus} \cup \{i_0\}$, $i_0 \notin K^{\oplus}$, and consider $\{g_i :\ i \in I\}$ as defined in (2.8). Then $S := \{x :\ g_i(x) \le 0\}$ is exactly $C \cap g^{-1}(-K)$, and the active index set $I(x)$ is exactly $\{i_0\} \cup \{\lambda \in K^{\oplus} :\ \lambda g(x) = 0\}$. Thus (6.2) simply means that the family $\{g_i :\ i \in I\}$ has the BCQ, and hence the result follows from Remark 6.1 and Theorem 4.1. □

COROLLARY 6.3. *If* g, C, K *are as in Example* 2.1 *and the condition* (**C**$*$) *holds, then for each* $x \in C \cap g^{-1}(-K)$ *the family* $\{C, g^{-1}(-K)\}$ *satisfies the strong CHIP at* x *if and only if it satisfies the sharpened strong CHIP at* x*.*

Proof. By the given assumption, (6.3) holds; that is, $N_{g^{-1}(-K)}(x) = Ng(x)_0$ for each $x \in g^{-1}(-K)$. Consequently, the following equivalence holds for each $x \in C \cap g^{-1}(-K)$:

$$N_{C\cap g^{-1}(-K)}(x) = N_C(x) + N_{g^{-1}(-K)}(x) \Leftrightarrow N_{C\cap g^{-1}(-K)}(x) = N_C(x) + Ng(x)_0.$$

Thus the result is clear. □

7. Applications to best approximation theory. Let us recall from [14] that for a system of finitely many closed convex sets $\{D, C_i :\ i \in I\}$ in a Hilbert space, where $C_i = \{x \in X :\ \langle a_i, x\rangle \le b_i\}$ for some $a_i \in X$ and $b_i \in \mathbb{R}$, $i \in I$, the following statements are equivalent for each $x_0 \in D \cap \bigcap_{i\in I} C_i$:

(i) $\{D, C_i :\ i \in I\}$ has the strong CHIP at each x_0.

(ii) For each $x \in X$, $P_{D\cap\bigcap_{i\in I} C_i}(x) = x_0$ if and only if there exists a finite set $I_0 \subseteq I$ such that $P_D(x - \sum_{i\in I_0} a_i) = x_0$,

where $P_A(x)$ denotes the projection of the point x onto a convex set A. This important result has been extended in many aspects. For example, [32] discussed an extension to the case of an infinite system, and [31] discussed a family of functions in place of that of closed convex sets.

Recall that for a Banach space X and its dual X^*, the duality map $\Phi : X \rightrightarrows X^*$ is defined by $\Phi(x) := \{x^* : \|x\|^2 = \|x^*\|^2 = \langle x^*, x\rangle\}$ (cf. [40, section 3.7]). Let $\{g_i : i \in I\}$, X, C, S, and I be as in section 2.

THEOREM 7.1. *Suppose that X is a Banach space, and let $x_0 \in C \cap S$. Consider the following statements:*

(i) *$\{\delta_C; g_i, i \in I\}$ satisfies the BCQ at x_0.*

(ii) *For each $x \in X$, $x_0 \in P_{C\cap S}(x)$ if and only if*

$$\Phi(x - x_0) \cap \left(N_C(x_0) + \operatorname{cone} \bigcup_{i\in I(x_0)} \partial\, g_i(x_0) \right) \neq \emptyset. \tag{7.1}$$

Then (i)⇒(ii). *If we further assume that X is reflexive and smooth, then* (ii)⇔(i).

Proof. We regard the family $\{\delta_C; g_i, i \in I\}$ as $\{g_j : j \in J\}$ by letting $J = I \cup \{i_+\}$ and $g_j := \delta_C$, where $i_+ \notin I$. It follows that

$$J(x_0) := \left\{ j \in J : g_j(x_0) = \max\left\{ \sup_{i\in I} g_i(x_0), \delta_C(x_0) \right\} \right\} = \{i_+\} \cup I(x_0).$$

Suppose that (i) holds and let $x \in X$. Note that $x_0 \in P_{C\cap S}(x)$ if and only if x_0 minimizes the function $\frac{1}{2}\|\cdot - x\|^2$ over the set $\{x : \delta_C(x) = 0,\ g_i(x) \leq 0\ \forall i \in I\}$. Since $\Phi(x - x_0) = -\partial\,(\frac{1}{2}\|\cdot - x\|^2)(x_0)$ (cf. [40, p. 230]) and since the family $\{g_j : j \in J\}$ satisfies the BCQ at x_0, it follows from (2.3) and Theorem 5.1(iii) that $x_0 \in P_{C\cap S}(x)$ if and only if

$$\begin{aligned} &\Phi(x - x_0) \cap \left(N_C(x_0) + \operatorname{cone} \bigcup_{i\in I(x_0)} \partial\, g_i(x_0) \right) \\ &\quad = \Phi(x - x_0) \cap \left(\operatorname{cone} \bigcup_{j\in J(x_0)} \partial\, g_j(x_0) \right) \neq \emptyset. \end{aligned}$$

Thus (ii) holds.

Now we assume in addition that X is reflexive and smooth, and turn to proving (ii)⇒(i). By (3.9), we need to show that

$$N_{C\cap S}(x_0) \subseteq N_C(x_0) + \operatorname{cone} \bigcup_{i\in I(x_0)} \partial\, g_i(x_0). \tag{7.2}$$

To do this, take $y^* \in N_{C\cap S}(x_0)$. By the given assumptions, Φ is bijective (cf. [40, Theorem 3.7.2(vi) and p. 230]). Thus there exists $u = \Phi^{-1}(y^*)$, and it follows that $x_0 \in P_{C\cap S}(x_0 + u)$ by a well-known result (cf. [40, Corollary 3.8.5]). Therefore, by (ii), (7.1) holds with $x = x_0 + u$. Thus we obtain from (7.1) that

$$y^* = \Phi(x_0 + u - x_0) \in N_C(x_0) + \operatorname{cone} \bigcup_{i\in I(x_0)} \partial\, g_i(x_0).$$

This shows that (7.2) holds, as required. This completes the proof. ∎

COROLLARY 7.1. *Suppose that X in Theorem 7.1 is a Hilbert space, and let $x_0 \in C \cap S$. Then the following statements are equivalent:*

(i) *$\{\delta_C;\ g_i,\ i \in I\}$ satisfies the BCQ at x_0.*

(ii) *For each $x \in X$, $x_0 = P_{C\cap S}(x)$ if and only if*

$$x - x_0 \in N_C(x_0) + \operatorname{cone} \bigcup_{i\in I(x_0)} \partial\, g_i(x_0). \tag{7.3}$$

(iii) *For each $x \in X$, $x_0 = P_{C\cap S}(x)$ if and only if there exist finite subset $J \subseteq I(x_0)$, $\lambda_i \geq 0$, $u_i \in \partial\, g_i(x_0)$, $i \in J$, such that*

$$x_0 = P_C\left(x - \sum_{i\in J} \lambda_i u_i\right).$$

Proof. The equivalence of (ii) and (iii) is standard in Hilbert spaces. The equivalence of (i) and (ii) follows from Theorem 7.1, since (7.1) and (7.3) are now identical (because Φ is the identity map for Hilbert spaces) ∎

COROLLARY 7.2. *Suppose in Corollary 7.1 that the family $\{\delta_C;\ g_i,\ i \in I\}$ has the conical EHP. Then for each $x \in X$ and $x_0 \in C \cap S$, $P_{C\cap S}(x) = x_0$ if and only if there exist a finite set $I_0 \subseteq I(x_0)$, $x_i \in \partial\, g_i(x)$, and $\lambda_i \geq 0$ for each $i \in I_0$ such that $P_C(x - \sum_{i\in I_0} \lambda_i x_i) = x_0$.*

Proof. Since the family $\{\delta_C;\ g_i,\ i \in I\}$ has the conical EHP, it satisfies the BCQ by Theorem 4.1(ii). The result now follows from Corollary 7.1. ∎

For the next two corollaries, let g, C, and K be as in Example 2.1. These two corollaries were established, respectively, in [28] and [29] but under the additional assumption that g is continuous.

COROLLARY 7.3. *Suppose that X is a Hilbert space, and let $x_0 \in C \cap g^{-1}(-K)$. Then the following statements are equivalent:*

(i) *$\{\delta_C, \lambda g :\ \lambda \in K^{\oplus}\}$ satisfies the BCQ at x_0.*

(ii) *The pair $\{C, g^{-1}(-K)\}$ has the sharpened strong CHIP at x_0, that is,*

$$N_{C\cap g^{-1}(-K)}(x_0) = N_C(x_0) + \bigcup_{\substack{\lambda\in K^{\oplus}\\ \lambda g(x_0)=0}} \partial\,(\lambda g)(x_0).$$

(iii) *For each $x \in X$, $x_0 = P_{C\cap g^{-1}(-K)}(x)$ if and only if $x_0 = P_C(x - l)$ for some $l \in \bigcup_{\substack{\lambda\in K^{\oplus}\\ \lambda g(x_0)=0}} \partial\,(\lambda g)(x_0)$.*

Proof. By Fact 6.3 and Remark 6.1, (i)⇔(ii). The equivalence (i)⇔(iii) follows from Corollary 7.1 as applied to $I := K^{\oplus}$ and $g_\lambda := \lambda g$ (so $I(x_0) = \{\lambda \in K^{\oplus} : \lambda g(x_0) = 0\}$). ∎

COROLLARY 7.4. *Suppose that X is a Hilbert space and that $(\mathbf{C}*)$ holds. Let $x_0 \in C \cap g^{-1}(-K)$ and $x \in X$. Assume that $\{C, g^{-1}(-K)\}$ has the strong CHIP at x_0. Then the following statements are equivalent:*

(i) *$x_0 = P_{C\cap g^{-1}(-K)}(x)$.*

(ii) *$x_0 = P_C(x - l)$ for some $l \in \bigcup_{\substack{\lambda\in K^{\oplus}\\ \lambda g(x_0)=0}} \partial\,(\lambda g)(x_0)$.*

Proof. By Corollary 6.3, the given assumptions imply that $\{C, g^{-1}(-K)\}$ has the sharpened strong CHIP at x_0. By the implication (ii)⇒(iii) in the preceding corollary, the result is now clear. ∎

Acknowledgment. The authors would like to express their sincere thanks to the three anonymous referees for many helpful comments and for pointing out the references [1, 4, 5, 6, 12, 16, 17, 37].

REFERENCES

[1] M. Aït Mansoour, A. Metrane, and M. Théra, *Lower semicontinuous regularization for vector-valued mappings*, J. Global Optim., 35 (2006), pp. 283–309.
[2] V. Barbu and T. Precupanu, *Convexity and Optimization in Banach Spaces*, Sijthoff & Noordhoff, Alphen aan den Rijn, The Netherlands, 1978.
[3] J. Borwein and R. Goebel, *Notions of relative interior in Banach spaces, Optimization and related topics*, J. Math. Sci., 115 (2003), pp. 2542–2553.
[4] R. I. Boţ, S. M. Grad, and G. Wanka, *A weaker regularity condition for subdifferential calculus and Fenchel duality in infinite dimensional spaces*, Nonlinear Anal., 64 (2006), pp. 2787–2804.
[5] R. I. Boţ, S. M. Grad, and G. Wanka, *A new constraint qualification for the formula of the subdifferential of composed convex functions in infinite dimensional spaces*, Math. Nachr., to appear.
[6] R. I. Boţ, S. M. Grad, and G. Wanka, *On strong and total Lagrange duality for convex optimization problems*, J. Math. Anal. Appl., 337 (2008), pp. 1315–1325.
[7] R. I. Boţ and G. Wanka, *An alternative formulation for a new closed cone constraint qualification*, Nonlinear Anal., 64 (2006), pp. 1367–1381.
[8] R. S. Burachik and V. Jeyakumar, *A simple closure condition for the normal cone intersection formula*, Proc. Amer. Math. Soc., 133 (2005), pp. 1741–1748.
[9] R. S. Burachik and V. Jeyakumar, *A new geometric condition for Fenchel's duality in infinite dimensional spaces*, Math. Program., 104 (2005), pp. 229–233.
[10] R. S. Burachik and V. Jeyakumar, *A dual condition for the convex subdifferential sum formula with applications*, J. Convex Anal., 12 (2005), pp. 279–290.
[11] F. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley & Sons, New York, 1983.
[12] C. Combari, M. Laghdir, and L. Thibault, *Sous-différentiels de foncitions convexes composées*, Ann. Sci. Math. Québec, 18 (1994), pp. 119–148.
[13] F. Deutsch, W. Li, and J. Ward, *A dual approach to constrained interpolation from a convex subset of Hilbert space*, J. Approx. Theory, 90 (1997), pp. 385–414.
[14] F. Deutsch, W. Li, and J. D. Ward, *Best approximation from the intersection of a closed convex set and a polyhedron in Hilbert space, weak Slater conditions, and the strong conical hull intersection property*, SIAM J. Optim., 10 (1999), pp. 252–268.
[15] N. Dinh, M. A. Goberna, and M. A. López, *From linear to convex systems: Consistency, Farkas' lemma and applications*, J. Convex Anal., 13 (2006), pp. 113–133.
[16] N. Dinh, M. A. Goberna, M. A. López, and T. Q. Son, *New Farkas-type constraint qualification in convex semi-infinite programming*, ESAIM Control Optim. Calc. Var., 13 (2007), pp. 580–597.
[17] M. A. Goberna, V. Jeyakumar, and M. A. López, *Necessary and sufficient conditions for solvability of systems of infinite convex inequalities*, Nonlinear Anal., to appear; doi: 10.1016/j.na.2006.12.014.
[18] M. A. Goberna and M. A. López, *Linear Semi-Infinite Optimization*, Wiley Series in Mathematical Methods in Practice, Wiley, Chichester, UK, 1998.
[19] J. Hiriart-Urruty and C. Lemarechal, *Convex Analysis and Minimization Algorithms* I, Grundlehren der Math. Wiss. 305, Springer-Verlag, New York, 1993.
[20] R. B. Holmes, *Geometric Functional Analysis and Its Applications*, Springer-Verlag, New York, 1975.
[21] H. Hu, *Characterizations of local and global error bounds for convex inequalities in Banach spaces*, SIAM J. Optim., 18 (2007), pp. 309–321.
[22] J. Jahn, *Mathematical Vector Optimization in Partially Ordered Linear Spaces*, Methoden und Verfahren der mathematischen Physik, Band 31, Verlag Peter Lang, Frankfurt am Main Bern, New York, 1986.
[23] V. Jeyakumar, *Characterizing set containments involving infinite convex constraints and reverse-convex constraints*, SIAM J. Optim., 13 (2003), pp. 947–959.
[24] V. Jeyakumar, *The strong conical hull intersection property for convex programming*, Math. Program., 106 (2006), pp. 81–92.
[25] V. Jeyakumar, G. M. Lee, and N. Dinh, *New sequential Lagrange multiplier conditions characterizing optimality without constraint qualification for convex programs*, SIAM J. Optim., 14 (2003), pp. 534–547.

[26] V. Jeyakumar, N. Dinh, and G. M. Lee, *A New Closed Cone Constraint Qualification for Convex Optimization*, Applied Mathematics Research Report AMR 04/8, Department of Applied Mathematics, University of New South Wales, 2004.
[27] V. Jeyakumar, N. Dinh, and G. M. Lee, *Liberating the subgradient optimality conditions from constraint qualifications*, J. Global Optim., 36 (2006), pp. 127–137.
[28] V. Jeyakumar and H. Mohebi, *Limiting ϵ-subgradient characterizations of constrained best approximation*, J. Approx. Theory, 135 (2005), pp. 145–159.
[29] V. Jeyakumar and H. Mohebi, *A global approach to nonlinearly constrained best approximation*, Numer. Funct. Anal. Optim., 26 (2005), pp. 205–227.
[30] V. Jeyakumar, A. M. Rubinov, B. M. Glover, and Y. Ishizuka, *Inequality systems and global optimization*, J. Math. Anal. Appl., 202 (1996), pp. 900–919.
[31] C. Li and X.-Q. Jin, *Nonlinearly constrained best approximation in Hilbert spaces: The strong CHIP and the basic constraint qualification*, SIAM J. Optim., 13 (2002), pp. 228–239.
[32] C. Li and K. F. Ng, *Constraint qualification, the strong CHIP, and best approximation with convex constraints in Banach spaces*, SIAM J. Optim., 14 (2003), pp. 584–607.
[33] C. Li and K. F. Ng, *On constraint qualification for an infinite system of convex inequalities in a Banach space*, SIAM J. Optim., 15 (2005), pp. 488–512.
[34] C. Li and K. F. Ng, *Strong CHIP for infinite system of closed convex sets in normed linear spaces*, SIAM J. Optim., 16 (2005), pp. 311–340.
[35] C. Li, K. F. Ng, and T. K. Pong, *The SECQ, linear regularity, and the strong CHIP for an infinite system of closed convex sets in normed linear spaces*, SIAM J. Optim., 18 (2007), pp. 643–665.
[36] W. Li, C. Nahak, and I. Singer, *Constraint qualifications for semi-infinite systems of convex inequalities*, SIAM J. Optim., 11 (2000), pp. 31–52.
[37] J. P. Penot and M. Théra, *Semi-continuous mappings in general topology*, Arch. Math., 38 (1982), pp. 158–166.
[38] H. L. Royden, *Real Analysis*, 3rd ed., MacMillan, New York, 1988.
[39] I. Singer, *Best Approximation in Normed Linear Spaces by Elements of Linear Spaces*, Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen, Band 171, Springer-Verlag, Berlin, 1970.
[40] C. Zălinescu, *Convex Analysis in General Vector Spaces*, World Scientific, River Edge, NJ, 2002.

SIAM J. OPTIM.
Vol. 19, No. 1, pp. 188–210

MULTIOBJECTIVE OPTIMIZATION THROUGH A SERIES OF SINGLE-OBJECTIVE FORMULATIONS*

CHARLES AUDET†, GILLES SAVARD†, AND WALID ZGHAL‡

Abstract. This work deals with bound constrained multiobjective optimization (MOP) of nonsmooth functions for problems where the structure of the objective functions either cannot be exploited, or are absent. Typical situations arise when the functions are computed as the result of a computer simulation. We first present definitions and optimality conditions as well as two families of single-objective formulations of MOP. Next, we propose a new algorithm called BiMads for the biobjective optimization (BOP) problem (i.e., MOP with two objective functions). The property that Pareto points may be ordered in BOP and not in MOP is exploited by our algorithm. BiMads generates an approximation of the Pareto front by solving a series of single-objective formulations of BOP. These single-objective problems are solved using the recent Mads (mesh adaptive direct search) algorithm for nonsmooth optimization. The Pareto front approximation is shown to satisfy some first order necessary optimality conditions based on the Clarke calculus. Finally, BiMads is tested on problems from the literature designed to illustrate specific difficulties encountered in biobjective optimization, such as a nonconvex or disjoint Pareto front, local Pareto fronts, or a nonuniform Pareto front.

Key words. biobjective programming problem, multiobjective programming problem, mesh adaptive direct search algorithms (Mads), convergence analysis

AMS subject classifications. 90C26, 90C29, 90C30, 90C56

DOI. 10.1137/060677513

1. Introduction.

In many real-world problems, decisions depend on multiple and conflicting criteria. For example, in portfolio management, two criteria are usually considered: the return of the portfolio and its volatility risk [22]. There is usually no unique solution that is simultaneously optimal for all criteria, multiobjective optimization (MOP) aims at identifying the best trade-offs between these criteria. In this paper, we consider multiobjective programming under bound constraints, which may be stated as

$$\text{MOP}: \quad \min_{x \in X} \; F(x) = (f_1(x),\; f_2(x), \ldots,\; f_p(x))$$

with

$$F : \mathbb{R}^n \to \mathbb{R}^p \;\text{ and }\; X = \{x \in \mathbb{R}^n : a \le x \le b\},$$

where n is the number of variables, p is the number of objective functions, and $a \in (\mathbb{R} \cup \{-\infty\})^n$ and $b \in (\mathbb{R} \cup \{+\infty\})^n$ are bound vectors. Throughout this paper, X is assumed to be full dimensional, i.e., $a < b$ componentwise. The solution of MOP consists of the set of best trade-off points selected according to the Pareto dominance

*Received by the editors December 12, 2006; accepted for publication (in revised form) July 11, 2007; published electronically February 20, 2008.
http://www.siam.org/journals/siopt/19-1/67751.html

†GERAD, HEC Montréal, 3000, chemin de la côte Sainte-Catherine, Montréal (Québec), H3T 2A7 Canada. Département de Mathématique et de Génie Industriel, École Polytechnique de Montréal, C.P. 6079, succ. Centreville, Montréal (Québec), H3C 3A7 Canada (charles.audet@gerad.ca, gilles.savard@polymtl.ca).

‡Département de Mathématique et de Génie Industriel, École Polytechnique de Montréal, C.P. 6079, succ. Centreville, Montréal (Québec), H3C 3A7 Canada (walid.zghal@gerad.ca).

relation presented in section 2.1. The image under the mapping F of these points is called the *Pareto front*.

In this paper, we present an algorithm that generates an approximation of the Pareto front by only using function values. We are interested in problems in which the structure of the objective functions is either absent, unreliable, or cannot be exploited. A typical example is when the evaluation of F requires a computer simulation, i.e., the value $F(x)$ is only returned after a long series of computer operations. Some examples in single-objective optimization of such blackbox problems are detailed in [1, 5, 7, 19, 23].

The method proposed in this work is called BiMads as it is designed to approximate the Pareto front of a biobjective optimization problem (BOP) by using the recent Mads (mesh adaptive direct search) algorithm [4]. BiMads essentially solves a series of bound-constrained single-objective formulations of BOP using Mads with increasingly stringent stopping criteria. The series of formulations is constructed in a way to attempt a uniform coverage of the Pareto front, even in the case where the Pareto front is nonconvex or disjoint.

This paper is organized as follows. A brief overview of multiobjective optimization and some methods from the literature are presented in section 2. Some necessary optimality conditions that make use of the Clarke calculus [8] are also presented. Two classes of single-objective formulations are introduced in section 3: the *single-objective normalized formulation* and the *single-objective product formulation*. Then, the algorithm BiMads is detailed in section 4 for solving BOP. Finally, in section 5, BiMads is tested on problems from the literature [15] designed to underline specific difficulties encountered in biobjective optimization, such as a nonconvex or disjoint Pareto front, local Pareto fronts, or a nonuniform Pareto front. A new performance measure is proposed. It provides an indication of the coverage of the Pareto front by the set of points produced by the algorithm.

2. Multiobjective optimization. This section summarizes some key notions in multiobjective optimization such as Pareto dominance and solution approaches from the literature.

2.1. Pareto dominance. Comparison of optimal solutions requires an order relation, called dominance relation, between the different points [26]. Several dominance relations such as the Geoffrion's dominance [18] and the lexicographical dominance [16] have been proposed. The most commonly used relies on Pareto dominance shown below.

Definition 2.1. *Let $u, v \in X$ be two decision vectors. We define the following:*

- *$u \preceq v$ (u weakly dominates v) if and only if $f_i(u) \leq f_i(v)$ for all $i \in \{1, 2, \ldots, p\}$.*
- *$u \prec v$ (u dominates v) if and only if $u \preceq v$ and $f_j(u) < f_j(v)$ for at least one $j \in \{1, 2, \ldots, p\}$.*
- *$u \sim v$ (u is indifferent to v) if and only if u does not dominate v and v does not dominate u.*

Definition 2.1 is illustrated in Figure 2.1 for a biobjective problem in which $X \subset \mathbb{R}^3$. The feasible region X is projected on the objective space; the image of X under the mapping F is denoted by $Y \subseteq \mathbb{R}^p$, and is delimited by the curve in the right part of the figure.

Figure 2.1 also highlights three zones in the objective space, relative to the feasible point $x_1 \in X$. The *dominated zone* is the set of points which are dominated by x_1. The *dominance zone* is the set of points that dominate x_1. The *indifference zone* is

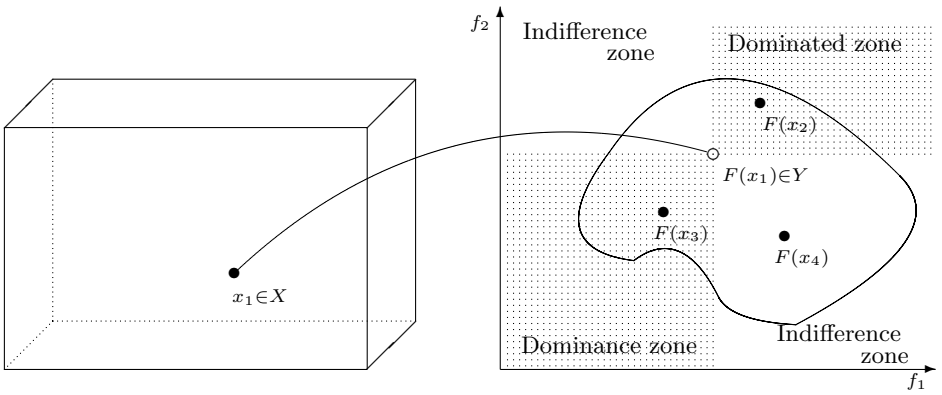


FIG. 2.1. *Pareto dominance illustrated on a biobjective problem with* 3 *variables.*

the set of points which are indifferent to x_1. In Figure 2.1, x_1 dominates x_2 but is dominated by x_3 and indifferent to x_4, i.e., $x_3 \prec x_1 \prec x_2$ and $x_1 \sim x_4$. Furthermore, x_4 is also indifferent to both x_2 and x_3, i.e., $x_2 \sim x_4 \sim x_3$.

The above dominance relation provides an optimality definition for multiobjective programming. We distinguish between local and global Pareto optimality.

DEFINITION 2.2. *A point $x^* \in X$ is said to be globally Pareto optimal if and only if there is no $x \in X$ such that $x \prec x^*$. If x^* is globally Pareto optimal, then $F(x^*)$ is called globally efficient.*

Some computational methods do not guarantee global optimality but ensure at best local optimality, as shown below.

DEFINITION 2.3. *A point $\tilde{x} \in X$ is said to be locally Pareto optimal if and only if there exits some $\epsilon > 0$ and $\sigma > 0$ for which the set $\{x \in B_\epsilon(\tilde{x}) \cap X \ : \ x \prec \tilde{x}$ and $F(x) \in B_\sigma(F(\tilde{x}))\}$ is empty, where $B_\epsilon(\tilde{x})$ denotes an open ball around $\tilde{x}$ of radius ϵ and $B_\sigma(F(\tilde{x}))$ denotes an open ball around $F(\tilde{x})$ of radius σ. If $\tilde{x}$ is locally Pareto optimal, then $F(\tilde{x})$ is called locally efficient.*

The two notions are illustrated in Figure 2.2. The point $\tilde{x} \in X$ is locally Pareto optimal while all points that map to the boundary of Y in the shaded area are globally Pareto optimal. Pareto optimality will henceforth refer to global Pareto optimality unless noted otherwise. The set of globally Pareto optimal points is called globally Pareto optimal set and denoted by $X_{\mathcal{P}}$. The image under the mapping F of $X_{\mathcal{P}}$ defines the solution set of the multiobjective problem. This set is called the *global Pareto front* or simply the *Pareto front* and is denoted by $Y_{\mathcal{P}} \in \mathbb{R}^p$. The image of a set of locally Pareto optimal points is called a *local Pareto front*.

Figure 2.2 also depicts two important points in the objective space. The *ideal point* $\ell \in (\mathbb{R} \cup -\infty)^p$ is defined as the vector whose components are the individual minima of each objective function

$$\ell \ = \ \left(\min_{x \in X} f_1(x), \ \min_{x \in X} f_2(x), \ldots, \ \min_{x \in X} f_p(x) \right)^T .$$

The *nadir point* $u \in (\mathbb{R} \cup \infty)^p$ is defined as the vector whose components are the

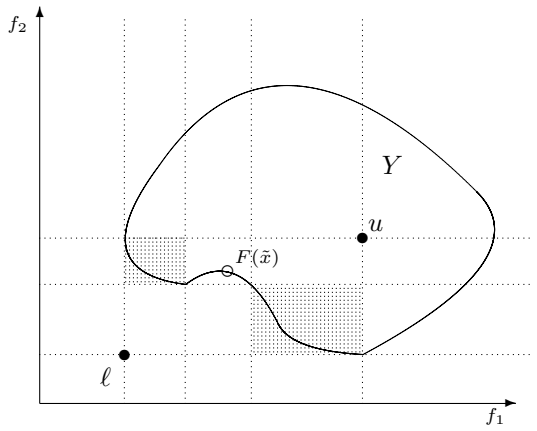


FIG. 2.2. *Pareto front and ideal and nadir points of a biobjective problem.*

individual maxima in the Pareto front of each objective function

$$u = \left(\max_{x \in X_{\mathcal{P}}} f_1(x), \max_{x \in X_{\mathcal{P}}} f_2(x)^T, \ldots, \max_{x \in X_{\mathcal{P}}} f_p(x) \right)^T.$$

2.2. Classes of methods for MOP. Several methods have been proposed for multiobjective optimization. There are some exact methods for linear multiobjective optimization [6, 2]. In the nonlinear case, several heuristic methods generate a set of points $Y_{\mathcal{L}}$ that gives an approximation $Y_{\mathcal{P}}$ of the Pareto front. A survey of heuristic approaches may be found in [17].

Two natural strategies arise when developing heuristics for MOP [28]. One strategy consists of making sure that the algorithm generates trial points whose images rapidly converge to some point of the Pareto front $Y_{\mathcal{P}}$. The second strategy is to spread out these approximations of Pareto points so that they achieve a well-distributed nondominated set $Y_{\mathcal{L}}$.

We next discuss different classes of methods for MOP. A first class of methods consists of reformulating MOP into a series of single-objective programs by aggregating the objective functions. Three methods of this class are presented as follows: A linear weighting method, an approximation to some reference point, and a weighted geometric mean approach. We also briefly discuss the normal-boundary intersection algorithm (NBI) [14].

Linear weighting method. The linear weighting method (LWM) [9] consists of converting the MOP into a single-objective optimization problem by minimizing a convex combination of objectives

$$\text{LWM}: \quad \min_{x \in X} \sum_{i=1}^{p} w_i f_i(x),$$

where $w_i \geq 0$ for $i = 1, 2, \ldots, p$ are weights such that $\sum_{i=1}^{p} w_i = 1$. Any optimal solution of LWM is Pareto optimal for MOP. Hence, solving LC for different weight combinations produces a subset of Pareto solutions. A well-known difficulty of this method is that it cannot generate any point in the nonconvex part of the Pareto front $Y_{\mathcal{P}}$ as illustrated in Figure 2.3. No value of the weights w_1 and w_2 can lead an algorithm on the optimization problem LWM to any point near the efficient point $F(x^*)$,

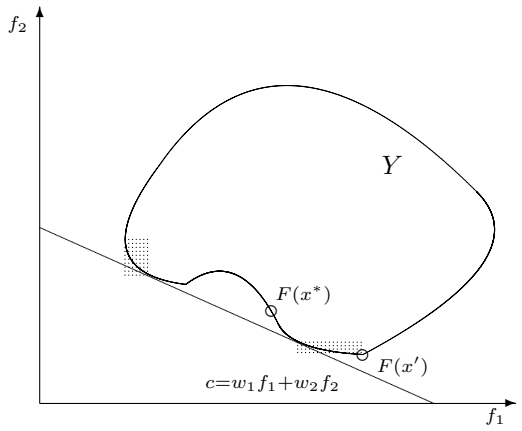


FIG. 2.3. *The linear weighting method may only generate a subset of the Pareto front.*

and the only points of the Pareto front that can be generated are those belonging to the shaded areas in Figure 2.3. These shaded areas are significantly smaller than those of Figure 2.2, delimiting the Pareto front.

A second difficulty with the linear weighting method is that the same point may be found by several weight combinations, as illustrated by $F(x')$ in Figure 2.3.

These two difficulties are such that the main drawback with the linear weighting method is the lack of uniformity of the Pareto front approximation distribution.

Approximation to a reference point. This method consists of finding a feasible solution $x \in X$ such that $F(x)$ is close to some reference point $r \in \mathbb{R}^p$ [26, 27]. Often, the ideal point defined in section 2.1 is used as the reference point, i.e., $r = \ell$. The problem is formulated as follows:

$$PI_r: \quad \min_{x \in X} \ \|F(x) - r\|_q \ = \ \left(\sum_{i=1}^{p} |f_i(x) - r_i|^q \right)^{1/q},$$

where $\|.\|_q$ is the q-norm with $1 \leq q \leq \infty$. The method using $\|.\|_2$ norm is illustrated in Figure 2.4 for a biobjective problem. Solving PI_r with different reference points produces a set of points which approximates the Pareto front. In Figure 2.4, the method applied with ℓ as a reference point generates an efficient point $F(x^*)$. As opposed to the weighting method, the approximation to the reference point may generate points in the nonconvex part of the Pareto front. However, it may generate nonefficient points, as illustrated in Figure 2.4 by the generation of $F(x)$ from the reference point r.

Weighted geometric mean and NBI approaches. The weighted geometric mean approach consists of converting the MOP to a single-objective optimization problem by maximizing the weighted geometric mean of differences between the components of the nadir point u defined in section 2.1 and the objective functions

$$\text{WGMP}: \quad \begin{array}{lll} \max\limits_{x} & \prod\limits_{i=1}^{p} (u_i - f_i(x))^{\lambda_i} & \\ \text{s.t.} & f_i(x) \leq u_i, & \text{for } i = 1, 2, \ldots, p, \\ & x \in X, & \end{array}$$

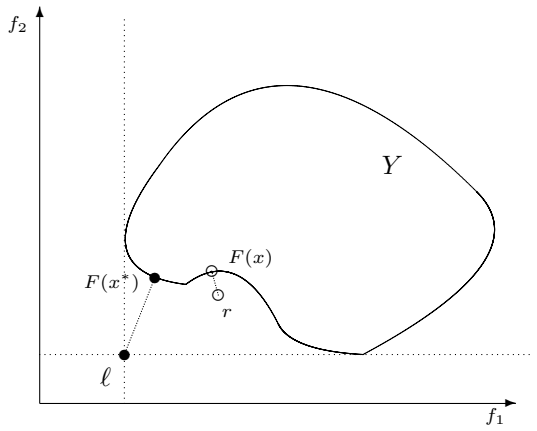


FIG. 2.4. *The approximation to a reference point may lead to a nonefficient point.*

where $\lambda_i > 0$ for $i = 1, 2, \ldots, p$. Lootsma, Athan, and Papalambros [21] show that if all objective functions $f_1, f_2, \ldots, f_p$ are convex, then a solution of WGMP is Pareto optimal. A difficulty with this approach is that the resulting problem WGMP now contains p general constraints instead of only bound constraints. This approach has some similarities to what we propose in section 3, but we do not add any supplementary constraints.

The NBI approach of Das [13] produces an approximation of the Pareto front by solving a series of single-objective optimization problems NBI_β, in which an additional equality constraints tying the objective function values are added:

$$\text{NBI}_\beta: \quad \begin{array}{ll} \max\limits_{x,t} & t \\ \text{s.t.} & \phi\beta + t\hat{n} = \quad F(x) \\ & x \in X, \end{array}$$

where β are barycentric coordinates, $\phi\beta$ represents a point in the convex hull of individual minima (CHIM), and $\hat{n}$ denotes the unit vector normal to the CHIM simplex, pointing towards origin. Solution of NBI_β is the intersection of the normal to CHIM and the boundary of Y closest to the origin. Solving NBI_β for various values of β gives an approximation of the Pareto front. This approach may be impracticable in the blackbox optimization context.

2.3. Necessary optimality conditions for multiobjective optimization. This section contains necessary optimality conditions for multiobjective optimization that extend the first order optimality condition for single-objective optimization. The convergence analysis makes use of the Clarke calculus [8] for nonsmooth functions. For a locally Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$, Clarke [8] defines the *generalized directional derivative* of f evaluated at $\tilde{x} \in X$ in the tangent direction $v \in \mathbb{R}^n$ as

$$f^\circ(\tilde{x}; v) = \limsup_{y \to \tilde{x},\ t \downarrow 0} \frac{f(y + tv) - f(y)}{t}. \tag{2.1}$$

A point $\tilde{x} \in X$ is said to be a Clarke stationary point of f over the bound constraints domain X if the generalized derivative is nonnegative for any direction in the tangent cone $T_X(\tilde{x})$, to X at $\tilde{x}$. The *generalized gradient* is defined to be the set $\partial f(\hat{x}) = \{s \in \mathbb{R}^n : f^\circ(\hat{x}; v) \geq v^T s \text{ for all } v \in \mathbb{R}^n\}$.

The following theorem presents a necessary condition for Pareto optimality of (MOP) analogue to optimality conditions for set-valued functions [11] rewritten with our notation.

THEOREM 2.4. *Let f_i be Lipschitz near $\tilde{x} \in X$, $i = 1, 2, \ldots, p$. If $\tilde{x}$ is locally Pareto optimal, then for any d in the tangent cone $T_X(\tilde{x})$ there exists $\hat{\jmath} \in \{1, 2, \ldots, p\}$ for which*

$$f_{\hat{\jmath}}^{\circ}(\tilde{x}; d) \geq 0.$$

Assuming strict differentiability [20] of each objective functions f_i at $\tilde{x} \in X$, for $i = 1, 2, \ldots, p$ leads to the following corollary [17]. Strict differentiability of f_i at $\tilde{x}$ is just the requirement that the generalized gradient is a singleton, i.e., that $\partial f_i(\tilde{x}) = \{\nabla f(\tilde{x})\}$ in addition to the requirement that f_i is Lipschitz near $\tilde{x}$. F is called strictly differentiable at $\tilde{x} \in X$ if f_i are strictly differentiable $\tilde{x}$ for $i = 1, 2, \ldots, p$.

COROLLARY 2.5. *Let F be strictly differentiable at $\tilde{x} \in X$. If $\tilde{x}$ is locally Pareto optimal, then for any d in the tangent cone $T_X(\tilde{x})$ there exists some $\hat{\jmath} \in \{1, 2, \ldots, p\}$ such that $\nabla f_{\hat{\jmath}}(\tilde{x})^T d \geq 0$. The point $\tilde{x}$ is called a* KKT-properly efficient solution for MOP.

3. Single-objective formulations of multiobjective optimization. We propose to solve MOP through a series of single-objective optimization problems. Each single-objective optimization problem relies on a reference point r in the objective space $\mathbb{R}^p$ and must satisfy the requirements presented in the following definition.

DEFINITION 3.1. *Consider the single-objective optimization problem:*

$$R_r \ : \ \min_{x \in X} \psi_r(x) \quad \textit{with} \quad \psi_r(x) = \phi_r(f_1(x), f_2(x), \ldots, f_p(x)),$$

where $\phi_r : \mathbb{R}^p \to \mathbb{R}$ is parameterized with respect to some reference point $r \in \mathbb{R}^p$. Then R_r is called a single-objective formulation at r of MOP *if the following conditions hold:*

- *If F is Lipschitz near some $\tilde{x} \in X$, then ψ_r is also Lipschitz near $\tilde{x} \in X$.*
- *If F is Lipschitz near some $\tilde{x} \in X$ with $F(\tilde{x}) < r$ componentwise, and if $d \in T_X(\tilde{x})$ is such that $f_i^{\circ}(\tilde{x}; d) < 0$ for $i = 1, 2, \ldots, p$, then $\psi_r^{\circ}(\tilde{x}; d) < 0$.*

The first condition ensures that the formulation preserves local Lipschitz continuity while the second involves Clarke descent directions for all f_i's and ψ_r. Assuming more smoothness on the function ϕ_r leads to the following theorem.

THEOREM 3.2. *Let R_r be a single-objective formulation at $r \in \mathbb{R}^p$ of MOP. If F and ψ_r are strictly differentiable at some $\tilde{x} \in X$ with $F(\tilde{x}) < r$ componentwise, and if $d \in T_X(\tilde{x})$ is such that $\nabla f_i(\tilde{x})^T d < 0$ for $i = 1, 2, \ldots, p$, then $\nabla \psi_r(\tilde{x})^T d < 0$.*

Proof. Strict differentiability of ϕ_r and F at $\tilde{x}$ ensures strict differentiability of ψ_r at $\tilde{x}$ with $\nabla \psi_r(\tilde{x})^T d = \psi_r^{\circ}(\tilde{x}; d)$. It follows from the second condition of Definition 3.1 that $\nabla \phi_r(\tilde{x})^T d < 0$ if $\nabla f_i(\tilde{x})^T d = f_i^{\circ}(\tilde{x}; d) < 0$ for $i = 1, 2, \ldots, p$. ∎

The next two subsections introduce two single-objective formulations: the *single-objective normalized formulation* and the *single-objective product formulation*. Both formulations are similar to the weighted geometric mean approach described in section 2.2 but have the advantage of not introducing nonlinear constraints to the bound-constrained domain X.

3.1. Single-objective normalized formulation. Let $r \in \mathbb{R}^p$ be a reference point in the objective space and $s \in \mathbb{R}^p$ be a positive scaling factor. The *single-objective normalized formulation* is defined as

$$\hat{R}_r \ : \quad \min_{x \in X} \hat{\psi}_r \quad \text{with} \quad \hat{\psi}_r(x) = \hat{\phi}_r(f_1(x), f_2(x), \ldots, f_p(x)) = \max_{i \in \{1,2,\ldots,p\}} \frac{f_i(x) - r_i}{s_i}.$$

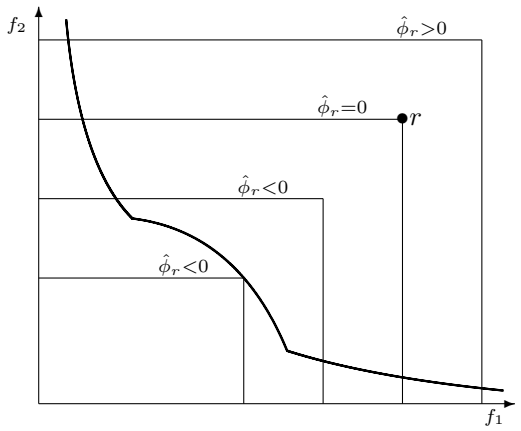


FIG. 3.1. *Level sets in objective space of the single-objective normalized formulation $\hat{R}_r$ for a BOP.*

Level sets of the function $\hat{\phi}_r$ in the objective function space are represented by horizontal and vertical lines in Figure 3.1 for a BOP.

The next theorem shows that $\hat{R}_r$ is a single-objective formulation of MOP.

THEOREM 3.3. *$\hat{R}_r$ is a single-objective formulation at r of MOP in the sense of Definition* 3.1.

Proof. Let F be Lipschitz near $x \in X$. Now, choose $j \in \operatorname{argmax}_{i\in\{1,2,\ldots,p\}} \frac{f_i(x)-r_i}{s_i}$ and for some $y \in X$, let $k \in \operatorname{argmax}_{i\in\{1,2,\ldots,p\}} \frac{f_i(y)-r_i}{s_i}$. It follows that

$$|\hat{\psi}_r(x) - \hat{\psi}_r(y)| = \begin{cases} \frac{f_j(x)-r_j}{s_j} - \frac{f_k(y)-r_k}{s_k} \le \frac{f_j(x)-r_j}{s_j} - \frac{f_j(y)-r_j}{s_j} = \frac{f_j(x)-f_j(y)}{s_j} \\ \text{or} \\ \frac{f_k(y)-r_k}{s_k} - \frac{f_j(x)-r_j}{s_j} \le \frac{f_k(y)-r_k}{s_k} - \frac{f_k(x)-r_k}{s_k} = \frac{f_k(y)-f_k(x)}{s_k}. \end{cases}$$

Hence, $|\hat{\psi}_r(x) - \hat{\psi}_r(y)| \le \max_{i\in\{1,2,\ldots,p\}} \frac{|f_i(x)-f_i(y)|}{s_i} \le \frac{\max(\lambda_1,\lambda_2,\ldots,\lambda_p)}{\min(s_1,s_2,\ldots,s_p)}\|x-y\|$, where λ_i denotes the Lipschitz constant of f_i for $i = 1, 2, \ldots, p$. The second condition of Definition 3.1 follows directly from [8, Proposition 2.3.12]. ∎

Based on the $\hat{R}_r$ formulation, the next proposition gives a necessary condition for a Pareto optimal solution for MOP.

PROPOSITION 3.4. *If there exists some vectors $r \in \mathbb{R}^p$ and $s \in \mathbb{R}^p_+$ such that $\tilde{x}$ is the unique optimal solution of $\hat{R}_r$, then $\tilde{x}$ is Pareto optimal for MOP.*

Proof. Let $\tilde{x}$ the unique optimal solution of $\hat{R}_r$ and $x \in X$, $x \neq \tilde{x}$. Then $\hat{\psi}(\tilde{x}) < \hat{\psi}(x)$ and, consequently, there exists some index $j \in \{1, 2, \ldots, p\}$ for which $\frac{f_j(\tilde{x})-r_j}{s_j} < \frac{f_j(x)-r_j}{s_j}$. It follows that $f_j(\tilde{x}) < f_j(x)$ and thus x does not dominate $\tilde{x}$. Hence, $\tilde{x}$ is Pareto optimal. ∎

The same argument ensures that a unique local optimal solution of $(\hat{R}_r)$ is locally Pareto optimal.

3.2. Single-objective product formulation. Let $r \in \mathbb{R}^p$ be a reference point in the objective space. The *single-objective product formulation* is defined as

$$\tilde{R}_r \quad : \quad \min_{x\in X} \tilde{\psi}_r \quad \text{with} \quad \hat{\psi}_r(x) = \tilde{\phi}_r(f_1(x), f_2(x), \ldots, f_p(x)) = -\prod_{i=1}^{p}((r_i - f_i(x))_+)^2,$$

where $(r_i - f_i(x))_+ = \max\{r_i - f_i(x), 0\}$ for $i = 1, 2, \ldots, p$. An advantage of $\tilde{R}_r$ over $\hat{R}_r$ is that the function of p variables ϕ_r is continuously differentiable in the entire space, and therefore, the formulation preserves the differentiability of the original problem. More precisely, if $F(x)$ is continuously differentiable near $\tilde{x} \in X$, then $\tilde{\psi}_r$ will also be continuously differentiable near $\tilde{x} \in X$; and if $F(x)$ is strictly differentiable near $\tilde{x} \in X$, then $\tilde{\psi}_r$ will also be strictly differentiable near $\tilde{x} \in X$.

$\tilde{R}_r$ is obtained from WGMP presented in section 2.2 with $\lambda_k = 2$ for $k = 1, 2, \ldots, p$ by treating the additional constraints introduced by WGMP in the objective function. The level sets of the $\tilde{R}_r$ formulation for a biobjective problem are depicted by thin curves in Figure 3.2. The level set $\tilde{R}_r = 0$ consists of the entire shaded region.

The next theorem shows that $\tilde{R}_r$ is a single-objective formulation of MOP.

THEOREM 3.5. *$\tilde{R}_r$ is a single-objective formulation at r of MOP in the sense of Definition* 3.1.

Proof. Let F be Lipschitz near $x \in X$. $\tilde{\psi}_r$ is Lipschitz near x as product of Lipschitz functions is also Lipschitz.

To prove the second condition of Definition 3.1, we first compute the generalized directional derivative of $\tilde{\psi}_r$ evaluated at $\tilde{x} \in X$ satisfying $F(\tilde{x}) < r$, componentwise, in some tangent direction $d \in T_X(\tilde{x})$. In order to simplify the presentation, let us define the function $c_i(x) = r_i - f_i(x)$ for $i = 1, 2, \ldots, p$. It follows from [8, Proposition 2.3.13] that

$$\begin{aligned}\tilde{\psi}_r^\circ(\tilde{x}; d) &\le -2\sum_{i=1}^{p}\Big((c_i(\tilde{x}; d)_+)^\circ (c_i(\tilde{x})_+)\prod_{j\neq i}(c_j(\tilde{x})_+)^2\Big)\\ &= -2\sum_{i=1}^{p}\Big(c_i^\circ(\tilde{x}; d)(c_i(\tilde{x})_+)\prod_{j\neq i}(c_j(\tilde{x})_+)^2\Big).\end{aligned}$$

Since, $c_i^\circ(\tilde{x}; d) = -f_i^\circ(\tilde{x}; d)$, we have

$$\tilde{\psi}_r^\circ(\tilde{x}; d) \;\le\; \sum_{i=1}^{p}\Big(f_i^\circ(\tilde{x}; d)((r_i - f_i(\tilde{x}))_+)\prod_{j\neq i}((r_j - f_j(\tilde{x}))_+)^2\Big).$$

It follows that if $f_i^\circ(\tilde{x}; d) < 0$ for $i = 1, 2, \ldots, p$, then $\tilde{\psi}_r^\circ(\tilde{x}; d) < 0$. ∎

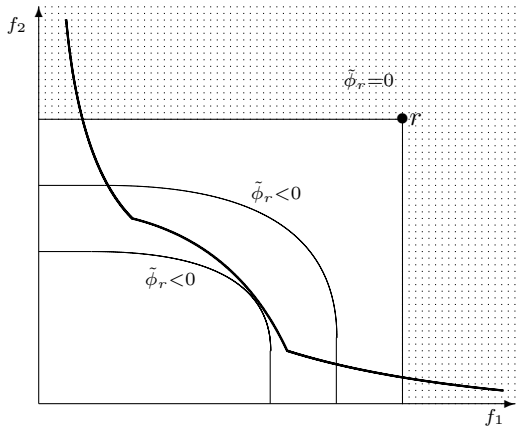


FIG. 3.2. *Level sets in objective space of the single-objective product formulation $\tilde{R}_r$ for a BOP.*

Based on the $\tilde{R}_r$ formulation, the next proposition shows that an optimal solution of $\tilde{R}_r$ with nonzero value is Pareto optimal.

PROPOSITION 3.6. *If there exists a vector $r \in \mathbb{R}^p$ such that $\tilde{x}$ is an optimal solution of $\tilde{R}_r$ with $\tilde{\psi}_r(\tilde{x}) < 0$, then $\tilde{x}$ is Pareto optimal for MOP.*

Proof. Assume that $x \in X$ satisfies $x \prec \tilde{x}$. Therefore, $f_i(x) \leq f_i(\tilde{x})$ for $i \in \{1, 2, \ldots, p\}$ with at least one strict inequality. It follows that $\tilde{\psi}_r(x) < \tilde{\psi}_r(\tilde{x}) < 0$, which contradicts the optimality of $\tilde{x}$ for the $\tilde{R}_r$ formulation. □

The same argument can be given to prove that a local optimal of $\tilde{R}_r$ with nonzero value is locally Pareto optimal.

4. A new algorithm for biobjective programming. This section focuses on the biobjective optimization problem BOP. BOP is the only instance of MOP that possesses an ordering property of the Pareto front. Based on this ordering property and on the aggregation approach presented in section 3, a new algorithm BIMADS is proposed to solve BOP. BIMADS essentially generates a sequence of single-objective formulations of BOP and solves them sequentially using the MADS algorithm [4] with increasingly stringent stopping criteria.

This section is divided as follows. An overview of the MADS algorithm is given in section 4.1 and BIMADS is presented in section 4.2. Convergence and uniformity analysis of BIMADS are proposed in section 4.3.

4.1. The MADS algorithm for single-objective optimization. MADS [4] is a direct search method for the minimization of a nonsmooth function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ under general constraints $x \in \Omega \neq \emptyset \subseteq \mathbb{R}^n$. MADS is a generalization of Torczon's [25] pattern search algorithms. Pattern search algorithms rely on a fixed finite set of directions to explore the space of variables and the convergence analysis [3] is confined to these directions. MADS overcomes this limitation by allowing an infinite set of directions. We next summarize the main steps of MADS. The following definitions are from [4]. (The reader is invited to consult [4] for the specific details of this method.)

MADS is an iterative algorithm that attempts at each iteration to improve the current incumbent value (i.e., the best feasible objective function value found so far) by evaluating f on some trial points that lie on the *current mesh*.

DEFINITION 4.1. *At iteration k, the current mesh is defined to be the following union:*

$$M_k = \bigcup_{x \in S_k} \{x + \Delta_k^m D z : z \in \mathbb{N}^{n_D}\},$$

where S_k is the set of points where the objective function f had been evaluated by the start of iteration k.

The mesh is constructed from a finite set of n_D directions $D \subset \mathbb{R}^n$ scaled by a positive *mesh size parameter* $\Delta_k^m \in \mathbb{R}_+$. Each iteration of MADS consists of two steps. The first, called the SEARCH step, allows evaluation of f at any finite number of feasible mesh points to get eventually a better incumbent. Then the second step, called the POLL, is invoked. The POLL step consists of a local exploration of the space of optimization variables. The set of trial points considered during the POLL is called MADS *frame* and is denoted by P_k. P_k is constructed using a current incumbent solution x_k (called the *frame center*) and the poll and mesh size parameters Δ_k^p and Δ_k^m to obtain a positive spanning set of directions D_k. A formal definition of MADS *frame* P_k is given below.

DEFINITION 4.2. *At iteration k, the* MADS *frame is defined to be the set*

$$P_k = \{x_k + \Delta_k^m d : d \in D_k\} \subset M_k,$$

where D_k is a positive spanning set such that $0 \notin D_k$ and for each $d \in D_k$,

- *d can be written as a nonnegative integer combination of the directions in D: $d = Du$ for some vector $u \in \mathbb{N}^{n_{D_k}}$ that may depend on the iteration number k,*
- *the distance from the frame center x_k to a frame point $x_k + \Delta_k^m d \in P_k$ is bounded above by a constant times the poll size parameter: $\Delta_k^m \|d\| \leq \Delta_k^p \max\{\|d'\| : d' \in D\}$,*
- *limits (as defined in Coope and Price [10]) of the normalized sets $\mathcal{D}_k = \left\{\frac{d}{\|d\|} : d \in D_k\right\}$ are positive spanning sets.*

If the value of f at a trial feasible point x is less than the current incumbent value, then x is called an *improved mesh point*, and the iteration is called a *successful iteration*. The next iteration will be initiated with the newly found incumbent solution and with a mesh size parameter equal to or larger than the previous one. When the iteration fails in generating an improved mesh point, the mesh size parameter is reduced to increase the mesh resolution in order to allow the evaluation of f at trial points closer to the incumbent solution. The MADS algorithm is stated formally below.

A GENERAL MADS ALGORITHM

- INITIALIZATION: Let $x_0 \in \Omega$, $\Delta_0^m \leq \Delta_0^p$, D satisfy the requirements given in [4]. Set the iteration counter $k \leftarrow 0$.
- SEARCH AND POLL STEP: Perform the SEARCH and possibly the POLL steps (or only part of them) until an improved mesh point x_{k+1} is found on the mesh M_k (see Definition 4.1).
  - OPTIONAL SEARCH: Evaluate f on a finite subset of trial points on the mesh M_k.
  - LOCAL POLL: Evaluate f on the frame P_k (see Definition 4.2).
- PARAMETER UPDATE: Update Δ_{k+1}^m according to [4]. Set $k \leftarrow k + 1$ and go back to the SEARCH and POLL step.

The MADS convergence analysis ensures that some optimality conditions hold at a limit point $\hat{x}$ which is the limit of mesh local optimizers on meshes that get infinitely fine. A hierarchy of convergence results based on local smoothness of f and Ω near $\hat{x}$ is developed in [4]. In particular, if f is Lipschitz near $\hat{x}$, then $f^\circ(\hat{x}; d) \geq 0$ for every $d \in T_\Omega^H(\hat{x})$, where $T_\Omega^H(\hat{x})$ denotes the hypertangent cone [24] to Ω at $\hat{x}$.

Our new algorithm uses MADS as a tool to solve single-objective blackbox optimization problems.

4.2. The BIMADS algorithm for biobjective optimization. The general scheme of BIMADS is presented in Figure 4.1, and is more fully described in the following paragraphs. BIMADS is an iterative algorithm that constructs sets of points that approximate the Pareto optimal set $X_\mathcal{P}$. At each iteration, the set of nondominated points (with respect to all points generated so far) is denoted by $X_\mathcal{L}$. The image under the mapping F of $X_\mathcal{L}$ is denoted by $Y_\mathcal{L} \in \mathbb{R}^p$. $Y_\mathcal{L}$ gives an approximation of the Pareto front $Y_\mathcal{P}$.

At the initialization step, the algorithm solves the two single-objective problems:

$$\min_{x \in X} f_1(x) \quad \text{and} \quad \min_{x \in X} f_2(x) \tag{4.1}$$

INITIALIZATION:

- Apply the MADS algorithm from x_0 to solve $\min_{x\in X} f_1(x)$ and $\min_{x\in X} f_2(x)$.
- Let $X_{\mathcal{L}} = \{x^1, x^2, \ldots, x^J\}$ be an ordered list of pairwise nondominated points such that $f_1(x^1) < f_1(x^2) < \cdots < f_1(x^J)$ and $f_2(x^1) > f_2(x^2) > \cdots > f_2(x^J)$.
  Initialize the weight $w(x) = 0$ for all $x \in X$ and, let $\delta > 0$.

MAIN ITERATIONS: Repeat

- REFERENCE POINT DETERMINATION:
  – If $J > 2$, let $\hat{\jmath} \in \operatorname{argmax}_{j=2,\ldots,J-1} \delta^j = \frac{||F(x^j)-F(x^{j-1})||^2+||F(x^j)-F(x^{j+1})||^2}{w(x^j)+1}$, and define the reference point $r = (f_1(x^{\hat{\jmath}+1}), f_2(x^{\hat{\jmath}-1}))$.
  – If $J = 2$, let $x^{\hat{\jmath}} = x^2$, define the reference point $r = (f_1(x^2), f_2(x^1))$, and set $\delta^{\hat{\jmath}} = \frac{||F(x^2)-F(x^1)||^2}{w(x^2)+1}$.
  – If $J = 1$, let $x^{\hat{\jmath}} = x^1$, $\delta^{\hat{\jmath}} = \frac{\delta}{w(x^{\hat{\jmath}})+1}$ and apply the MADS algorithm from $x^{\hat{\jmath}}$ to solve $\min_{x\in X} f_1(x)$ and $\min_{x\in X} f_2(x)$. Terminate MADS when the mesh size parameter Δ^m drops below $\Delta(\delta^{\hat{\jmath}}) = O(\delta^{\hat{\jmath}})$ and continue to the step UPDATE $X_{\mathcal{L}}$.
- SINGLE-OBJECTIVE FORMULATION MINIMIZATION:
  Solve a single-objective formulation R_r using the MADS algorithm from starting point $x^{\hat{\jmath}}$. Terminate MADS when the mesh size parameter Δ^m drops below $\Delta(\delta^{\hat{\jmath}}) = O(\delta^{\hat{\jmath}})$ or if a maximal number of objective evaluations is attained.
- UPDATE $X_{\mathcal{L}}$:
  Add to $X_{\mathcal{L}}$ all nondominated points found in the current iteration, remove dominated points from $X_{\mathcal{L}}$, and order the resulting list of points.
  Increase weights: $w(x^{\hat{\jmath}}) \leftarrow w(x^{\hat{\jmath}}) + 1$ for each $x \in X_{\mathcal{L}}$.

FIG. 4.1. *Scheme of the* BIMADS *algorithm for the biobjective programming.*

using the MADS algorithm from a user-defined starting point $x_0 \in X$. A first list $X_{\mathcal{L}}$ of nondominated points is obtained from the set of all trial mesh points generated by the two runs of MADS. The cardinality of this set is denoted by J. The sets $X_{\mathcal{L}}$ and $Y_{\mathcal{L}}$ are sorted in ascending order of f_1 value. The ordering property gives a simple way to measure the gaps between nondominated points by evaluating Euclidean distances between successive solutions in $Y_{\mathcal{L}}$. This strategy allows the evaluation of the solutions coverage along $Y_{\mathcal{L}}$ in order to determine a reference point. Note that solving problems with more than two objective functions requires the use of other techniques to measure the coverage in the absence of ordering property. Alternatively, a recursive application of BIMADS could be considered. In future work, we aim at studying the different alternatives and selecting the best one. Each iteration of BIMADS consists of three steps. First, the ordered list $Y_{\mathcal{L}}$ is used to identify a reference point r in the space of objectives. If $J > 2$, the strategy considers the weighted sum δ^j of squared distances from each nondominated point $F(x_j) \in Y_{\mathcal{L}}$ to its predecessor $F(x_{j-1})$ and successor $F(x_{j+1})$ for $j = 2, 3, \ldots, J-1$. Hence, a point $F(x^{\hat{\jmath}})$ is identified from the list $Y_{\mathcal{L}}$ that maximizes the measure δ^j. If $J = 2$, $x^{\hat{\jmath}}$ is set to x^2, the reference point r is set to $(f_1(x^2), f_2(x^1))$, and $\delta^{\hat{\jmath}}$ is set to be equal to the weighted squared distance between the two nondominated points. The weights are updated in a way to reduce frequent definitions of r around the same point by increasing $w(x^j)$ by one at the end

of the iteration. If $J = 1$, i.e., if a single point x^1 dominates all others generated so far, then the algorithm solves again the two single-objective problems (4.1).

Two measures of the uniformity distribution are introduced:

- the *coverage measure* $\tilde{c} = \max_{j=1,\ldots,J} \delta^j$ and
- the *weighted coverage measure* $\tilde{c}_w = \max_{j=1,\ldots,J} \frac{\delta^j}{w(x^j)+1}$.

A small value of the coverage measure indicates that the Pareto front approximation does not contain large gaps.

The left part of Figure 4.2 illustrates the reference point r selection when $J > 2$. The symbol "$\cdot$" is used for points generated by BiMads in the objective space, the symbol "$*$" is used for nondominated points found by the algorithm.

The second step of the iteration consists of solving the single-objective formulation R_r using the Mads algorithm. The image of the trial points produced by this algorithm will most likely lie in the dominance zone with respect to r. The right part of Figure 4.2 illustrates some trial points generated by Mads. Each run of Mads terminates when the mesh size parameter Δ^m drops below $\Delta(\delta^j) = O(\delta^j)$ or if a maximal number of objective evaluations is attained.

Finally, at the end of each iteration the set of nondominated points $X_{\mathcal{L}}$ is updated. New nondominated points are added and dominated ones are removed. The new set is represented in Figure 4.2 by stars.

These three steps are iterated by BiMads. In practice, termination is either set to a fixed number of iterations, or when δ^j drops below a predetermined threshold. Observe that other single-objective optimization algorithms could be applied for solving R_r. Nevertheless, we choose Mads to solve the single-objective formulation since it is designed for blackbox optimization problems, which corresponds to the practical problems that we plan to solve in future works and is still providing rigorous convergence analysis for these blackbox problems.

4.3. Convergence analysis. In this section, the quality of points produced by BiMads is studied and convergence results for BiMads are presented. We will make the standard assumptions of blackbox optimization that all trial points generated by the algorithm lie in a bounded set. The following theorem shows that BiMads produces points satisfying the necessary optimality conditions for biobjective opti-

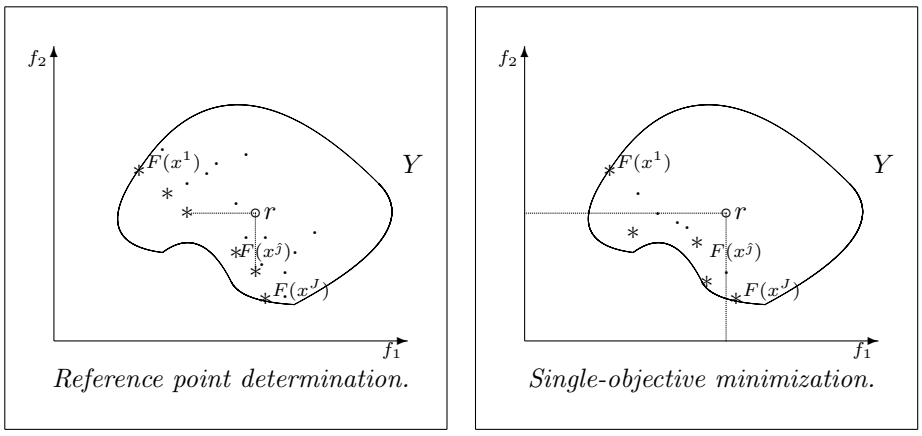


FIG. 4.2. *An iteration of* BiMads.

mization presented in section 2.3 when the objective functions are Lipschitz. We denote by R_r a single-objective formulation of BOP at some reference point $r \in \mathbb{R}^p$, as defined in Definition 3.1.

THEOREM 4.3. *Let f_1 and f_2 be Lipschitz near a limit point $\hat{x} \in X$ generated by* MADS *applied to a single-objective formulation R_r of BOP at some reference point $r \in \mathbb{R}^p$, then*

$$\text{for any } d \in T_X(\hat{x}), \text{there exists } j \in \{1,2\} \text{ such that } f_j^\circ(\hat{x};d) \geq 0,$$

where $T_X(\hat{x})$ is the tangent cone at $\hat{x}$.

Proof. The solution $\hat{x}$ produced by MADS is a stationary point of R_r on X [4], i.e., for all $d \in T_X(\hat{x}), R_r^\circ(\hat{x},d) \geq 0$. The second condition appearing in the definition of single-objective formulation in subsection 3.1 ensures that for any $d \in T_X(\hat{x})$, there exists an index $j \in \{1,2\}$ for which $f_j^\circ(\hat{x};d) \geq 0$. ∎

A corollary to this result is derived when f_1 and f_2 are strictly differentiable at $\hat{x}$.

COROLLARY 4.4. *Let f_1, f_2, and ψ_r be strictly differentiable at a limit $\hat{x} \in X$ generated by* MADS *applied to a single-objective formulation R_r of BOP at some reference point $r \in \mathbb{R}^p$, then $\hat{x}$ is KKT-properly efficient solution of BOP:*

$$\text{for any } d \in T_X(\hat{x}), \text{there exists } j \in \{1,2\} \text{ such that } \nabla f_j(\hat{x})^T d \geq 0,$$

where $T_X(\hat{x})$ is the tangent cone at $\hat{x}$.

Proof. Let $d \in T_X(\hat{x})$. According to [4], $\hat{x}$ is a KKT stationary point of R_r on X. Hence, $\nabla \psi_r(\hat{x})^T d \geq 0$. The contraposition of Theorem 3.2 ensures that there exists $j \in \{1,2\}$ such that $\nabla f_j(\hat{x})^T d \geq 0$. ∎

In addition to convergence analysis, an analysis of uniformity of solutions distribution in $Y_{\mathcal{L}}$ may be derived. The uniformity analysis makes use of the *weighted coverage measure* $\tilde{c}_w$ introduced in section 4.2. We introduce the index η to represent the BIMADS iteration number. The next theorem shows that the $\tilde{c}_w^\eta$ goes to zero with an infinite number of MADS runs.

THEOREM 4.5. *The weighted coverage measure $\tilde{c}_w^\eta$ of the pairwise nondominated points list $Y_{\mathcal{L}}^\eta$ found at* MADS *run η satisfies $\lim_{\eta\to\infty} \tilde{c}_w^\eta = 0$.*

Proof. By contradiction, suppose that there exists some scalar $L > 0$ for which $c_w^\eta > L$ for all iterations $\eta > 0$. Therefore, we get $\delta^{\hat{\jmath}^\eta} \geq \frac{\delta^{\hat{\jmath}^\eta}}{w(\hat{\jmath}^\eta)+1} > L$ for all iterations $\eta > 0$. Furthermore, since each run η of MADS terminates when the mesh size parameter Δ^m drops below $\Delta(\delta^{\hat{\jmath}^\eta}) = O(\delta^{\hat{\jmath}^\eta})$, it follows that

$$\text{there exists } \underline{\delta} > 0 : \Delta_k^\eta \geq \underline{\delta} \text{ for all iterations } k \text{ and for all runs } \eta, \tag{4.2}$$

where Δ_k^η is the mesh size at iteration k for the MADS run number η. Moreover, since the feasible set X is bounded, there exists $\bar{\delta} > 0$ such that $\Delta_k^\eta \leq \bar{\delta}$ for all $\eta > 0$ and $k > 0$. Let x_0^η be the starting point of run η. We want to show that under the aforementioned assumptions, all trial points lie on a mesh which is independent of the run and iteration numbers.

Consider $x_{k_0}^{\eta_0}$ the k_0th trial point generated in run η_0. According to [4], we have

$$x_{k_0}^{\eta_0} = x_0^{\eta_0} + D \sum_{i=0}^{k_0-1} \Delta_i^{\eta_0} z_i^{\eta_0},$$

where D is the set of $n_D \in \mathbb{N}$ directions satisfying the requirements in [4], $z_i^{\eta_0} \in \mathbb{N}^{n_D}$, and $\Delta_i^{\eta_0}$ is the mesh size at iteration i of the MADS run η_0. The starting point $x_0^{\eta_0}$ was generated as η_1th trial point of a prior run $0 \le \eta_1 < \eta_0$. Indeed, from $x_{k_0}^{\eta_0}$, we can construct the finite series of points

$$x_{k_0}^{\eta_0}, x_0^{\eta_0} = x_{k_1}^{\eta_1}, x_0^{\eta_1} = x_{k_2}^{\eta_2}, \ldots, x_0^{\eta_p} = x_{k_{p+1}}^{0}, x_0^0,$$

such that $\eta_0 > \eta_1 > \cdots > \eta_p > 0$ are decreasing integers and $x_{k_l}^{\eta_l}$ was generated at run η_l from the starting point $x_0^{\eta_l}$. This starting point was generated as the η_{l+1}th trial point of run η_{l+1}, i.e., $x_0^{\eta_l} = x_{k_{l+1}}^{\eta_{l+1}}$. $x_0^{\eta_p}$ was generated as the k_{p+1}th trial point of the first run. With this notation, we get that for $l = 1, 2, \ldots, p$,

$$x_{k_l}^{\eta_l} = x_0^{\eta_l} + D \sum_{i=0}^{k_l-1} \Delta_i^{\eta_l} z_i^{\eta_l},$$

where $z_i^{\eta_l} \in \mathbb{N}$. Therefore, we get

$$\begin{aligned}
x_{k_0}^{\eta_0} &= x_0^{\eta_0} + D \sum_{i=0}^{k_0-1} \Delta_i^{\eta_0} z_i^{\eta_0} \\
&= x_{k_1}^{\eta_1} + D \sum_{i=0}^{k_0-1} \Delta_i^{\eta_0} z_i^{\eta_0} \\
&= x_0^{\eta_1} + D \left(\sum_{i=0}^{k_0-1} \Delta_i^{\eta_0} z_i^{\eta_0} + \sum_{i=0}^{k_1-1} \Delta_i^{\eta_1} z_i^{\eta_1} \right) \\
&= x_{k_2}^{\eta_2} + D \left(\sum_{i=0}^{k_0-1} \Delta_i^{\eta_0} z_i^{\eta_0} + \sum_{i=0}^{k_1-1} \Delta_i^{\eta_1} z_i^{\eta_1} \right) \\
&\quad \cdots \\
&= x_0^0 + D \left(\sum_{l=0}^{p+1} \sum_{i=0}^{k_l-1} \Delta_i^{\eta_l} z_i^{\eta_l} \right).
\end{aligned}$$

Torczon's [25] showed that if $\Delta_i^{\eta_l}$ are multiple of integer powers of some rational number (which is the case in the present algorithm since Δ_0^{η} are identical for all runs η), and if all iterates lie in a bounded set, and if (4.2) holds, then $x_{k_0}^{\eta_0}$ belongs to a mesh that depends on $D, \underline{\delta}, \bar{\delta}$. Thus Torczon's proof shows that $x_{k_0}^{\eta_0}$ belongs to a mesh that is independent of the iteration and run numbers. Consequently, there exists a point selected infinitely many times by BIMADS around which to refine $Y_{\mathcal{L}}^{\eta}$ without varying $\delta^{\hat{\jmath}^{\eta}}$ value. Hence, $w(\hat{\jmath}^{\eta}) \to \infty$ and thus $c_w^{\eta} \to 0$. This contradicts the assumption that c_w^{η} is bounded below by $L > 0$. ∎

5. Numerical results on test problems. The behavior of BIMADS is evaluated using test problems from Deb [15]. The functions of these biobjective problems are constructed in such a way that

$$\begin{aligned}
&f_1(x) = f_1(x_1, x_2, \ldots, x_m) \quad \text{and} \\
&f_2(x) = g(x_{m+1}, x_{m+2}, \ldots, x_n) h(f_1(x_1, x_2, \ldots, x_m), g(x_{m+1}, x_{m+2}, \ldots, x_n)).
\end{aligned}$$

The function f_1 depends on m variables $x_1, x_2, \ldots, x_m$ with $m < n$ and f_2 is a function of n variables. The function g depends on $n - m$ variables $x_{m+1}, x_{m+2}, \ldots, x_n$ which do not appear in the function f_1; h is a function of the function values f_1 and g.

Difficulties in test problems are introduced by choosing appropriate functions f_1, g, and h [15]:

- Convexity and discontinuity in the Pareto front is handled by the function h.
- Convergence to the true Pareto front is handled by the function g.
- Diversity in the Pareto front is handled by the function f_1.

The test problems that we consider are defined through specific choices of f_1, g, and h.

The algorithm is coded in C++ and uses the NOMAD 0.6 [12] implementation of MADS as a subroutine to solve each single-objective formulation. The point $x_0 = (0.51, 0.51)$ is used as a starting point to initialize the algorithm, and the initial mesh is set to 0.01. Two termination criteria are selected for each run of MADS on the single-objective formulations:

- Poll size termination: the run ends when the mesh size parameter Δ^m drops below $\frac{\delta^j}{1000}$, where δ^j is defined in Figure 4.1.
- Truth evaluations termination: the run ends after 50 evaluations of the reformulated objective function unless indicated explicitly.

In our numerical experiments, BIMADS generates either 10 or 30 single-objective formulations, and therefore, calls MADS 10 or 30 times. This strategy generates at most 500 and 1500 evaluations of objective function, respectively.

For each test problem, a figure containing four graphs is presented: two of them for each single-objective formulations, and two of them for the runs involving 10 or 30 calls to MADS.

Each graph represents the function values of all trial points generated by BIMADS using the "$\cdot$" symbol. The symbol "$*$" is used for nondominated points found by the algorithm. The curves represent the global and local Pareto fronts, or the boundary of the image Y of X. For each series of experiments, the quality of the approximation $Y_{\mathcal{L}}$ provided by the algorithm BIMADS is discussed according to the two key issues presented in section 2: convergence of solutions to the Pareto curve $Y_{\mathcal{P}}$ and uniformity of the solutions distribution in $Y_{\mathcal{L}}$. The coverage measure $\tilde{c}$ introduced in (4.2) is displayed for each test problems.

5.1. Convex and nonconvex of Pareto fronts. Convexity of Pareto front is affected by the function h. Deb [15] proposes the following function:

$$h(f_1, g) = \begin{cases} 1 - (\frac{f_1}{g})^\alpha & \text{if } f_1 \leq g, \\ 0 & \text{otherwise.} \end{cases} \tag{5.1}$$

The parameter α controls the shape of the Pareto front. The Pareto front is nonconvex if $\alpha > 1$ and convex otherwise. Note that the global Pareto solutions are obtained at the global minimum of the function g. We consider the following functions for g and f_1:

$$g(x_2) = \begin{cases} 4 - 3\exp\left(-\dfrac{x_2 - 0.2}{0.02}\right)^2 & \text{if } 0 \leq x_2 \leq 0.4, \\ 4 - 2\exp\left(-\dfrac{x_2 - 0.7}{0.2}\right)^2 & \text{if } 0.4 \leq x_2 \leq 1, \end{cases}$$

$$f_1(x_1) = 4x_1,$$

where both x_1 and $x_2 \in [0, 1]$.

Three series of tests are generated with different values of α. In the first test, the global Pareto front is convex. In the second one, the global Pareto front is nonconvex. Finally, in the third test, the global Pareto front is convex but the local Pareto front is nonconvex.

Convex global Pareto front. By setting α to 0.25 in problem (5.1), we obtain a biobjective problem with the convex global Pareto front. Results obtained by applying BiMads with 10 runs of Mads and 30 runs of Mads are illustrated in Figure 5.1. The figure shows that all nondominated points $Y_{\mathcal{L}}$ found by BiMads belong to the global Pareto curve $Y_{\mathcal{P}}$ for all series of test. Furthermore, the figure suggests that BiMads achieves a well distributed and well spread nondominated points in the global Pareto front for each formulation.

Nonconvex global Pareto front. By setting α to 4 in problem (5.1), we obtain a biobjective problem with nonconvex global Pareto front. Results obtained by applying BiMads with 10 runs of Mads and 30 runs of Mads are shown in Figure 5.2. Except for the first test corresponding to results of BiMads with 10 runs of Mads applied to single-objective normalized formulation, all tests generate nondominated solutions $Y_{\mathcal{L}}$ in the global Pareto front $Y_{\mathcal{P}}$. A well spread Pareto solutions is obtained using 30 runs of Mads.

Nonconvex local optimal front and convex global front. We set $\alpha = 0.25 + 3.75(g(x_2) - 1)$ [15] in problem (5.1). The resulting problem is hard to solve for algorithms that exploit the shape of the Pareto curve [15]. Results obtained by applying BiMads with 10 runs of Mads and 30 runs of Mads are shown in Figure 5.3. Using 10 runs of Mads, BiMads generates a set of nondominated points $Y_{\mathcal{L}}$ close to the Pareto curve $Y_{\mathcal{P}}$. The approximation quality is better using the single-objective product formulation. Using 30 runs of Mads, all nondominated points found by

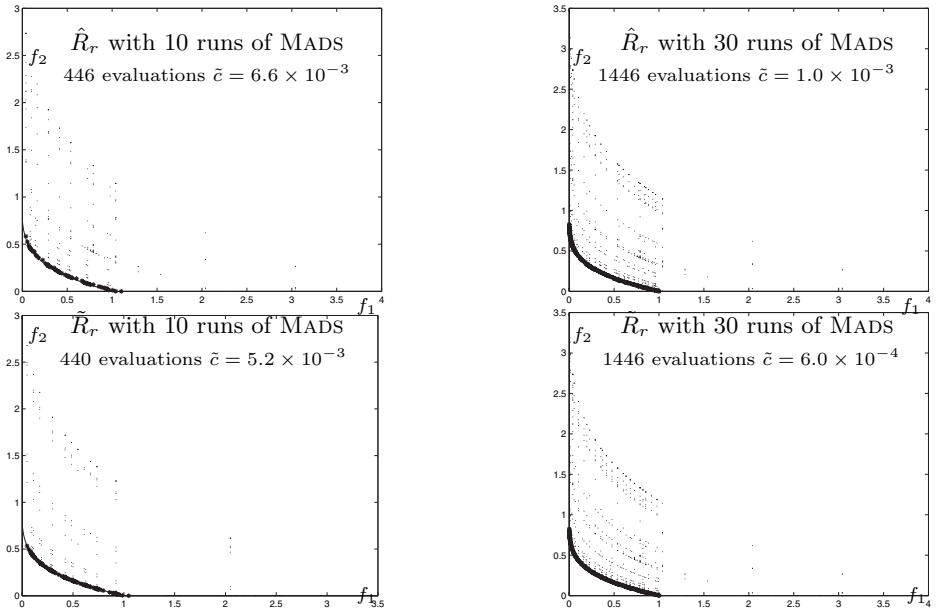


Fig. 5.1. *Convex global Pareto front:* BiMads *with* 10 *and* 30 *optimization runs of* Mads.

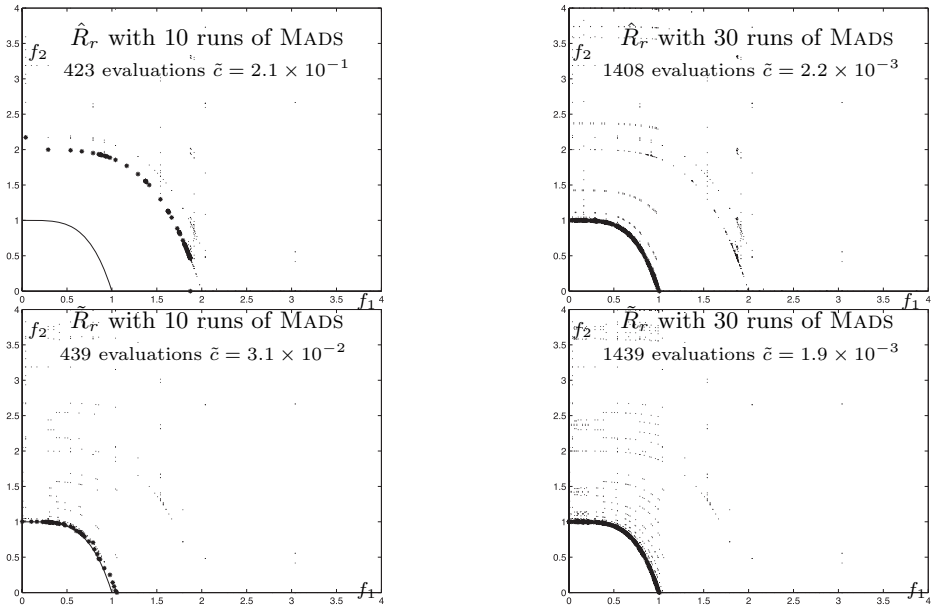


FIG. 5.2. *Nonconvex global Pareto front:* BIMADS *with* 10 *and* 30 *optimization runs of* MADS.

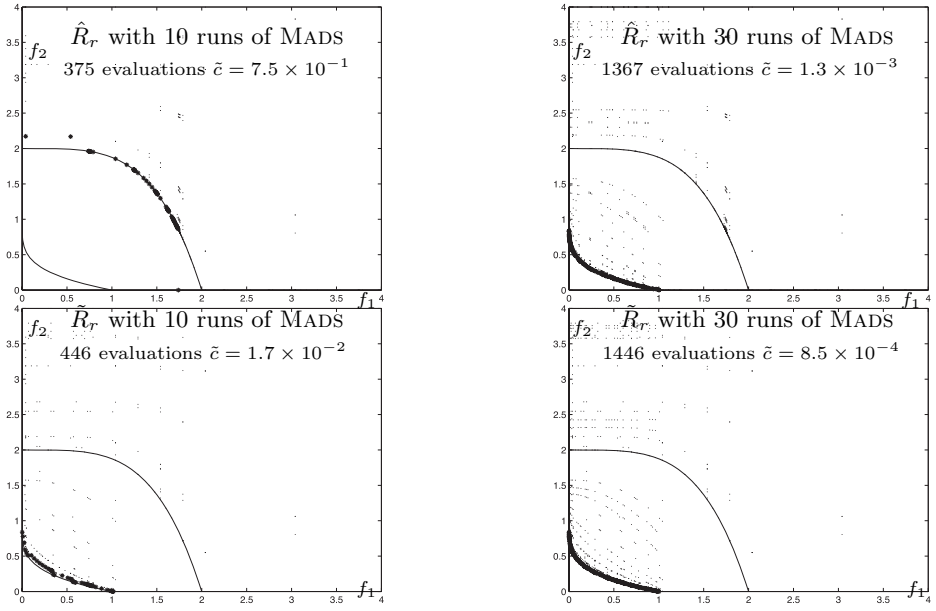


FIG. 5.3. *Nonconvex local optimal front and convex global front:* BIMADS *with* 10 *and* 30 *optimization runs of* MADS.

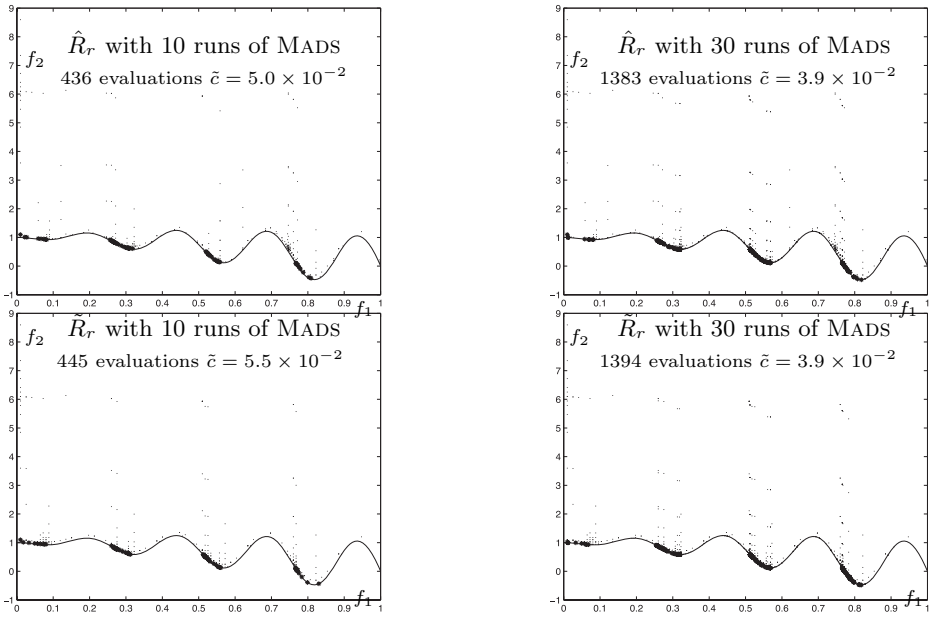


FIG. 5.4. *Discontinuous global front:* BIMADS *with* 10 *and* 30 *optimization runs of* MADS.

BIMADS lie on the global Pareto curve, i.e., $Y_{\mathcal{L}} \subset Y_{\mathcal{P}}$. The nondominated points $Y_{\mathcal{L}}$ are well spread among the Pareto curve.

5.2. Discontinuous Pareto front. Deb [15] proposes a family of test problems in which the Pareto front is discontinuous:

$$\begin{aligned} h(f_1, g) &= 1 - \left(\frac{f_1}{g}\right)^{\alpha} - \tfrac{f_1}{g}\sin(2\pi q f_1), \\ f_1(x_1) &= x_1, \\ g(x_2) &= 1 + 10x_2, \end{aligned}$$

where x_1, $x_2 \in [0, 1]$, $\alpha > 0$, and q is the number of discontinuous regions. We use the same values for α and q as in [15], i.e., $\alpha = 2$ and $q = 4$. Results obtained by applying BIMADS with 10 and 30 runs of MADS are shown in Figure 5.4.

Figure 5.4 shows that BIMADS generates nondominated solutions $Y_{\mathcal{L}}$ in all four discontinuous Pareto regions. The solutions $Y_{\mathcal{L}}$ are well distributed among the Pareto curve $Y_{\mathcal{P}}$. Table 5.1 displays the coverage measure $\tilde{c}$ for each of the four disjoint regions. The table also reports the mean value μ and standard deviation σ of $\tilde{c}$ over these regions. Moreover, using 30 optimization runs of MADS, results obtained by applying $\tilde{R}_r$ formulation are slightly better than those obtained by applying $\hat{R}_r$ formulation.

5.3. Biased search space. In order to make the convergence to the Pareto front more problematic, Deb [15] proposes the following function g:

$$g(x_2) = 1 + x_2^{\gamma},$$

TABLE 5.1
Coverage measure $\tilde{c}$ *of each discontinuous Pareto region.*

$\tilde{c}$	$\hat{R}_r$ formulation					
Region	1	2	3	4	μ	σ
10 runs	.0078	.0039	.0098	.0433	.0162	.0182
30 runs	.0100	.0054	.0040	.0025	.0054	.0032
$\tilde{c}$	$\tilde{R}_r$ formulation					
Region	1	2	3	4	μ	σ
10 runs	.0076	.0033	.0091	.0552	.0188	.0243
30 runs	.0012	.0030	.0029	.0059	.0032	.0019

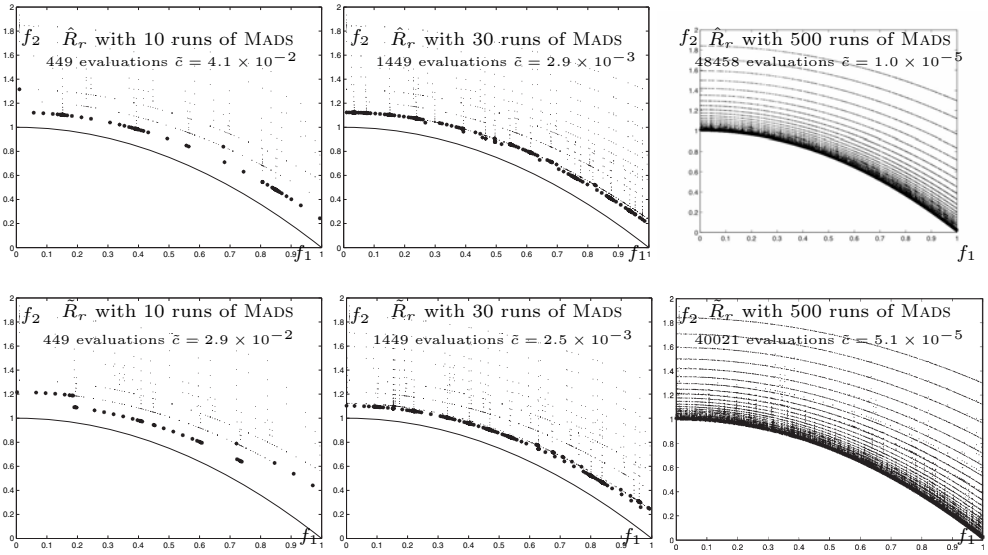


FIG. 5.5. *Biased search:* BIMADS *with* 10, 30, *and* 500 *optimization runs of* MADS.

where γ is a real parameter that controls the biasness in the search space. To complete the illustrative example, the following functions for f_1 and h are used [15]:

$$f_1(x_1) = x_1,$$
$$h(f_1, g) = 1 - \left(\frac{f_1}{g}\right)^2.$$

Results obtained for $\gamma = 0.25$ by applying BIMADS with 10 and 30 runs of MADS are shown in the top part of Figure 5.5. The nondominated solutions found by the algorithm approach the global Pareto curve; but no solution lies on the Pareto curve. Deb [15] observes that if $\gamma < 1$, then the density of solutions increases by moving away from the Pareto front. Hence, by randomly generating 50,000 solutions for $\gamma = 0.25$, not even one solution lies near the Pareto optimal front [15]. For comparison purposes, the maximal number of function evaluations was raised to 50,000 (i.e., the number of runs of MADS was increased to 500 and the maximal number of function evaluations was raised to 100 for each run). Results are shown in the bottom of Figure 5.5. By increasing the objective evaluations number, all nondominated points $Y_{\mathcal{L}}$ found by BIMADS lie on the global Pareto front $Y_{\mathcal{P}}$. The points are well spread along $Y_{\mathcal{P}}$.

5.4. Nonuniformly represented Pareto front. Nonuniformity of solution along the Pareto front is achieved by choosing a nonlinear function for f_1. An example

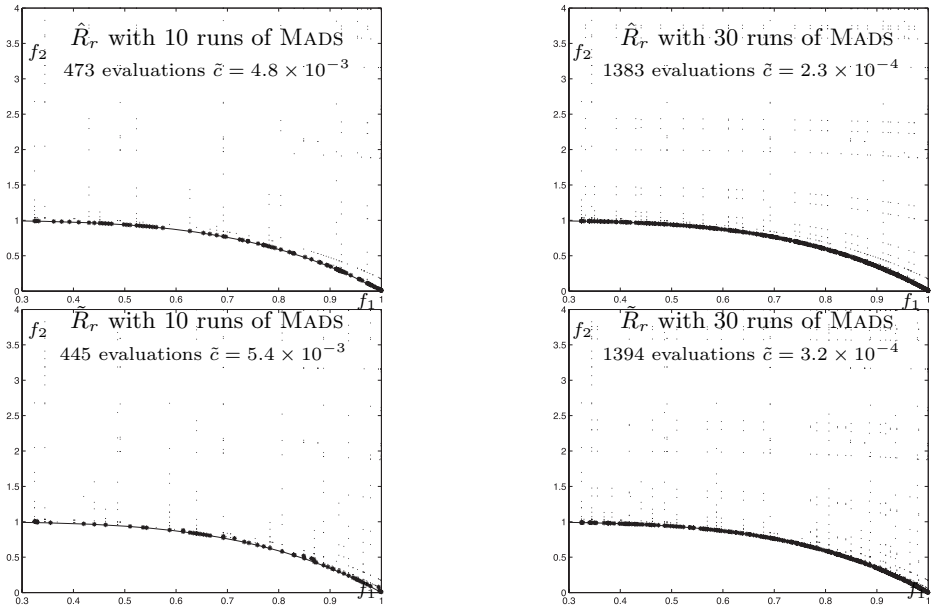


FIG. 5.6. *Nonuniformly represented Pareto front:* BIMADS *with* 10 *and* 30 *optimization runs of* MADS.

is given in [15] with the following functions:

$$f_1(x_1) = 1 - \exp(-4x_1)\sin^4(5\pi x_1),$$

$$h(f_1, g) = \begin{cases} 1 - (\frac{f_1}{g})^4 & \text{if } f_1 \leq g, \\ 0 & \text{otherwise,} \end{cases}$$

$$g(x_2) = \begin{cases} 4 - 3\exp(-\frac{x_2 - 0.2}{0.02})^2 & \text{if } 0 \leq x_2 \leq 0.4, \\ \\ 4 - 2\exp(-\frac{x_2 - 0.7}{0.2})^2 & \text{if } 0.4 \leq x_2 \leq 1. \end{cases}$$

Using 500 uniformly-spaced points in x_1, Deb [15] shows that the corresponding Pareto curve is biased for solutions for which f_1 value is near 1: most of the generated solutions have the function value $f_1 = 1$. Thereby, solutions cluster around $f_1 = 1$ values. Results obtained by applying BIMADS with 10 and 30 runs of MADS are shown in Figure 5.6. BIMADS overcomes the bias around $f_1 = 1$ values. Both methods find well spread nondominated solutions $Y_{\mathcal{L}}$ belonging to the Pareto curve $Y_{\mathcal{P}}$. With 30 iterations run, the Pareto solutions are more uniformly distributed.

6. Discussion. We proposed a new solution approach for MOP ensuring some first-order necessary optimality conditions for nonsmooth functions. In addition to the convergence analysis, a new analysis of uniformity is proposed. Our new algorithm BIMADS for biobjective optimization BOP was presented and applied to six problems from the literature designed to highlight some intrinsic difficulties of BOP. The algorithm performance is evaluated by studying the quality of solution set in terms of proximity to the Pareto front and uniformity of solutions distribution.

Results for all test problems are summarized in Table 6.1. The entries of Table 6.1 are mean value μ and standard deviation σ of the coverage measure $\tilde{c}$ over the six test problems.

TABLE 6.1
Coverage measure $\tilde{c}$ *for test problems.*

$\tilde{c}$	$\hat{R}_r$ formulation		$\tilde{R}_r$ formulation	
Runs	μ	σ	μ	σ
10 runs	.1714	.349	.0177	.0105
30 runs	.0025	.002	.0017	.0013

Table 6.1 shows that the uniformity of Pareto solutions distribution is sensitive to the number of runs and the formulation adopted. The best results are obtained by applying 30 runs of MADS using the $\tilde{R}_r$ formulation. This may be in part due to the fact that $\tilde{R}_r$ does not introduce additional nonsmoothness as $\hat{R}_r$ does.

In future work, we plan to apply BIMADS to larger and nonsmooth real engineering problems. We also plan to study the case where blackbox constraints are present. A simple way to do so would be to treat the constraints by the barrier approach to reject all infeasible points as done by MADS. However, the convergence analysis would not be trivial for nonsmooth functions. Another aspect that we wish to study is the use of parallelism in the biobjective framework, and the extension to problems with more than two objective functions.

Acknowledgments. The authors thank J. E. Dennis Jr. for his invaluable suggestions. They also thank the anonymous authors for their helpful comments.

REFERENCES

[1] M. A. ABRAMSON, *Mixed variable optimization of a load-bearing thermal insulation system using a filter pattern search algorithm*, Optim. Eng., 5 (2004), pp. 157–177.
[2] P. ARMAND, *Finding all maximal efficient faces in multiobjective linear programming*, Math. Programming, 61 (1993), pp. 357–377.
[3] C. AUDET AND J. E. DENNIS, JR., *Analysis of generalized pattern searches*, SIAM J. Optim., 13 (2002), pp. 889–903.
[4] C. AUDET AND J. E. DENNIS, JR., *Mesh adaptive direct search algorithms for constrained optimization*, SIAM J. Optim., 17 (2006), pp. 188–217.
[5] C. AUDET AND D. ORBAN, *Finding optimal algorithmic parameters using derivative-free optimization*, SIAM J. Optim., 17 (2006), pp. 642–664.
[6] H. P. BENSON, *An outer approximation algorithm for generating all efficient extreme points in the outcome set of a multiple objective linear programming problem*, J. Global Optim., 13 (1998), pp. 1–24.
[7] A. J. BOOKER, J. E. DENNIS, JR., P. D. FRANK, D. B. SERAFINI, AND V. TORCZON, *Optimization using surrogate objectives on a helicopter test example*, in Computational Methods Optimal Design and Control, J. Borggaard, J. Burns, E. Cliff, and S. Schreck, eds., Progr. Systems Control Theory 24, Birkhäuser, Boston, MA, 1998, pp. 49–58.
[8] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983. Reissued in 1990 Classics in Applied Mathematics, SIAM, Philadelphia.
[9] J. L. COHON, *Multiobjective Programming and Planning*, Academic Press, New York, 1978.
[10] I. D. COOPE AND C. J. PRICE, *Frame based methods for unconstrained optimization*, J. Optim. Theory Appl., 107 (2000), pp. 261–274.
[11] H.W. CORLEY, *Optimality conditions for maximizations of set-valued functions*, J. Optim. Theory Appl., 58 (1988), pp. 1–10.
[12] G. COUTURE, C. AUDET, J. E. DENNIS, JR., AND M. A. ABRAMSON, *The NOMAD Project*, available online at http://www.gerad.ca/NOMAD/.
[13] I. DAS, *Nonlinear Multicriteria Optimization and Robust Optimality*, Ph.D. thesis, Rice University, Houston, TX, 1997.
[14] I. DAS AND J. E. DENNIS, JR., *Normal-boundary intersection: A new method for generating the Pareto surface in nonlinear multicriteria optimization problems*, SIAM J. Optim., 8 (1998), pp. 631–657.
[15] K. DEB, *Multi-objective genetic algorithms: Problem difficulties and construction of test problems*, Evol. Comput., 7 (1999), pp. 205–230.

[16] M. Ehrgott, *Multicriteria Optimization*, in Lecture Notes in Economics and Mathematical Systems 491, Springer-Verlag, Berlin, 2000.
[17] M. Ehrgott and M. M. Wiecek, *Multiobjective Programming*, in Multiple Criteria Decision Analysis State of the Art Surveys, Springer-Verlag, Berlin, 2005, pp. 667–772.
[18] A. M. Geoffrion, *Proper efficiency and the theory of vector maximization*, J. Math. Anal. Appl., 22 (1968), pp. 618–630.
[19] M. Kokkolaras, C. Audet, and J. E. Dennis, Jr., *Mixed variable optimization of the number and composition of heat intercepts in a thermal insulation system*, Optim. Eng., 2 (2001), pp. 5–29.
[20] E. B. Leach, *A note on inverse function theorems*, Proc. Amer. Math. Soc., 12 (1961), pp. 694–697.
[21] F. A. Lootsma, T. W. Athan, and P. W. Papalambros, *Controling the search for a compromise solution in multi-objective optimization*, Eng. Optim., 25 (1998), pp. 65–81.
[22] H. Markowitz, *Portfolio selection*, J. Finance, 7 (1952), pp. 77–91.
[23] A. L. Marsden, M. Wang, J. E. Dennis, Jr., and P. Moin, *Optimal aeroacoustic shape design using the surrogate management framework*, Optim. Eng., 5 (2004), pp. 235–262.
[24] R. T. Rockafellar, *Generalized directional derivatives and subgradients of nonconvex functions*, Canad. J. Math., 32 (1980), pp. 257–280.
[25] V. Torczon, *On the convergence of pattern search algorithms*, SIAM J. Optim., 7 (1997), pp. 1–25.
[26] P. L. Yu, *Cone convexity, cone extreme points, and nondominated solutions in decision problems with multiobjectives*, J. Optim. Theory Appl., 14 (1974), pp. 319–377.
[27] M. Zeleny, *Compromise programming*, in Multiple Criteria Decision Making, J. L. Cochrane and M. Zeleny, eds., University of South Carolina Press, Columbia, SC, 1973, pp. 262–301.
[28] E. Zitzler, *Evolutionary Algorithms for Multiobjective Optimization: Methods and Applications*, Ph.D. thesis, Institute of Technology, Zurich, Switzerland, 1999.

SIAM J. OPTIM.
Vol. 19, No. 1, pp. 211–216

WELL POSED OPTIMIZATION PROBLEMS AND NONCONVEX CHEBYSHEV SETS IN HILBERT SPACES*

FRANCESCA FARACI† AND ANTONIO IANNIZZOTTO†

Abstract. A result on the existence and uniqueness of metric projection for certain sets is proved, by means of a saddle point theorem. A conjecture, based on such a result and aiming for the construction of a nonconvex Chebyshev set in a Hilbert space, is presented.

Key words. Chebyshev sets, Hilbert spaces, metric projections

AMS subject classifications. 41A65, 41A52

DOI. 10.1137/06067496X

1. Introduction.

In the present paper we deal with the classical problem of convexity of Chebyshev sets: we are going to present a conjecture, aimed at the construction of a nonconvex Chebyshev set in a Hilbert space. First, let us introduce some definitions and results which we are going to employ.

In what follows, we follow the surveys by Vlasov [10] and Balaganskiĭ and Vlasov [2]; we also refer the reader to a recent paper by Cobzaş [3] for a state-of-the-art description of the theory. Let X be a (real) Banach space, M be a subset of X: for every $x \in X$, we define the *distance* from x to M as

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

Also, we define the *metric projection* onto M as the set-valued function $P_M : X \to 2^M$ mapping $x \in X$ into the (possibly empty) set

$$P_M(x) := \{y \in M \ : \ \|x - y\| = d(x, M)\}.$$

DEFINITION 1. *A subset M of X is a* Chebyshev set *if $P_M(x)$ is a singleton for every $x \in X$.*

An important family of Chebyshev sets, in certain Banach spaces, is that of *convex closed* sets, as the following classical result proves (in the form of a characterization).

THEOREM 1 (see [10, Theorem 0.6]). *The following are equivalent:*

- *X is strictly convex and reflexive;*
- *every convex closed subset of X is Chebyshev.*

Obviously, every Chebyshev set in a Banach space is closed, but it is not known, in general, whether a Chebyshev set is necessarily convex. So, the following question is a natural one and represents one of the most celebrated and challenging open problems in the theory of best approximation: *does a class of Banach spaces containing nonconvex Chebyshev sets exist?*

For strictly convex, smooth, finite-dimensional spaces, the answer is negative, as proved by Efimov and Stechkin in [5]. However, the problem is still open in the case of infinite-dimensional Banach spaces. Note that completeness is a fundamental

*Received by the editors November 15, 2006; accepted for publication (in revised form) November 6, 2007; published electronically February 20, 2008.
http://www.siam.org/journals/siopt/19-1/67496.html
†Dipartimento di Matematica e Informatica, Università degli Studi di Catania, Viale A. Doria 6, 95125 Catania, Italy (ffaraci@dmi.unict.it, iannizzotto@dmi.unict.it).

assumption here: for instance, Johnson in [7] has proved the existence of a nonconvex Chebyshev set in a pre-Hilbert space with infinite dimension.

A partial answer is known under a twofold restriction, that is, for the case of *approximatively compact* Chebyshev sets in uniformly convex, smooth spaces. We will rapidly overview this result, since it will be useful in what follows. Let us give the following definition.

DEFINITION 2. *Let M be a subset of X, x be a point of X: a* minimizing sequence *in M for x is a sequence $\{y_n\}$ in M such that $\|x - y_n\| \to d(x, M)$. M is* approximatively compact *if, for every $x \in X$, each minimizing sequence in M for x admits a subsequence $\{y_{n_k}\}$ converging to an element of M.*

We note that, in the above framework, the limit of $\{y_{n_k}\}$ lies in $P_M(x)$. Under such restrictions, Efimov and Stechkin proved in [6] the following result.

THEOREM 2. *Let X be a uniformly convex, smooth Banach space, and M be a Chebyshev set in X. Then, the following are equivalent:*

- *M is convex;*
- *M is approximatively compact;*
- *M is sequentially weakly closed.*

We observe that, whenever M is a Chebyshev set, approximative compactness is equivalent to the following property for all $x \in X$:

(S) each minimizing sequence in M for x converges to an element of M.

Obviously, the limit of each minimizing sequence in M for x is the unique point of $P_M(x)$.

In particular, we are interested in the problem of convexity of Chebyshev sets in Hilbert spaces, which was studied by Asplund in [1]. In [8], Klee put forward the conjecture that, in an infinite-dimensional Hilbert space, nonconvex Chebyshev sets do exist. To support his conjecture, he proposed the following example: in the Hilbert space ℓ^2, let

$$K := \left\{ \{x_n\} \in \ell^2 \ : \ \sum_{n=1}^{\infty} n x_n^2 < 1 \right\}$$

and define

$$M := \left\{ \{x_n\} \in \ell^2 \ : \ d(\{x_n\}, K) \geq 1 \right\};$$

then M is not convex and, for every $\{x_n\} \notin K$, $P_M(\{x_n\})$ is a singleton, while for $\{x_n\} \in K$, $P_M(\{x_n\}) = \emptyset$.

Our approach is similar to that of Klee, though set in a more general framework. It is based on the following result (see section 2 for the proof).

THEOREM 3. *Let X be a Hilbert space, C be a nonempty subset of X, K be the closed convex hull of C, and $x_0 \in X \setminus K$. Then, for all real $t > 0$, denoting*

$$M_t := \{ y \in X \ : \ d(y, C) \geq t \},$$

the set $P_{M_t}(x_0)$ is a singleton, towards which each minimizing sequence in M_t for x_0 converges.

In connection with the above result, we present a conjecture.

CONJECTURE 1. *There exist a Hilbert space X, a subset C of X, and a real number $t > 0$ such that the set M_t satisfies the following conditions:*

(A_1) *M_t is not convex;*

(A_2) *for every x lying in the closed convex hull of C, $P_{M_t}(x)$ is a singleton.*

Our motivation is the following: suppose X, C, and t comply with Conjecture 1; then, due to Theorem 3, M_t would be a nonconvex Chebyshev set in X.

We believe that our conjecture is likely to be proved, as, due to the definition of M_t, condition (A_1) is very easy to fulfill (the same is not true for condition (A_2), of course). Also, we observe that, by Theorem 2, if Conjecture 1 is true, then M_t cannot be approximatively compact (or, equivalently, sequentially weakly closed): that is, since condition (S) holds for all $x \notin K$, there must be at least one point $x \in K$ and a minimizing sequence in M_t for x which does not converge.

Our result relies on the general method introduced by Ricceri in [9], where some minimization problems, related to C^1 functionals with locally Lipschitz derivative, are studied following a new approach based on a classical saddle point theorem, which we recall here in a form suitable to our purposes.

THEOREM 4 (see [11, Theorem 49.A]). *Let Λ be a compact real interval, and $\Phi : X \times \Lambda \to \mathbb{R}$ be a function such that*

- *$\Phi(\cdot, \lambda)$ is continuous and convex for all $\lambda \in \Lambda$;*
- *$\Phi(x, \cdot)$ is continuous and concave for all $x \in X$;*
- *there exists $\lambda_0 \in \Lambda$ such that $\Phi(\cdot, \lambda_0)$ is coercive.*

Then there exists a pair $(\bar{x}, \bar{\lambda}) \in X \times \Lambda$ such that

$$\Phi(\bar{x}, \bar{\lambda}) = \min_{x \in X} \Phi(x, \bar{\lambda}) = \max_{\lambda \in \Lambda} \Phi(\bar{x}, \lambda).$$

2. The results. Our first step is the following general result.

THEOREM 5. *Let X be a Hilbert space, C be a nonempty subset of X, $x_0 \in X$, and for all $\lambda \in [0, 1]$ let $I_\lambda : X \to \mathbb{R}$ be defined by*

$$I_\lambda(x) = \|x - x_0\|^2 - \lambda d^2(x, C).$$

Moreover, assume the following condition:

(D) *x_0 is not a global minimizer of the functional I_1.*

Then there exists a positive $\tau \in \mathbb{R} \cup \{+\infty\}$, $\tau > d(x_0, C)$ such that for all $t \in]d(x_0, C), \tau[$, denoting

$$M_t := \{y \in X \ : \ d(y, C) \geq t\},$$

the set $P_{M_t}(x_0)$ is a singleton, towards which each minimizing sequence in M_t for x_0 converges.

Proof. First we prove that the functional I_1 is convex: indeed, it can be expressed as

$$I_1(x) = \sup_{y \in C} \left[2\langle y - x_0, x\rangle + \|x_0\|^2 - \|y\|^2\right],$$

so it is convex as the supremum of a family of convex functions. From this, it readily follows that I_λ is strictly convex for all $\lambda \in [0, 1[$.

We denote by Q the set of global minimizers of I_1, and we put

$$\rho := d(x_0, Q) \quad (\rho = +\infty \text{ if } Q = \emptyset), \tag{2.1}$$

so by condition (D), since Q is closed, we have $\rho > 0$. Then we set

$$\tau := \sup_{x \in B(x_0, \rho)} d(x, C) \quad (\tau = +\infty \text{ if } \rho = +\infty).$$

It is easily seen that

$$\tau > d(x_0, C).$$

Indeed, clearly $\tau \geq d(x_0, C)$; moreover, assuming $\tau = d(x_0, C)$, for all $x \in B(x_0, \rho)$ we would have

$$I_1(x) = \|x - x_0\|^2 - d^2(x, C) \geq -d^2(x_0, C) = I_1(x_0),$$

so x_0 would be a local minimizer of the convex functional I_1, which, in turn, would imply $x_0 \in Q$, against condition (D).

Choose an arbitrary $t \in]d(x_0, C), \tau[$. We observe that $d(x_0, M_t) \in]0, \rho[$: indeed, $d(x_0, M_t) > 0$ as $x_0 \notin M_t$ and the latter is a closed set; on the other hand, $d(x_0, M_t) < \rho$ since, as $t < \tau$, there exists $x \in B(x_0, \rho) \cap M_t$, so

$$d(x_0, M_t) \leq \|x_0 - x\| < \rho.$$

We are going to apply Theorem 4 with $\Lambda = [0, 1]$ and

$$\Phi(x, \lambda) = \|x - x_0\|^2 + \lambda \left(t^2 - d^2(x, C)\right).$$

Such a function complies with all the hypotheses of Theorem 4 (with an arbitrary $\lambda_0 \in [0, 1[$), and hence there exists a pair $(\bar{x}, \bar{\lambda}) \in X \times \Lambda$ satisfying

$$\begin{aligned} \|\bar{x} - x_0\|^2 + \bar{\lambda}\left(t^2 - d^2(\bar{x}, C)\right) &= \inf_{x \in X} \left[\|x - x_0\|^2 - \bar{\lambda} d^2(x, C)\right] + \bar{\lambda} t^2 \\ &= \|\bar{x} - x_0\|^2 + \sup_{\lambda \in \Lambda} \left[\lambda \left(t^2 - d^2(\bar{x}, C)\right)\right]. \end{aligned} \tag{2.2}$$

The above equality has very important consequences for the proof. First we prove that

$$d(\bar{x}, C) \leq t. \tag{2.3}$$

Arguing by contradiction, we suppose $d(\bar{x}, C) > t$: then, from (2.2) it would follow that $\bar{\lambda} = 0$, which, in turn, would imply $\bar{x} = x_0$ and so $d(\bar{x}, C) < t$, against our assumption. Then we prove that

$$\bar{\lambda} < 1. \tag{2.4}$$

Again, we argue by contradiction and assume $\bar{\lambda} = 1$: hence, from (2.2) we would deduce that $\bar{x} \in Q$, so by (2.1), $\|\bar{x} - x_0\| \geq \rho$; the latter inequality, together with (2.3), yields

$$I_1(\bar{x}) \geq \rho^2 - t^2.$$

Recalling that $t < \tau$, we note that there exists $y \in B(x_0, \rho)$ such that $d(y, C) > t$ and so

$$I_1(y) < \rho^2 - t^2,$$

against the fact that $\bar{x} \in Q$. Now we can improve (2.3) and get

$$d(\bar{x}, C) = t \tag{2.5}$$

(in particular, $\bar{x} \in M_t$). Indeed, if $d(\bar{x}, C) < t$, from (2.2) it would follow that $\bar{\lambda} = 1$, against (2.4).

Now we are in a position to prove our assertion. With this aim in mind, we observe that, by (2.2), $\bar{x}$ is a global minimizer of the functional $I_{\bar{\lambda}}$; moreover, by (2.4), $I_{\bar{\lambda}}$ is strictly convex, continuous, and coercive, so $\bar{x}$ is its only minimizer, and every sequence $\{x_n\}$ in X, such that $I_{\bar{\lambda}}(x_n) \to I_{\bar{\lambda}}(\bar{x})$, weakly converges to $\bar{x}$ (see [4, Example 8, p. 3]).

We now prove that $P_{M_t}(x_0) = \{\bar{x}\}$, that is, that for every $y \in M_t$, $y \neq \bar{x}$, we have $\|y - x_0\| > \|\bar{x} - x_0\|$. Indeed, by what has been argued above, $I_{\bar{\lambda}}(y) > I_{\bar{\lambda}}(\bar{x})$, which, by (2.5), implies

$$\|y - x_0\|^2 > \|\bar{x} - x_0\|^2 + \bar{\lambda}\left(d^2(y, C) - t^2\right) \geq \|\bar{x} - x_0\|^2.$$

Finally, we prove that condition (S) holds for the point x_0, that is, that every sequence $\{y_n\}$ in M_t such that $\|y_n - x_0\| \to \|\bar{x} - x_0\|$ converges to $\bar{x}$. Indeed, for all $n \in \mathbb{N}$ we have

$$I_{\bar{\lambda}}(y_n) \leq \|y_n - x_0\|^2 - \bar{\lambda} t^2,$$

and the right-hand side tends to $I_{\bar{\lambda}}(\bar{x})$, so, by what has been stated above, $\{y_n\}$ weakly converges to $\bar{x}$; then, by well-known results, the weak convergence implies the strong convergence. Thus, the proof is complete. □

Our thesis, in the language of optimization theory, can be expressed by saying that, for every $t \in]d(x_0, C), \tau[$, the problem of minimizing on M_t the distance from x_0 is *Tykhonov well posed* (see [4]).

Remark 1. Under the same assumptions as in Theorem 5, an analogous result is achieved for the set

$$N_t := \{y \in X \ : \ d(y, C) = t\}$$

for all $t \in]d(x_0, C), \tau[$.

Our main hypothesis, namely condition (D), is indeed very general. In particular, it is fulfilled whenever x_0 does not belong to the closed convex hull of C. Thus, Theorem 3 is easily deduced from Theorem 5, as follows.

Proof of Theorem 3. Let us fix $t > 0$; then one of the following cases occurs:

- If $d(x_0, C) \geq t$, then $x_0 \in M_t$, so obviously $P_{M_t}(x_0) = \{x_0\}$, and each minimizing sequence in M_t for x_0 tends to x_0.
- If $d(x_0, C) < t$, then we apply Theorem 5. With this aim in mind, we prove that the functional I_1 is unbounded from below. Indeed, since $x_0 \notin K$, we can apply the separation theorem in its strongest form, assuring the existence of an element $\bar{y} \in X$, $\bar{y} \neq 0$, and of a positive ε such that

$$\langle \bar{y}, x \rangle \leq \langle \bar{y}, x_0 \rangle - \varepsilon \quad \text{for all } x \in K.$$

For all $\mu > 0$ we get

$$\begin{aligned} I_1(x_0 + \mu\bar{y}) &= \|\mu\bar{y}\|^2 - \inf_{x \in C} \|x_0 + \mu\bar{y} - x\|^2 \\ &= \sup_{x \in C} \left(-\|x_0 - x\|^2 + 2\mu\langle \bar{y}, x - x_0 \rangle\right) \\ &\leq -d^2(x_0, C) - 2\mu\varepsilon, \end{aligned}$$

so clearly

$$\lim_{\mu\to+\infty} I_1(x_0+\mu\bar{y}) = -\infty.$$

Thus, the set Q is empty: in particular, condition (D) is fulfilled. The thesis of Theorem 5 follows, with $\tau = +\infty$, which concludes the proof. □

Acknowledgment. We would like to thank the anonymous referees for useful comments and suggestions.

REFERENCES

[1] E. Asplund, *Čebyšev sets in Hilbert space*, Trans. Amer. Math. Soc., 144 (1969), pp. 235–240.
[2] V. S. Balaganskiĭ and L. P. Vlasov, *The problem of the convexity of Chebyshev sets*, Russian Math. Surveys, 51 (1996), pp. 1127–1190.
[3] S. Cobzaş, *Geometric properties of Banach spaces and the existence of nearest and farthest points*, Abstr. Appl. Anal., 3 (2005), pp. 259–285.
[4] A. L. Dontchev and T. Zolezzi, *Well-Posed Optimization Problems*, Springer-Verlag, Berlin, 1993.
[5] N. V. Efimov and S. B. Stechkin, *Support properties of sets in Banach spaces and Čebyšev sets*, Dokl. Akad. Nauk SSSR, 127 (1959), pp. 254–257.
[6] N. V. Efimov and S. B. Stechkin, *Approximative compactness and Chebyshev sets*, Dokl. Akad. Nauk SSSR, 140 (1961), pp. 522–524.
[7] G. G. Johnson, *A nonconvex set which has the unique nearest point property*, J. Approx. Theory, 51 (1987), pp. 289–332.
[8] V. L. Klee, *Remarks on nearest points in normed linear spaces*, in Proceedings of the Colloquium on Convexity (Copenhagen, 1965), Kobenhavns Univ. Mat. Inst., Copenhagen, pp. 168–176.
[9] B. Ricceri, *The problem of minimizing locally a C^2 functional around noncritical points is well-posed*, Proc. Amer. Math. Soc., 135 (2007), pp. 2187–2191.
[10] L. P. Vlasov, *Approximative properties of sets in normed linear spaces*, Russian Math. Surveys, 28 (1973), pp. 1–66.
[11] E. Zeidler, *Nonlinear Functional Analysis and Its Applications*, *Vol.* III, Springer-Verlag, New York, 1985.

SIAM J. OPTIM.
Vol. 19, No. 1, pp. 217–233

REGULARITY CONDITIONS VIA QUASI-RELATIVE INTERIOR IN CONVEX PROGRAMMING*

RADU IOAN BOŢ†, ERNÖ ROBERT CSETNEK†, AND GERT WANKA†

Abstract. We give some new regularity conditions for Fenchel duality in separated locally convex vector spaces, written in terms of the notion of quasi interior and quasi-relative interior, respectively. We provide also an example of a convex optimization problem for which the classical generalized interior-point conditions given so far in the literature cannot be applied, while the one given by us is applicable. By using a technique developed by Magnanti, we derive some duality results for the optimization problem with cone constraints and its Lagrange dual problem, and we show that a duality result recently given in the literature for this pair of problems has self-contradictory assumptions.

Key words. convex programming, Fenchel duality, Lagrange duality, quasi-relative interior

AMS subject classifications. 90C25, 46A20, 90C51

DOI. 10.1137/07068432X

1. Introduction. Usually there is a so-called duality gap between the optimal objective values of a primal convex optimization problem and its dual problem. A challenge in convex analysis is to give sufficient conditions which guarantee strong duality, the situation when the optimal objective values of the two problems are equal and the dual problem has an optimal solution. Several generalized interior-point conditions were given in the past in order to eliminate the above-mentioned duality gap. Along the classical interior, some generalized interior notions were used, such as the core [14], the intrinsic core [9], or the strong quasi-relative interior [2], in order to give regularity conditions which guarantee strong duality. For an overview of these conditions we invite the reader to consult [8], [16] (see also [17] for more on this subject).

Unfortunately, for infinite-dimensional convex optimization problems, also in practice, it can happen that the duality results given in the past cannot be applied because, for instance, the interior of the set involved in the regularity condition is empty. This is the case, for example, when we deal with the positive cones

$$l^p_+ = \{x = (x_n)_{n\in\mathbb{N}} \in l^p : x_n \geq 0 \ \forall n \in \mathbb{N}\}$$

and

$$L^p_+(T,\mu) = \{u \in L^p(T,\mu) : u(t) \geq 0, \text{ a.e.}\}$$

of the spaces l^p and $L^p(T,\mu)$, respectively, where (T,μ) is a σ-finite measure space and $p \in [1,\infty)$. Moreover, also the strong quasi-relative interior (which is the weakest generalized interior notion from the one mentioned above) of these cones is empty. For this reason, for a convex set, Borwein and Lewis introduced the notion of the quasi-relative interior [3], which generalizes all of the above-mentioned interior notions. They proved that the quasi-relative interiors of l^p_+ and $L^p_+(T,\mu)$ are nonempty.

*Received by the editors March 5, 2007; accepted for publication (in revised form) October 1, 2007; published electronically March 5, 2008.
http://www.siam.org/journals/siopt/19-1/68432.html
†Faculty of Mathematics, Chemnitz University of Technology, D-09107 Chemnitz, Germany (radu.bot@mathematik.tu-chemnitz.de, robert.csetnek@mathematik.tu-chemnitz.de, gert.wanka@mathematik.tu-chemnitz.de). The research of the second author was supported by a Graduate Fellowship of the Free State Saxony, Germany.

In this paper, we start by considering the primal optimization problem with the objective function being the sum of two proper convex functions defined on a separated locally convex vector space, to which we attach its Fenchel dual problem, stated in terms of the conjugates of the two functions. We give a new regularity condition for Fenchel duality based on the notion of the quasi-relative interior of a convex set using a separation theorem given by Cammaroto and Di Bella in [4]. Further, two stronger regularity conditions are also given. We provide an appropriate example for which our duality results are applicable, while the other generalized interior-point conditions given in the past fail, justifying the theory developed in this paper. Then we state duality results for the case when the objective function of the primal problem is the sum of a proper convex function with the composition of another proper convex function with a continuous linear operator. Let us notice that for this case Borwein and Lewis in [3] also gave some conditions by means of the quasi-relative interior, but they considered a more restrictive case, namely, that the codomain of the linear operator is finite-dimensional. We consider the more general case, when both of the spaces are infinite-dimensional.

In 1974 Magnanti proved that "Fenchel and Lagrange duality are equivalent" in the sense that the classical Fenchel duality result can be deduced from the classical Lagrange duality result, and vice versa (see [13]). By using this technique we derive some Lagrange duality results for the convex optimization problem with cone constraints, written in terms of the quasi-relative interior. Let us notice that another condition for Lagrange duality, stated also in terms of the quasi-relative interior, was given recently by Cammaroto and Di Bella in [4]. We show that this result has self-contradictory assumptions. Let us mention that also in [11] some regularity conditions, in terms of the quasi-relative interior, have been introduced. However, most of these conditions require the interior of a cone to be nonempty, and this fails for many optimization problems as we pointed out above.

The paper is structured as follows. In the next section we give some definitions and results which will be used later in the paper. Section 3 is devoted to the theory of Fenchel duality. We give here the announced regularity conditions written in terms of the quasi-relative interior. By using an idea due to Magnanti we derive in section 4 some duality results for the optimization problem with cone constraints and its Lagrange dual problem.

2. Preliminary notions and results. Consider X, a separated locally convex vector space, and X^*, its topological dual space. We denote by $\langle x^*, x\rangle$ the value of the linear continuous functional $x^* \in X^*$ at $x \in X$. Further, let $\mathrm{id}_X : X \to X$, $\mathrm{id}_X(x) = x$, for all $x \in X$, be the *identity function* of X. The *indicator function* of $C \subseteq X$, denoted by δ_C, is defined as $\delta_C : X \to \overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$,

$$\delta_C(x) = \begin{cases} 0 & \text{if } x \in C, \\ +\infty & \text{otherwise.} \end{cases}$$

For a function $f : X \to \overline{\mathbb{R}}$ we denote by $\mathrm{dom}(f) = \{x \in X : f(x) < +\infty\}$ its *domain* and by $\mathrm{epi}(f) = \{(x, r) \in X \times \mathbb{R} : f(x) \leq r\}$ its *epigraph*. We call f *proper* if $\mathrm{dom}(f) \neq \emptyset$ and $f(x) > -\infty$ for all $x \in X$. We also denote by $\widehat{\mathrm{epi}}(f) = \{(x, r) \in X \times \mathbb{R} : (x, -r) \in \mathrm{epi}(f)\}$ the symmetric of $\mathrm{epi}(f)$ with respect to the x-axis. For a given real number α, $f - \alpha : X \to \overline{\mathbb{R}}$ is, as usual, the function defined by $(f - \alpha)(x) = f(x) - \alpha$ for all $x \in X$. Given two functions $f : M_1 \to M_2$ and $g : N_1 \to N_2$, where M_1, M_2, N_1, N_2 are nonempty sets, we define the function $f \times g : M_1 \times N_1 \to M_2 \times N_2$ by $f \times g(m, n) = (f(m), g(n))$ for all $(m, n) \in M_1 \times N_1$.

The *Fenchel–Moreau conjugate* of f is the function $f^* : X^* \to \overline{\mathbb{R}}$ defined by

$$f^*(x^*) = \sup_{x \in X} \{\langle x^*, x\rangle - f(x)\} \ \forall x^* \in X^*.$$

For a subset C of X we denote by $\operatorname{co} C$, $\operatorname{aff} C$, $\operatorname{cl} C$, and $\operatorname{int} C$ its *convex hull*, *affine hull*, *closure*, and *interior*, respectively. The set $\operatorname{cone} C := \bigcup_{\lambda \geq 0} \lambda C$ is the *cone generated by* C. The following property, the proof of which we omit since it presents no difficulty, will be used throughout the paper: If C is convex, then

$$\operatorname{cone} \operatorname{co}(C \cup \{0\}) = \operatorname{cone} C. \tag{1}$$

The *normal cone* of C at $x \in C$ is defined as $N_C(x) = \{x^* \in X^* : \langle x^*, y - x\rangle \leq 0, \ \forall y \in C\}$.

DEFINITION 2.1 (see [3]). *Let C be a convex subset of X. The* quasi-relative interior *of C is the set*

$$\operatorname{qri} C = \{x \in C : \operatorname{cl} \operatorname{cone}(C - x) \text{ is a linear subspace of } X\}.$$

We give the following useful characterization of the quasi-relative interior of a convex set.

PROPOSITION 2.2 (see [3]). *Let C be a convex subset of X and $x \in C$. Then $x \in \operatorname{qri} C$ if and only if $N_C(x)$ is a linear subspace of X^*.*

In the following we consider another interior notion for a convex set, which is close to the one of the quasi-relative interior.

DEFINITION 2.3. *Let C be a convex subset of X. The* quasi interior *of C is the set*

$$\operatorname{qi} C = \{x \in C : \operatorname{cl} \operatorname{cone}(C - x) = X\}.$$

The following characterization of the quasi interior of a convex set was given in [6], where the space X was considered a reflexive Banach space. One can prove that this property is true even in a separated locally convex vector space.

PROPOSITION 2.4. *Let C be a convex subset of X and $x \in C$. Then $x \in \operatorname{qi} C$ if and only if $N_C(x) = \{0\}$.*

Proof. Assume first that $x \in \operatorname{qi} C$, and take an arbitrary element $x^* \in N_C(x)$. One can easily see that $\langle x^*, z\rangle \leq 0$ for all $z \in \operatorname{cl} \operatorname{cone}(C - x)$. Thus $\langle x^*, z\rangle \leq 0$ for all $z \in X$, which is nothing else than $x^* = 0$.

In order to prove the opposite implication we consider an arbitrary $\bar{x} \in X$ and prove that $\bar{x} \in \operatorname{cl} \operatorname{cone}(C - x)$. By assuming the contrary, by a separation theorem (see, for instance, Theorem 1.1.5 in [17]), one has that there exists $x^* \in X^* \setminus \{0\}$ and $\alpha \in \mathbb{R}$ such that

$$\langle x^*, z\rangle < \alpha < \langle x^*, \bar{x}\rangle \ \forall z \in \operatorname{cl} \operatorname{cone}(C - x).$$

Let $y \in C$ be fixed. For all $\lambda > 0$ it holds that $\langle x^*, y - x\rangle < \frac{1}{\lambda}\alpha$, and this implies that $\langle x^*, y - x\rangle \leq 0$. As this inequality is true for every arbitrary $y \in C$, we obtain that $x^* \in N_C(x)$. But this leads to a contradiction, and in this way the conclusion follows. □

It follows from the definitions above that $\operatorname{qi} C \subseteq \operatorname{qri} C$ and $\operatorname{qri}\{x\} = \{x\}$ for all $x \in X$. Moreover, if $\operatorname{qi} C \neq \emptyset$, then $\operatorname{qi} C = \operatorname{qri} C$. Although this property is given in [12] in the case of a real normed space, it holds also in an arbitrary separated

locally convex vector space, as follows by the properties given above. If X is a finite-dimensional space, then $\operatorname{qi} C = \operatorname{int} C$ (cf. [12]) and $\operatorname{qri} C = \operatorname{ri} C$ (cf. [3]), where $\operatorname{ri} C$ is the *relative interior* of C.

Useful properties of the quasi-relative interior are listed below. For the proof of (i)–(viii) we refer to [1] and [3].

PROPOSITION 2.5. *Let us consider C and D two convex subsets of X, $x \in X$, and $\alpha \in \mathbb{R}$. Then:*

(i) $\operatorname{qri} C + \operatorname{qri} D \subseteq \operatorname{qri}(C + D)$;
(ii) $\operatorname{qri}(C \times D) = \operatorname{qri} C \times \operatorname{qri} D$;
(iii) $\operatorname{qri}(C - x) = \operatorname{qri} C - x$;
(iv) $\operatorname{qri}(\alpha C) = \alpha \operatorname{qri} C$;
(v) $t \operatorname{qri} C + (1-t)C \subseteq \operatorname{qri} C$, $\forall t \in (0,1]$, *and hence* $\operatorname{qri} C$ *is a convex set;*
(vi) *if C is an affine set, then* $\operatorname{qri} C = C$;
(vii) $\operatorname{qri}(\operatorname{qri} C) = \operatorname{qri} C$.

If $\operatorname{qri} C \neq \emptyset$, *then:*

(viii) $\operatorname{cl} \operatorname{qri} C = \operatorname{cl} C$;
(ix) $\operatorname{cl} \operatorname{cone} \operatorname{qri} C = \operatorname{cl} \operatorname{cone} C$.

Proof. (ix) The inclusion $\operatorname{cl} \operatorname{cone} \operatorname{qri} C \subseteq \operatorname{cl} \operatorname{cone} C$ is obvious. We prove that $\operatorname{cone} C \subseteq \operatorname{cl} \operatorname{cone} \operatorname{qri} C$. Consider $x \in \operatorname{cone} C$ arbitrary. There exist $\lambda \geq 0$ and $c \in C$ such that $x = \lambda c$. Take $x_0 \in \operatorname{qri} C$. By applying property (v) we get $tx_0 + (1-t)c \in \operatorname{qri} C$ for all $t \in (0,1]$, so $\lambda t x_0 + (1-t)x = \lambda(tx_0 + (1-t)c) \in \operatorname{cone} \operatorname{qri} C$ for all $t \in (0,1]$. By passing to the limit as $t \searrow 0$ we obtain $x \in \operatorname{cl} \operatorname{cone} \operatorname{qri} C$, and hence the desired conclusion follows. □

The next lemma plays an important role in this paper.

LEMMA 2.6. *Let A and B be nonempty convex subsets of X such that* $\operatorname{qri} A \cap B \neq \emptyset$. *If* $0 \in \operatorname{qi}(A - A)$, *then* $0 \in \operatorname{qi}(A - B)$.

Proof. Take $x \in \operatorname{qri} A \cap B$, and let $x^* \in N_{A-B}(0)$ be arbitrary. We get $\langle x^*, a - b \rangle \leq 0$, for all $a \in A$, for all $b \in B$. This implies that

$$\langle x^*, a - x \rangle \leq 0 \ \forall a \in A, \tag{2}$$

that is, $x^* \in N_A(x)$. As $x \in \operatorname{qri} A$, $N_A(x)$ is a linear subspace of X^*, and hence $-x^* \in N_A(x)$, which is nothing else than

$$\langle x^*, x - a \rangle \leq 0 \ \forall a \in A. \tag{3}$$

The relations (2) and (3) give us $\langle x^*, a' - a'' \rangle \leq 0$, for all $a', a'' \in A$, so $x^* \in N_{A-A}(0)$. Since $0 \in \operatorname{qi}(A - A)$ we have $N_{A-A}(0) = \{0\}$ (cf. Proposition 2.4), and we get $x^* = 0$. As x^* was arbitrary chosen we obtain $N_{A-B}(0) = \{0\}$, and, by using again Proposition 2.4, the conclusion follows. □

Next we give useful separation theorems in terms of the notion of the quasi-relative interior.

THEOREM 2.7. *Let C be a convex subset of X and $x_0 \in C$. If $x_0 \notin \operatorname{qri} C$, then there exists $x^* \in X^*, x^* \neq 0$, such that*

$$\langle x^*, x \rangle \leq \langle x^*, x_0 \rangle \ \forall x \in C.$$

Vice versa, if there exists $x^* \in X^*$, $x^* \neq 0$, such that

$$\langle x^*, x \rangle \leq \langle x^*, x_0 \rangle \ \forall x \in C$$

and

$$0 \in \operatorname{qi}(C - C),$$

then $x_0 \notin \operatorname{qri} C$.

Proof. Suppose that $x_0 \notin \operatorname{qri} C$. According to Proposition 2.2, $N_C(x_0)$ is not a linear subspace of X^*, and hence there exists $x^* \in N_C(x_0)$, $x^* \neq 0$. By using the definition of the normal cone, we get that $\langle x^*, x\rangle \leq \langle x^*, x_0\rangle$ for all $x \in C$.

Conversely, assume that there exists $x^* \in X^*$, $x^* \neq 0$, such that $\langle x^*, x\rangle \leq \langle x^*, x_0\rangle$ for all $x \in C$ and $0 \in \operatorname{qi}(C - C)$. We obtain

$$\langle x^*, x - x_0\rangle \leq 0 \ \forall x \in C, \tag{4}$$

that is, $x^* \in N_C(x_0)$. If we suppose that $x_0 \in \operatorname{qri} C$, then $N_C(x_0)$ is a linear subspace of X^*, and hence $-x^* \in N_C(x_0)$. By combining this with (4) we get $\langle x^*, x - x_0\rangle = 0$ for all $x \in C$. The last relation implies $\langle x^*, x\rangle = 0$ for all $x \in C - C$, and from here one has further that $\langle x^*, x\rangle = 0$ for all $x \in \operatorname{cl}\operatorname{cone}(C - C) = X$. But this can be the case just if $x^* = 0$, which is a contradiction. In conclusion, $x_0 \notin \operatorname{qri} C$. □

Remark 2.8. In [5], [6] a similar separation theorem in the case when X is a real normed space is given. For the second part of the above theorem the authors require that the following condition must be fulfilled:

$$\operatorname{cl}(T_C(x_0) - T_C(x_0)) = X,$$

where

$$T_C(x_0) = \big\{y \in X : y = \lim_{n\to\infty} \lambda_n(x_n - x_0), \lambda_n > 0 \ \forall n \in \mathbb{N},$$
$$x_n \in C \ \forall n \in \mathbb{N} \text{ and } \lim_{n\to\infty} x_n = x_0\big\}$$

is called the *contingent cone* to C at $x_0 \in C$. In general, we have the following inclusion: $T_C(x_0) \subseteq \operatorname{cl}\operatorname{cone}(C - x_0)$. If the set C is convex, then $T_C(x_0) = \operatorname{cl}\operatorname{cone}(C - x_0)$ (cf. [10]). As $\operatorname{cl}(\operatorname{cl} E + \operatorname{cl} F) = \operatorname{cl}(E + F)$, for arbitrary sets E, F in X and $\operatorname{cone} A - \operatorname{cone} A = \operatorname{cone}(A - A)$, if A is a convex subset of X such that $0 \in A$, the condition $\operatorname{cl}(T_C(x_0) - T_C(x_0)) = X$ can be reformulated as follows: $\operatorname{cl}\operatorname{cone}(C - C) = X$ or, equivalently, $0 \in \operatorname{qi}(C - C)$. Indeed, we have

$$\operatorname{cl}[\operatorname{cl}\operatorname{cone}(C - x_0) - \operatorname{cl}\operatorname{cone}(C - x_0)] = X \Leftrightarrow \operatorname{cl}[\operatorname{cone}(C - x_0) - \operatorname{cone}(C - x_0)] = X$$
$$\Leftrightarrow \operatorname{cl}\operatorname{cone}(C - C) = X \Leftrightarrow 0 \in \operatorname{qi}(C - C).$$

This means that Theorem 2.7 is a generalization to the case of separated locally convex vector spaces of the separation theorem given in [5], [6] in the framework of real normed spaces.

The condition $x_0 \in C$ in Theorem 2.7 is essential (see [6]). However, if x_0 is an arbitrary element of X, we can also give a separation theorem by using the following result due to Cammaroto and Di Bella (Theorem 2.1 in [4]). The mentioned authors use this theorem in order to prove their strong duality result (Theorem 2.2 in [4]). Unfortunately, as we will show in section 4, this result has self-contradictory assumptions.

THEOREM 2.9 (see [4]). *Let S and T be nonempty convex subsets of X with $\operatorname{qri} S \neq \emptyset$, $\operatorname{qri} T \neq \emptyset$, and such that $\operatorname{cl}\operatorname{cone}(\operatorname{qri} S - \operatorname{qri} T)$ is not a linear subspace of X. Then there exists $x^* \in X^*$, $x^* \neq 0$, such that $\langle x^*, s\rangle \leq \langle x^*, t\rangle$ for all $s \in S$, $t \in T$.*

The following result is a direct consequence of Theorem 2.9.

COROLLARY 2.10. *Let C be a convex subset of X such that $\operatorname{qri} C \neq \emptyset$ and $\operatorname{cl}\operatorname{cone}(C - x_0)$ is not a linear subspace of X, where $x_0 \in X$. Then there exists $x^* \in X^*, x^* \neq 0$, such that $\langle x^*, x\rangle \leq \langle x^*, x_0\rangle$ for all $x \in C$.*

Proof. We take, in Theorem 2.9, $S := C$ and $T := \{x_0\}$. Then we apply Proposition 2.5 (iii) and (ix) to obtain the conclusion. □

3. Fenchel duality. In this section we give some new Fenchel duality results stated in terms of the quasi interior and quasi-relative interior, respectively.

Consider the convex optimization problem

$$(P_F) \quad \inf_{x\in X}\{f(x)+g(x)\},$$

where X is a separated locally convex vector space and $f, g : X \to \overline{\mathbb{R}}$ are proper convex functions such that $\operatorname{dom}(f) \cap \operatorname{dom}(g) \neq \emptyset$. The Fenchel dual problem to (P_F) is the followiing:

$$(D_F) \quad \sup_{x^*\in X^*}\{-f^*(-x^*)-g^*(x^*)\}.$$

We denote by $v(P_F)$ and $v(D_F)$ the optimal objective values of the primal and the dual problem, respectively. Weak duality always holds; that is, $v(D_F) \leq v(P_F)$. For strong duality, the case when $v(P_F) = v(D_F)$ and (D_F) has an optimal solution, several generalized interior-point regularity conditions were given in the literature. In order to recall them we need the following generalized interior notions. For a convex subset C of X we have:

- $\operatorname{core} C := \{x \in C : \operatorname{cone}(C - x) = X\}$, the *core* of C [14], [17];
- $\operatorname{icr} C := \{x \in C : \operatorname{cone}(C - x)$ is a linear subspace$\}$, the *intrinsic core* of C [1], [9], [17];
- $\operatorname{sqri} C := \{x \in C : \operatorname{cone}(C - x)$ is a closed linear subspace$\}$, the *strong quasi-relative interior* of C [2], [17].

We have the following inclusions:

$$\operatorname{core} C \subseteq \operatorname{sqri} C \subseteq \operatorname{qri} C \text{ and } \operatorname{core} C \subseteq \operatorname{qi} C \subseteq \operatorname{qri} C.$$

If X if finite-dimensional, then $\operatorname{qri} C = \operatorname{sqri} C = \operatorname{icr} C = \operatorname{ri} C$ [3], [8] and $\operatorname{core} C = \operatorname{qi} C = \operatorname{int} C$ [12], [14].

Consider now the following regularity conditions:

(i) $0 \in \operatorname{int}(\operatorname{dom}(f) - \operatorname{dom}(g))$;
(ii) $0 \in \operatorname{core}(\operatorname{dom}(f) - \operatorname{dom}(g))$ (cf. [14]);
(iii) $0 \in \operatorname{icr}(\operatorname{dom}(f) - \operatorname{dom}(g))$ and $\operatorname{aff}(\operatorname{dom}(f) - \operatorname{dom}(g))$ is a closed linear subspace (cf. [8]);
(iv) $0 \in \operatorname{sqri}(\operatorname{dom}(f) - \operatorname{dom}(g))$ (cf. [15]).

Let us notice that all of these conditions guarantee strong duality if we suppose the additional hypotheses that the functions f and g are lower semicontinuous and X is a Fréchet space. Between the above conditions we have the following relation: (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii) $\Leftrightarrow$ (iv) [8].

Trying to give a similar regularity condition for strong duality by means of the notion of the quasi-relative interior of a convex set, a natural question arises: Is the condition $0 \in \operatorname{qri}(\operatorname{dom}(f) - \operatorname{dom}(g))$ sufficient for strong duality? The following example (which can be found in [8]) gives us a negative answer, and this means that we need additional assumptions in order to guarantee Fenchel duality (see Theorem 3.5).

Example 3.1. As in [8], we consider $X = l^2$, the Hilbert space consisting of all sequences $x = (x_n)_{n\in\mathbb{N}}$ such that $\sum_{n=1}^{\infty} x_n^2 < \infty$. Consider also the sets

$$C = \{x \in l^2 : x_{2n-1} + x_{2n} = 0 \ \forall n \in \mathbb{N}\},$$

$$S = \{x \in l^2 : x_{2n} + x_{2n+1} = 0 \ \forall n \in \mathbb{N}\}.$$

The sets C and S are closed linear subspaces of l^2 and $C \cap S = \{0\}$. Define the functions $f, g : l^2 \to \overline{\mathbb{R}}$ by $f = \delta_C$ and $g(x) = x_1$ if $x \in S$ and $+\infty$ otherwise. One can see that f and g are proper, convex, and lower semicontinuous functions with $\mathrm{dom}(f) = C$ and $\mathrm{dom}(g) = S$. As was shown in [8], $v(P_F) = 0$ and $v(D_F) = -\infty$, so we have a duality gap between the optimal objective values of the primal problem and its Fenchel dual. Moreover, $S - C$ is dense in l^2; thus $\mathrm{cl}\,\mathrm{cone}(\mathrm{dom}(f) - \mathrm{dom}(g)) = \mathrm{cl}(C - S) = l^2$. The last relation implies that $0 \in \mathrm{qi}(\mathrm{dom}(f) - \mathrm{dom}(g))$, hence $0 \in \mathrm{qri}(\mathrm{dom}(f) - \mathrm{dom}(g))$.

Let us notice that if $v(P_F) = -\infty$, by the weak duality follows that also strong duality holds. This is the reason why we suppose in the following that $v(P_F) \in \mathbb{R}$.

LEMMA 3.2. *The following relation is always true:*

$$0 \in \mathrm{qri}(\mathrm{dom}(f) - \mathrm{dom}(g)) \Rightarrow (0,1) \in \mathrm{qri}[\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))].$$

Proof. One can see that $\widehat{\mathrm{epi}}(g - v(P_F)) = \{(x,r) \in X \times \mathbb{R} : r \leq -g(x) + v(P_F)\}$. Let us prove first that $(0,1) \in \mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))$. Since $\inf_{x \in X}[f(x) + g(x)] = v(P_F) < v(P_F) + 1$, there exists $x' \in X$ such that $f(x') + g(x') < v(P_F) + 1$. Then $(0,1) = (x', v(P_F) + 1 - g(x')) - (x', -g(x') + v(P_F)) \in \mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))$.

Now let $(x^*, r^*) \in N_{\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))}(0,1)$. We have

$$\langle x^*, x - x' \rangle + r^*(\mu - \mu' - 1) \leq 0 \ \forall (x,\mu) \in \mathrm{epi}(f) \ \forall (x',\mu') \in \widehat{\mathrm{epi}}(g - v(P_F)). \tag{5}$$

For $(x,\mu) := (x_0, f(x_0))$ and $(x',\mu') := (x_0, -g(x_0) + v(P_F) - 2)$ in (5), where $x_0 \in \mathrm{dom}(f) \cap \mathrm{dom}(g)$ is fixed, we get $r^*(f(x_0) + g(x_0) - v(P_F) + 1) \leq 0$, and hence $r^* \leq 0$. As $\inf_{x \in X}[f(x) + g(x)] = v(P_F) < v(P_F) + 1/2$, there exists $x_1 \in X$ such that $f(x_1) + g(x_1) < v(P_F) + 1/2$. By taking now $(x,\mu) := (x_1, f(x_1))$ and $(x',\mu') := (x_1, -g(x_1) + v(P_F) - 1/2)$ in (5) we obtain $r^*(f(x_1) + g(x_1) - v(P_F) - 1/2) \leq 0$, and so $r^* \geq 0$. Thus $r^* = 0$, and (5) gives: $\langle x^*, x - x' \rangle \leq 0$ for all $x \in \mathrm{dom}(f)$ for all $x' \in \mathrm{dom}(g)$. Hence $x^* \in N_{\mathrm{dom}(f) - \mathrm{dom}(g)}(0)$. Since $N_{\mathrm{dom}(f) - \mathrm{dom}(g)}(0)$ is a linear subspace of X^* (cf. Proposition 2.2), we have $\langle -x^*, x - x' \rangle \leq 0$ for all $x \in \mathrm{dom}(f)$ for all $x' \in \mathrm{dom}(g)$, and so $-(x^*, r^*) = (-x^*, 0) \in N_{\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))}(0,1)$, showing that $N_{\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))}(0,1)$ is a linear subspace of $X^* \times \mathbb{R}$. Hence, by applying again Proposition 2.2, we get $(0,1) \in \mathrm{qri}[\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))]$. □

PROPOSITION 3.3. *Assume that* $0 \in \mathrm{qi}[(\mathrm{dom}(f) - \mathrm{dom}(g)) - (\mathrm{dom}(f) - \mathrm{dom}(g))]$. *Then* $N_{\mathrm{co}[(\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))) \cup \{(0,0)\}]}(0,0)$ *is a linear subspace of* $X^* \times \mathbb{R}$ *if and only if* $N_{\mathrm{co}[(\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))) \cup \{(0,0)\}]}(0,0) = \{(0,0)\}$.

Proof. The sufficiency is trivial. In the following let us suppose that the set $N_{\mathrm{co}[(\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))) \cup \{(0,0)\}]}(0,0)$ is a linear subspace of $X^* \times \mathbb{R}$. Take $(x^*, r^*) \in N_{\mathrm{co}[(\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))) \cup \{(0,0)\}]}(0,0)$. Then

$$\langle x^*, x - x' \rangle + r^*(\mu - \mu') \leq 0 \ \forall (x,\mu) \in \mathrm{epi}(f) \ \forall (x',\mu') \in \widehat{\mathrm{epi}}(g - v(P_F)). \tag{6}$$

Let $x_0 \in \mathrm{dom}\, f \cap \mathrm{dom}(g)$ be fixed. By taking $(x,\mu) := (x_0, f(x_0)) \in \mathrm{epi}(f)$ and $(x',\mu') := (x_0, -g(x_0) + v(P_F) - 1/2) \in \widehat{\mathrm{epi}}(g - v(P_F))$ in the previous inequality, we get $r^*(f(x_0) + g(x_0) - v(P_F) + 1/2) \leq 0$, implying $r^* \leq 0$. As the set $N_{\mathrm{co}[(\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))) \cup \{(0,0)\}]}(0,0)$ is a linear subspace of $X^* \times \mathbb{R}$, the same argument applies also for $(-x^*, -r^*)$, implying $-r^* \leq 0$. In this way we get $r^* = 0$. The inequality (6) and the relation $(-x^*, 0) \in N_{\mathrm{co}[(\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))) \cup \{(0,0)\}]}(0,0)$ imply that

$$\langle x^*, x - x' \rangle = 0 \ \forall (x,\mu) \in \mathrm{epi}(f) \ \forall (x',\mu') \in \widehat{\mathrm{epi}}(g - v(P_F)),$$

which is nothing else than $\langle x^*, x - x' \rangle = 0$ for all $x \in \mathrm{dom}(f)$ for all $x' \in \mathrm{dom}(g)$, and thus $\langle x^*, x \rangle = 0$ for all $x \in \mathrm{dom}(f) - \mathrm{dom}(g)$. Since x^* is linear and continuous, the last relation implies that $\langle x^*, x \rangle = 0$ for all $x \in \mathrm{cl}\,\mathrm{cone}[(\mathrm{dom}(f) - \mathrm{dom}(g)) - (\mathrm{dom}(f) - \mathrm{dom}(g))] = X$; hence $x^* = 0$, and the conclusion follows. □

Remark 3.4. (a) By (1) one can see that $\mathrm{cl}\,\mathrm{cone}\,\mathrm{co}[(\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))) \cup \{(0,0)\}] = \mathrm{cl}\,\mathrm{cone}[\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))]$. Hence one has the following sequence of equivalences: $N_{\mathrm{co}[(\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))) \cup \{(0,0)\}]}(0,0)$ is a linear subspace of $X^* \times \mathbb{R} \Leftrightarrow (0,0) \in \mathrm{qri}\,\mathrm{co}[(\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))) \cup \{(0,0)\}] \Leftrightarrow \mathrm{cl}\,\mathrm{cone}\,\mathrm{co}[(\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))) \cup \{(0,0)\}]$ is a linear subspace of $X \times \mathbb{R} \Leftrightarrow \mathrm{cl}\,\mathrm{cone}(\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F)))$ is a linear subspace of $X \times \mathbb{R}$. The relation $N_{\mathrm{co}[(\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))) \cup \{(0,0)\}]}(0,0) = \{(0,0)\}$ is equivalent to $(0,0) \in \mathrm{qi}\,\mathrm{co}[(\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))) \cup \{(0,0)\}]$ (cf. Proposition 2.4), so in the case $0 \in \mathrm{qi}[(\mathrm{dom}(f) - \mathrm{dom}(g)) - (\mathrm{dom}(f) - \mathrm{dom}(g))]$ the conclusion of the previous proposition can be reformulated as follows:

$$\mathrm{cl}\,\mathrm{cone}(\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))) \text{ is a linear subspace of } X \times \mathbb{R} \Leftrightarrow$$
$$(0,0) \in \mathrm{qi}\,\mathrm{co}[(\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))) \cup \{(0,0)\}]$$

or, equivalently,

$$(0,0) \in \mathrm{qri}\,\mathrm{co}[(\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))) \cup \{(0,0)\}] \Leftrightarrow$$
$$(0,0) \in \mathrm{qi}\,\mathrm{co}[(\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))) \cup \{(0,0)\}].$$

(b) One can prove that the primal problem (P_F) has an optimal solution if and only if $(0,0) \in \mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))$. This means that if we suppose that the primal problem has an optimal solution and $0 \in \mathrm{qi}[(\mathrm{dom}(f) - \mathrm{dom}(g)) - (\mathrm{dom}(f) - \mathrm{dom}(g))]$, then the conclusion of the previous proposition can be rewritten as follows: $N_{(\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F)))}(0,0)$ is a linear subspace of $X^* \times \mathbb{R}$ if and only if $N_{(\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F)))}(0,0) = \{(0,0)\}$ or, equivalently,

$$(0,0) \in \mathrm{qri}[\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))] \Leftrightarrow (0,0) \in \mathrm{qi}[\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))].$$

We give now the first strong duality result for (P_F) and its Fenchel dual (D_F). Let us notice that for the functions f and g we suppose just convexity properties, and we do not use any closedness type of condition.

THEOREM 3.5. *Suppose that* $0 \in \mathrm{qi}[(\mathrm{dom}(f) - \mathrm{dom}(g)) - (\mathrm{dom}(f) - \mathrm{dom}(g))]$, $0 \in \mathrm{qri}(\mathrm{dom}(f) - \mathrm{dom}(g))$, *and* $(0,0) \notin \mathrm{qri}\,\mathrm{co}[(\mathrm{epi}\, f - \widehat{\mathrm{epi}}(g - v(P_F))) \cup \{(0,0)\}]$. *Then* $v(P_F) = v(D_F)$, *and* (D_F) *has an optimal solution.*

Proof. Lemma 3.2 ensures that $(0,1) \in \mathrm{qri}[\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))]$, and hence $\mathrm{qri}[\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))] \neq \emptyset$. The condition $(0,0) \notin \mathrm{qri}\,\mathrm{co}[(\mathrm{epi}\, f - \widehat{\mathrm{epi}}(g - v(P_F))) \cup \{(0,0)\}]$, together with the relation $\mathrm{cl}\,\mathrm{cone}\,\mathrm{co}[(\mathrm{epi}\, f - \widehat{\mathrm{epi}}(g - v(P_F))) \cup \{(0,0)\}] = \mathrm{cl}\,\mathrm{cone}[\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))]$, implies that $\mathrm{cl}\,\mathrm{cone}[\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))]$ is not a linear subspace of $X \times \mathbb{R}$. We apply Corollary 2.10 with $C := \mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))$ and $x_0 = (0,0)$. Thus there exists $(x^*, \lambda) \in X^* \times \mathbb{R}$, $(x^*, \lambda) \neq (0,0)$ such that

$$\langle x^*, x \rangle + \lambda\mu \geq \langle x^*, x' \rangle + \lambda\mu' \ \forall (x, \mu) \in \widehat{\mathrm{epi}}(g - v(P_F)) \ \forall (x', \mu') \in \mathrm{epi}(f). \tag{7}$$

We claim that $\lambda \leq 0$. Indeed, if $\lambda > 0$, then for $(x, \mu) := (x_0, -g(x_0) + v(P_F))$ and $(x', \mu') := (x_0, f(x_0) + n), n \in \mathbb{N}$, where $x_0 \in \mathrm{dom}(f) \cap \mathrm{dom}(g)$ is fixed, we obtain

from (7): $\langle x^*, x_0\rangle + \lambda(-g(x_0) + v(P_F)) \geq \langle x^*, x_0\rangle + \lambda(f(x_0) + n)$ for all $n \in \mathbb{N}$. By passing to the limit as $n \to +\infty$ we obtain a contradiction. Next we prove that $\lambda < 0$. Suppose that $\lambda = 0$. Then from (7) we have $\langle x^*, x\rangle \geq \langle x^*, x'\rangle$ for all $x \in \operatorname{dom}(g)$ for all $x' \in \operatorname{dom}(f)$, and hence $\langle x^*, x\rangle \leq 0$ for all $x \in \operatorname{dom}(f) - \operatorname{dom}(g)$. By using the second part of Theorem 2.7, we obtain $0 \notin \operatorname{qri}(\operatorname{dom}(f) - \operatorname{dom}(g))$, which contradicts the hypothesis. Thus we must have $\lambda < 0$, and so we obtain from (7):

$$\left\langle \frac{1}{\lambda}x^*, x \right\rangle + \mu \leq \left\langle \frac{1}{\lambda}x^*, x' \right\rangle + \mu', \forall (x, \mu) \in \widehat{\operatorname{epi}}(g - v(P_F)), \forall (x', \mu') \in \operatorname{epi}(f).$$

Let be $r \in \mathbb{R}$ such that

$$\mu' + \langle x_0^*, x'\rangle \geq r \geq \mu + \langle x_0^*, x\rangle \ \forall (x, \mu) \in \widehat{\operatorname{epi}}(g - v(P_F)) \ \forall (x', \mu') \in \operatorname{epi}(f),$$

where $x_0^* := \frac{1}{\lambda}x^*$. The first inequality shows that $f(x) \geq \langle -x_0^*, x\rangle + r$ for all $x \in X$, that is, $f^*(-x_0^*) \leq -r$. The second one gives us $-g(x) + v(P_F) + \langle x_0^*, x\rangle \leq r$ for all $x \in X$; hence $g^*(x_0^*) \leq r - v(P_F)$, and so we have $-f^*(-x_0^*) - g^*(x_0^*) \geq r + v(P_F) - r = v(P_F)$. This implies that $v(D_F) \geq v(P_F)$. As the opposite inequality is always true, we get $v(P_F) = v(D_F)$, and x_0^* is an optimal solution of the problem (D_F). □

The above theorem combined with Remark 3.4(b) gives us the following result.

COROLLARY 3.6. *Suppose that the primal problem (P_F) has an optimal solution, $0 \in \operatorname{qi}[(\operatorname{dom}(f) - \operatorname{dom}(g)) - (\operatorname{dom}(f) - \operatorname{dom}(g))]$, $0 \in \operatorname{qri}(\operatorname{dom}(f) - \operatorname{dom}(g))$, and $(0, 0) \notin \operatorname{qri}[\operatorname{epi}(f) - \widehat{\operatorname{epi}}(g - v(P_F))]$. Then $v(P_F) = v(D_F)$, and (D_F) has an optimal solution.*

Remark 3.7. The condition $0 \in \operatorname{qi}[(\operatorname{dom}(f) - \operatorname{dom}(g)) - (\operatorname{dom}(f) - \operatorname{dom}(g))]$ implies that

$$0 \in \operatorname{qri}(\operatorname{dom}(f) - \operatorname{dom}(g)) \Leftrightarrow 0 \in \operatorname{qi}(\operatorname{dom}(f) - \operatorname{dom}(g)).$$

Indeed, denote by $C := \operatorname{dom}(f) - \operatorname{dom}(g)$. Obviously $0 \in \operatorname{qi} C$ implies that $0 \in \operatorname{qri} C$. Suppose now that $0 \in \operatorname{qri} C$, and let $x^* \in N_C(0)$ be arbitrary. We have $\langle x^*, x\rangle \leq 0$ for all $x \in C$. Since $N_C(0)$ is a linear subspace of X^*, we obtain $\langle x^*, x\rangle = 0$ for all $x \in C$. We get further $\langle x^*, x\rangle = 0$ for all $x \in \operatorname{cl}\operatorname{cone}(C - C) = X$, which implies that $x^* = 0$. Thus $N_C(0) = \{0\}$, and the conclusion follows.

Some stronger versions of Theorem 3.5 and Corollary 3.6, respectively, follow.

THEOREM 3.8. *Suppose that $0 \in \operatorname{qi}(\operatorname{dom}(f) - \operatorname{dom}(g))$ and $(0, 0) \notin \operatorname{qri}\operatorname{co}[(\operatorname{epi}(f) - \widehat{\operatorname{epi}}(g - v(P_F))) \cup \{(0, 0)\}]$. Then $v(P_F) = v(D_F)$, and (D_F) has an optimal solution.*

Proof. We have $\operatorname{dom}(f) - \operatorname{dom}(g) \subseteq (\operatorname{dom}(f) - \operatorname{dom}(g)) - (\operatorname{dom}(f) - \operatorname{dom}(g))$, so the condition $0 \in \operatorname{qi}(\operatorname{dom}(f) - \operatorname{dom}(g))$ implies that $0 \in \operatorname{qi}[(\operatorname{dom}(f) - \operatorname{dom}(g)) - (\operatorname{dom}(f) - \operatorname{dom}(g))]$ and $0 \in \operatorname{qri}(\operatorname{dom}(f) - \operatorname{dom}(g))$. Then we apply Theorem 3.5 to obtain the conclusion. □

COROLLARY 3.9. *Suppose that the primal problem (P_F) has an optimal solution, $0 \in \operatorname{qi}(\operatorname{dom}(f) - \operatorname{dom}(g))$, and $(0, 0) \notin \operatorname{qri}[\operatorname{epi}(f) - \widehat{\operatorname{epi}}(g - v(P_F))]$. Then $v(P_F) = v(D_F)$, and (D_F) has an optimal solution.*

THEOREM 3.10. *Suppose that $\operatorname{dom}(f) \cap \operatorname{qri}\operatorname{dom}(g) \neq \emptyset$, $0 \in \operatorname{qi}(\operatorname{dom}(g) - \operatorname{dom}(g))$, and $(0, 0) \notin \operatorname{qri}\operatorname{co}[(\operatorname{epi}(f) - \widehat{\operatorname{epi}}(g - v(P_F))) \cup \{(0, 0)\}]$. Then $v(P_F) = v(D_F)$, and (D_F) has an optimal solution.*

Proof. We apply Lemma 2.6 with $A := \operatorname{dom}(g)$ and $B := \operatorname{dom}(f)$. We get $0 \in \operatorname{qi}(\operatorname{dom}(g) - \operatorname{dom}(f))$ or, equivalently, $0 \in \operatorname{qi}(\operatorname{dom}(f) - \operatorname{dom}(g))$. We obtain the result by applying Theorem 3.8. □

COROLLARY 3.11. *Suppose that the primal problem* (P_F) *has an optimal solution,* $\mathrm{dom}(f) \cap \mathrm{qri}\,\mathrm{dom}(g) \neq \emptyset$, $0 \in \mathrm{qi}(\mathrm{dom}(g) - \mathrm{dom}(g))$, *and* $(0,0) \notin \mathrm{qri}[\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))]$. *Then* $v(P_F) = v(D_F)$, *and* (D_F) *has an optimal solution.*

Remark 3.12. (a) We introduced above three new regularity conditions for Fenchel duality. As one can easily see from the proof of these results, the relation between these conditions is the following one: The regularity condition given in Theorem 3.10 (Corollary 3.11) implies the one given in Theorem 3.8 (Corollary 3.9), which implies the one given in Theorem 3.5 (Corollary 3.6).

(b) If we renounce the condition $(0,0) \notin \mathrm{qri}\,\mathrm{co}[(\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F))) \cup \{(0,0)\}]$, or, respectively, $(0,0) \notin \mathrm{qri}(\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F)))$, in the case when the primal problem has an optimal solution, then the duality results given above may fail. By using again Example 3.1 we show that these conditions are essential in our theory. Let us notice that for the problem in Example 3.1 the conditions $0 \in \mathrm{qi}[(\mathrm{dom}(f) - \mathrm{dom}(g)) - (\mathrm{dom}(f) - \mathrm{dom}(g))]$ and $0 \in \mathrm{qri}(\mathrm{dom}(f) - \mathrm{dom}(g))$ are fulfilled. We prove in the following that in the aforementioned example we have $(0,0) \in \mathrm{qri}(\mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F)))$. Note that the scalar product on l^2, $\langle \cdot, \cdot \rangle : l^2 \times l^2 \to \mathbb{R}$, is given by $\langle x, y \rangle = \sum_{n=1}^{\infty} x_n y_n$ for all $x = (x_n)_{n \in \mathbb{N}}, y = (y_n)_{n \in \mathbb{N}} \in l^2$. Also, for $k \in \mathbb{N}$, we denote by $e^{(k)}$ the element in l^2 which has on the kth position 1 and on the other positions 0, that is, $e_n^{(k)} = 1$, if $n = k$ and $e_n^{(k)} = 0$, for all $n \in \mathbb{N} \backslash \{k\}$. We have $\mathrm{epi}(f) = C \times [0, \infty)$. Further, $\widehat{\mathrm{epi}}(g - v(P_F)) = \{(x,r) \in l^2 \times \mathbb{R} : r \leq -g(x)\} = \{(x,r) \in l^2 \times \mathbb{R} : x = (x_n)_{n \in \mathbb{N}} \in S, r \leq -x_1\} = \{(x, -x_1 - \varepsilon) \in l^2 \times \mathbb{R} : x = (x_n)_{n \in \mathbb{N}} \in S, \varepsilon \geq 0\}$. Then $A := \mathrm{epi}(f) - \widehat{\mathrm{epi}}(g - v(P_F)) = \{(x - x', x_1' + \varepsilon) : x \in C, x' = (x_n')_{n \in \mathbb{N}} \in S, \varepsilon \geq 0\}$. Take $(x^*, r) \in N_A(0,0)$, where $x^* = (x_n^*)_{n \in \mathbb{N}}$. We have

$$\langle x^*, x - x' \rangle + r(x_1' + \varepsilon) \leq 0 \ \forall x \in C \ \forall x' = (x_n')_{n \in \mathbb{N}} \in S \ \forall \varepsilon \geq 0. \tag{8}$$

By taking in (8) $x' = 0$ and $\varepsilon = 0$ we get $\langle x^*, x \rangle \leq 0$ for all $x \in C$. As C is a linear subspace of X we have

$$\langle x^*, x \rangle = 0 \ \forall x \in C. \tag{9}$$

Since $e^{(2k-1)} - e^{(2k)} \in C$, for all $k \in \mathbb{N}$, the relation (9) implies that

$$x_{2k-1}^* - x_{2k}^* = 0 \ \forall k \in \mathbb{N}. \tag{10}$$

From (8) and (9) we obtain

$$\langle -x^*, x' \rangle + r(x_1' + \varepsilon) \leq 0 \ \forall x' = (x_n')_{n \in \mathbb{N}} \in S \ \forall \varepsilon \geq 0. \tag{11}$$

By taking $\varepsilon = 0$ and $x' := me^1 \in S$ in (11), where $m \in \mathbb{Z}$ is arbitrary, we get $m(-x_1^* + r) \leq 0$ for all $m \in \mathbb{Z}$, and thus $r = x_1^*$. For $\varepsilon = 0$ in (11) we obtain $-\sum_{n=1}^{\infty} x_n^* x_n' + r x_1' \leq 0$ for all $x' \in S$. By taking into account that $r = x_1^*$, we get $-\sum_{n=2}^{\infty} x_n^* x_n' \leq 0$ for all $x' \in S$. As S is a linear subspace of X it follows that $\sum_{n=2}^{\infty} x_n^* x_n' = 0$ for all $x' \in S$, but, since $e^{(2k)} - e^{(2k+1)} \in S$ for all $k \in \mathbb{N}$, the above relation shows that

$$x_{2k}^* - x_{2k+1}^* = 0 \ \forall k \in \mathbb{N}. \tag{12}$$

By combining (10) with (12) we get $x^* = 0$ (since $x^* \in l^2$). Because $r = x_1^*$, we also have $r = 0$. Thus $N_A(0,0) = \{(0,0)\}$, and Proposition 2.4 gives us the desired conclusion.

(c) Since in all of the strong duality results given above one must have that $0 \in \operatorname{qi}[(\operatorname{dom}(f) - \operatorname{dom}(g)) - (\operatorname{dom}(f) - \operatorname{dom}(g))]$, in view of Remark 3.4, the condition $(0,0) \notin \operatorname{qri}\operatorname{co}[(\operatorname{epi}(f) - \widehat{\operatorname{epi}}(g - v(P_F))) \cup \{(0,0)\}]$ (respectively, $(0,0) \notin \operatorname{qri}[\operatorname{epi}(f) - \widehat{\operatorname{epi}}(g - v(P_F))]$) is equivalent to $(0,0) \notin \operatorname{qi}\operatorname{co}[(\operatorname{epi}(f) - \widehat{\operatorname{epi}}(g - v(P_F))) \cup \{(0,0)\}]$ (respectively, $(0,0) \notin \operatorname{qi}[\operatorname{epi}(f) - \widehat{\operatorname{epi}}(g - v(P_F))]$).

(d) We have the following relation:

$$(0,0) \in \operatorname{qi}\operatorname{co}[(\operatorname{epi}(f) - \widehat{\operatorname{epi}}(g - v(P_F))) \cup \{(0,0)\}] \Rightarrow 0 \in \operatorname{qi}(\operatorname{dom}(f) - \operatorname{dom}(g)).$$

Indeed, $(0,0) \in \operatorname{qi}\operatorname{co}[(\operatorname{epi}(f) - \widehat{\operatorname{epi}}(g - v(P_F))) \cup \{(0,0)\}] \Leftrightarrow \operatorname{cl}\operatorname{cone}\operatorname{co}[(\operatorname{epi}(f) - \widehat{\operatorname{epi}}(g - v(P_F))) \cup \{(0,0)\}] = X \times \mathbb{R}$; hence $\operatorname{cl}\operatorname{cone}(\operatorname{epi}(f) - \widehat{\operatorname{epi}}(g - v(P_F))) = X \times \mathbb{R}$. Since $\operatorname{cl}\operatorname{cone}(\operatorname{epi}(f) - \widehat{\operatorname{epi}}(g - v(P_F))) \subseteq \operatorname{cl}\operatorname{cone}(\operatorname{dom}(f) - \operatorname{dom}(g)) \times \mathbb{R}$, this implies that $\operatorname{cl}\operatorname{cone}(\operatorname{dom}(f) - \operatorname{dom}(g)) = X$, that is, $0 \in \operatorname{qi}(\operatorname{dom}(f) - \operatorname{dom}(g))$. Hence

$$0 \notin \operatorname{qi}(\operatorname{dom}(f) - \operatorname{dom}(g)) \Rightarrow (0,0) \notin \operatorname{qi}\operatorname{co}[(\operatorname{epi}(f) - \widehat{\operatorname{epi}}(g - v(P_F))) \cup \{(0,0)\}].$$

Nevertheless, in the regularity conditions given above one cannot substitute the condition $(0,0) \notin \operatorname{qri}\operatorname{co}[(\operatorname{epi}(f) - \widehat{\operatorname{epi}}(g - v(P_F))) \cup \{(0,0)\}]$ by the "nice-looking" one $0 \notin \operatorname{qi}(\operatorname{dom}(f) - \operatorname{dom}(g))$, since in all three strong duality theorems the other hypotheses we consider imply that $0 \in \operatorname{qi}(\operatorname{dom}(f) - \operatorname{dom}(g))$ (cf. Remark 3.7).

Example 3.13. Consider again the space $X = l^2$ equipped with the norm $\|\cdot\| : l^2 \to \mathbb{R}$, $\|x\|^2 = \sum_{n=1}^{\infty} x_n^2$ for all $x = (x_n)_{n\in\mathbb{N}} \in l^2$. We define the functions $f, g : l^2 \to \overline{\mathbb{R}}$ by

$$f(x) = \begin{cases} \|x\| & \text{if } x \in x_0 - l^2_+, \\ +\infty & \text{otherwise} \end{cases}$$

and

$$g(x) = \begin{cases} \langle c, x\rangle & \text{if } x \in l^2_+, \\ +\infty & \text{otherwise,} \end{cases}$$

where $l^2_+ = \{(x_n)_{n\in\mathbb{N}} \in l^2 : x_n \geq 0, \forall n \in \mathbb{N}\}$ is the positive cone, $x_0 = (\frac{1}{n})_{n\in\mathbb{N}}$, and $c = (\frac{1}{2^n})_{n\in\mathbb{N}}$. Note that $v(P_F) = \inf_{x\in l^2}\{f(x) + g(x)\} = 0$, and the infimum is attained at $x = 0$. We have $\operatorname{dom}(f) = x_0 - l^2_+ = \{(x_n)_{n\in\mathbb{N}} \in l^2 : x_n \leq \frac{1}{n}, \forall n \in \mathbb{N}\}$ and $\operatorname{dom}(g) = l^2_+$. Since $\operatorname{qri} l^2_+ = \{(x_n)_{n\in\mathbb{N}} \in l^2 : x_n > 0, \forall n \in \mathbb{N}\}$ (cf. [3]), we get $\operatorname{dom}(f) \cap \operatorname{qri}\operatorname{dom}(g) = \{(x_n)_{n\in\mathbb{N}} \in l^2 : 0 < x_n \leq \frac{1}{n}, \forall n \in \mathbb{N}\} \neq \emptyset$. Also, $\operatorname{cl}\operatorname{cone}(\operatorname{dom}(g) - \operatorname{dom}(g)) = l^2$, so $0 \in \operatorname{qi}(\operatorname{dom}(g) - \operatorname{dom}(g))$. Further, $\operatorname{epi}(f) = \{(x,r) \in l^2 \times \mathbb{R} : x \in x_0 - l^2_+, \|x\| \leq r\} = \{(x, \|x\| + \varepsilon) \in l^2 \times \mathbb{R} : x \in x_0 - l^2_+, \varepsilon \geq 0\}$ and $\widehat{\operatorname{epi}}(g - v(P_F)) = \{(x,r) \in l^2 \times \mathbb{R} : r \leq -g(x)\} = \{(x,r) \in l^2 \times \mathbb{R} : r \leq -\langle c, x\rangle, x \in l^2_+\} = \{(x, -\langle c, x\rangle - \varepsilon) : x \in l^2_+, \varepsilon \geq 0\}$. We get $\operatorname{epi}(f) - \widehat{\operatorname{epi}}(g - v(P_F)) = \{(x - x', \|x\| + \varepsilon + \langle c, x'\rangle + \varepsilon') : x \in x_0 - l^2_+, x' \in l^2_+, \varepsilon, \varepsilon' \geq 0\} = \{(x - x', \|x\| + \langle c, x'\rangle + \varepsilon) : x \in x_0 - l^2_+, x' \in l^2_+, \varepsilon \geq 0\}$.

In the following we prove that $(0,0) \notin \operatorname{qri}(\operatorname{epi}(f) - \widehat{\operatorname{epi}}(g - v(P_F)))$. By assuming the contrary we would have that the set $\operatorname{cl}(\operatorname{cone}(\operatorname{epi}(f) - \widehat{\operatorname{epi}}(g - v(P_F))))$ is a linear subspace. Since $(0,1) \in \operatorname{cl}(\operatorname{cone}(\operatorname{epi}(f) - \widehat{\operatorname{epi}}(g - v(P_F))))$ (take $x = x' = 0$ and $\varepsilon = 1$) we must have that also $(0,-1)$ belongs to this set. On the other hand, one can easily see that for all (x,r) belonging to $\operatorname{cl}(\operatorname{cone}(\operatorname{epi}(f) - \widehat{\operatorname{epi}}(g - v(P_F))))$ it holds that $r \geq 0$. This leads to the desired contradiction.

Hence the conditions of Corollary 3.11 are fulfilled, and thus strong duality holds. Let us notice that the regularity conditions given in Corollaries 3.6 and 3.9 are also fulfilled (see Remark 3.12(a)).

On the other hand, l^2 is a Fréchet space (being a Hilbert space), the functions f and g are lower semicontinuous, and, as $\operatorname{sqri}(\operatorname{dom}(f) - \operatorname{dom}(g)) = \operatorname{sqri}(x_0 - l^2_+) = \emptyset$, none of the constraint qualifications (i)–(iv) presented in the beginning of this section can be applied for this optimization problem.

As for all $x^* \in l^2$ it holds that

$$g^*(x^*) = \begin{cases} 0 & \text{if } x^* \in c - l^2_+, \\ +\infty & \text{otherwise} \end{cases}$$

and (see Theorem 2.8.7 in [17])

$$f^*(-x^*) = \inf_{x_1^*+x_2^*=-x^*} \{\|\cdot\|^*(x_1^*) + \delta^*_{x_0-l^2_+}(x_2^*)\} = \inf_{\substack{x_1^*+x_2^*=-x^*, \\ \|x_1^*\|\leq 1, x_2^* \in l^2_+}} \{\langle x_2^*, x_0\rangle\},$$

the optimal objective value of the Fenchel dual problem is

$$\begin{aligned} v(D_F) &= \sup_{x^* \in X^*} \{-f^*(-x^*) - g^*(x^*)\} \\ &= \sup_{\substack{x_2^* \in l^2_+ - c - x_1^*, \\ \|x_1^*\|\leq 1, x_2^* \in l^2_+}} \{\langle -x_2^*, x_0\rangle\} = \sup_{x_2^* \in l^2_+} \{\langle -x_2^*, x_0\rangle\} = 0, \end{aligned}$$

and $x_2^* = 0$ is the optimal solution of the dual.

In the following, by using the results introduced above, we give regularity conditions for the following convex optimization problem:

$$(P_A) \quad \inf_{x \in X} \{f(x) + (g \circ A)(x)\},$$

where X and Y are separated locally convex vector spaces with their topological dual spaces X^* and Y^*, respectively, $A : X \to Y$ is a linear continuous mapping, $f : X \to \overline{\mathbb{R}}$, and $g : Y \to \overline{\mathbb{R}}$ are proper convex functions such that $A(\operatorname{dom}(f)) \cap \operatorname{dom}(g) \neq \emptyset$. The Fenchel dual problem to (P_A) is (cf. [17])

$$(D_A) \quad \sup_{y^* \in Y^*} \{-f^*(-A^*y^*) - g^*(y^*)\},$$

where $A^* : Y^* \to X^*$ is the *adjoint operator* of A, defined in the usual way: $\langle A^*y^*, x\rangle = \langle y^*, Ax\rangle$ for all $(y^*, x) \in Y^* \times X$. We denote by $v(P_A)$ and $v(D_A)$ the optimal objective values of the primal and the dual problem, respectively. We suppose also that $v(P_A) \in \mathbb{R}$. In the following theorem the set

$$A \times \mathrm{id}_{\mathbb{R}}(\operatorname{epi}(f)) = \{(Ax, r) \in Y \times \mathbb{R} : f(x) \leq r\}$$

is the image of $\operatorname{epi}(f)$ through the operator $A \times \mathrm{id}_{\mathbb{R}}$.

THEOREM 3.14. *Suppose that* $0 \in \operatorname{qi}[(A(\operatorname{dom}(f)) - \operatorname{dom}(g)) - (A(\operatorname{dom}(f)) - \operatorname{dom}(g))]$, $0 \in \operatorname{qri}(A(\operatorname{dom}(f)) - \operatorname{dom}(g))$, *and* $(0,0) \notin \operatorname{qri}\operatorname{co}[(A \times \mathrm{id}_{\mathbb{R}}(\operatorname{epi}(f)) - \widehat{\operatorname{epi}}(g - v(P_A))) \cup \{(0,0)\}]$. *Then* $v(P_A) = v(D_A)$, *and* (D_A) *has an optimal solution.*

Proof. Let us introduce the following functions: $F, G : X \times Y \to \overline{\mathbb{R}}$, $F(x,y) = f(x) + \delta_{\{x \in X : Ax = y\}}(x)$, and $G(x,y) = g(y)$. The functions F and G are proper and convex, and $\inf_{(x,y)\in X\times Y}[F(x,y) + G(x,y)] = \inf_{x\in X}\{f(x) + (g \circ A)(x)\} = v(P_A)$. Moreover, $\operatorname{dom}(F) = \operatorname{dom}(f) \times A(\operatorname{dom}(f))$ and $\operatorname{dom}(G) = X \times \operatorname{dom}(g)$, so $\operatorname{dom}(F) \cap \operatorname{dom}(G) \neq \emptyset$. Further,

$$\operatorname{dom}(F) - \operatorname{dom}(G) = X \times (A(\operatorname{dom}(f)) - \operatorname{dom}(g)).$$

By combining the last relation with the hypotheses, we obtain $(0,0) \in \mathrm{qi}[(\mathrm{dom}(F) - \mathrm{dom}(G)) - (\mathrm{dom}(F) - \mathrm{dom}(G))]$ and $(0,0) \in \mathrm{qri}(\mathrm{dom}(F) - \mathrm{dom}(G))$. Since $\mathrm{epi}(F) = \{(x, Ax, r) : f(x) \leq r\}$ and $\widehat{\mathrm{epi}}(G - v(P_A)) = \{(x, y, r) : r \leq -G(x,y) + v(P_A)\} = X \times \widehat{\mathrm{epi}}(g - v(P_A))$, we obtain

$$\mathrm{epi}(F) - \widehat{\mathrm{epi}}(G - v(P_A)) = X \times (A \times \mathrm{id}_{\mathbb{R}}(\mathrm{epi}(f)) - \widehat{\mathrm{epi}}(g - v(P_A))),$$

and this means that $(0,0,0) \notin \mathrm{qri}\,\mathrm{co}[(\mathrm{epi}(F) - \widehat{\mathrm{epi}}(G - v(P_A))) \cup \{(0,0,0)\}]$. Theorem 3.5 yields for F and G:

$$\inf_{(x,y)\in X\times Y}[F(x,y) + G(x,y)] = \max_{(x^*,y^*)\in X^*\times Y^*}\{-F^*(-x^*,-y^*) - G^*(x^*,y^*)\}.$$

On the other hand, $F^*(x^*,y^*) = f^*(x^* + A^*y^*)$ for all $(x^*,y^*) \in X^* \times Y^*$, and

$$G^*(x^*,y^*) = \begin{cases} g^*(y^*) & \text{if } x^* = 0, \\ +\infty & \text{otherwise.} \end{cases}$$

Therefore, $\max_{(x^*,y^*)\in X^*\times Y^*}\{-F^*(-x^*,-y^*) - G^*(x^*,y^*)\} = \max_{y^*\in Y^*}\{-f^*(-A^*y^*) - g^*(y^*)\}$, and the conclusion follows. □

COROLLARY 3.15. *Suppose that the primal problem (P_A) has an optimal solution, $0 \in \mathrm{qi}[(A(\mathrm{dom}(f)) - \mathrm{dom}(g)) - (A(\mathrm{dom}(f)) - \mathrm{dom}(g))]$, $0 \in \mathrm{qri}(A(\mathrm{dom}(f)) - \mathrm{dom}(g))$, and $(0,0) \notin \mathrm{qri}[A \times \mathrm{id}_{\mathbb{R}}(\mathrm{epi}(f)) - \widehat{\mathrm{epi}}(g - v(P_A))]$. Then $v(P_A) = v(D_A)$, and (D_A) has an optimal solution.*

THEOREM 3.16. *Suppose that $0 \in \mathrm{qi}(A(\mathrm{dom}(f)) - \mathrm{dom}(g))$ and $(0,0) \notin \mathrm{qri}\,\mathrm{co}[(A \times \mathrm{id}_{\mathbb{R}}(\mathrm{epi}(f)) - \widehat{\mathrm{epi}}(g - v(P_A))) \cup \{(0,0)\}]$. Then $v(P_A) = v(D_A)$, and (D_A) has an optimal solution.*

Proof. By considering the functions F and G from the proof of Theorem 3.14, we have $\mathrm{cl}\,\mathrm{cone}(\mathrm{dom}(F) - \mathrm{dom}(G)) = X \times \mathrm{cl}\,\mathrm{cone}(A(\mathrm{dom}(f)) - \mathrm{dom}(g)) = X \times Y$, and thus $(0,0) \in \mathrm{qi}(\mathrm{dom}(F) - \mathrm{dom}(G))$. Also we have $(0,0,0) \notin \mathrm{qri}\,\mathrm{co}[(\mathrm{epi}(F) - \widehat{\mathrm{epi}}(G - v(P_A))) \cup \{(0,0,0)\}]$. Theorem 3.8 yields for F and G:

$$\inf_{(x,y)\in X\times Y}[F(x,y) + G(x,y)] = \max_{(x^*,y^*)\in X^*\times Y^*}\{-F^*(-x^*,-y^*) - G^*(x^*,y^*)\},$$

and the conclusion follows. □

COROLLARY 3.17. *Suppose that the primal problem (P_A) has an optimal solution, $0 \in \mathrm{qi}(A(\mathrm{dom}(f)) - \mathrm{dom}(g))$, and $(0,0) \notin \mathrm{qri}[A \times \mathrm{id}_{\mathbb{R}}(\mathrm{epi}(f)) - \widehat{\mathrm{epi}}(g - v(P_A))]$. Then $v(P_A) = v(D_A)$, and (D_A) has an optimal solution.*

THEOREM 3.18. *Suppose that $A(\mathrm{dom}(f)) \cap \mathrm{qri}\,\mathrm{dom}(g) \neq \emptyset$, $0 \in \mathrm{qi}(\mathrm{dom}(g) - \mathrm{dom}(g))$ and $(0,0) \notin \mathrm{qri}\,\mathrm{co}[(A \times \mathrm{id}_{\mathbb{R}}(\mathrm{epi}(f)) - \widehat{\mathrm{epi}}(g - v(P_A))) \cup \{(0,0)\}]$. Then $v(P_A) = v(D_A)$, and (D_A) has an optimal solution.*

Proof. Consider again the functions F and G defined as in the proof of Theorem 3.14. We have $\mathrm{dom}(F) \cap \mathrm{qri}\,\mathrm{dom}(G) = (\mathrm{dom}(f) \times (A(\mathrm{dom}(f))) \cap (X \times \mathrm{qri}\,\mathrm{dom}(g))) = \mathrm{dom}(f) \times (A(\mathrm{dom}(f)) \cap \mathrm{qri}\,\mathrm{dom}(g)) \neq \emptyset$. Also, $\mathrm{cl}\,\mathrm{cone}(\mathrm{dom}(G) - \mathrm{dom}(G)) = X \times \mathrm{cl}\,\mathrm{cone}(\mathrm{dom}(g) - \mathrm{dom}(g)) = X \times Y$, and hence $(0,0) \in \mathrm{qi}(\mathrm{dom}(G) - \mathrm{dom}(G))$. Moreover, $(0,0,0) \notin \mathrm{qri}\,\mathrm{co}[(\mathrm{epi}(F) - \widehat{\mathrm{epi}}(G - v(P_A))) \cup \{(0,0,0)\}]$. Theorem 3.10 yields for F and G:

$$\inf_{(x,y)\in X\times Y}[F(x,y) + G(x,y)] = \max_{(x^*,y^*)\in X^*\times Y^*}\{-F^*(-x^*,-y^*) - G^*(x^*,y^*)\},$$

and the conclusion follows. □

COROLLARY 3.19. *Suppose that the primal problem* (P_A) *has an optimal solution,* $A(\mathrm{dom}(f)) \cap \mathrm{qri}\,\mathrm{dom}(g) \neq \emptyset$, $0 \in \mathrm{qi}(\mathrm{dom}(g) - \mathrm{dom}(g))$, *and* $(0,0) \notin \mathrm{qri}[A \times \mathrm{id}_{\mathbb{R}}(\mathrm{epi}(f)) - \widehat{\mathrm{epi}}(g - v(P_A))]$. *Then* $v(P_A) = v(D_A)$, *and* (D_A) *has an optimal solution.*

4. Lagrange duality. By using an approach due to Magnanti (cf. [13]), in this section we derive from the results in the previous section some duality results concerning the Lagrange dual problem. We work in the following setting. Let X be a real linear topological space and S a nonempty subset of X. Let Y be a separated locally convex space partially ordered by a convex cone C. Let $f : S \to \mathbb{R}$ and $g : S \to Y$ be two functions such that the function $(f,g) : S \to \mathbb{R} \times Y$, defined by $(f,g)(x) = (f(x), g(x))$, for all $x \in S$, is convexlike with respect to the cone $\mathbb{R}_+ \times C \subseteq \mathbb{R} \times Y$; that is, the set $(f,g)(S) + \mathbb{R}_+ \times C$ is convex. Let us notice that this property implies that the sets $f(S) + [0,\infty)$ and $g(S) + C$ are convex (the reverse implication does not always hold). Consider the optimization problem

$$(P_L) \qquad \inf_{\substack{x \in S \\ g(x) \in -C}} f(x),$$

where the constraint set $T = \{x \in S : g(x) \in -C\}$ is assumed to be nonempty. The Lagrange dual problem associated to (P_L) is

$$(D_L) \qquad \sup_{\lambda \in C^*} \inf_{x \in S} [f(x) + \langle \lambda, g(x) \rangle],$$

where $C^* = \{x^* \in X^* : \langle x^*, x \rangle \geq 0, \forall x \in C\}$ is the *dual cone* of C. Let us denote by $v(P_L)$ and $v(D_L)$ the optimal objective values of the primal and the dual problem, respectively. A regularity condition for strong duality between (P_L) and (D_L) was proposed in Theorem 2.2 in [4]. We show first that this theorem has self-contradictory assumptions. To this end we prove the following lemma.

LEMMA 4.1. *Suppose that* $\mathrm{cl}(C - C) = Y$ *and there exists* $\overline{x} \in S$ *such that* $g(\overline{x}) \in -\mathrm{qri}\, C$. *Then the following assertions are true:*

(a) $0 \in \mathrm{qi}(g(S) + C)$;

(b) $\mathrm{cl}\,\mathrm{cone}[\mathrm{qri}(g(S) + C)]$ *is a linear subspace of* Y.

Proof. (a) We apply Lemma 2.6 with $A := -C$ and $B := g(S) + C$. The condition $\mathrm{cl}(C - C) = Y$ implies that $0 \in \mathrm{qi}(A - A)$, while the Slater-type condition $g(\overline{x}) \in -\mathrm{qri}\, C$ ensures that $g(\overline{x}) \in \mathrm{qri}\, A \cap B$. Hence, by Lemma 2.6 we obtain $0 \in \mathrm{qi}(A - B)$, that is, $0 \in \mathrm{qi}(-g(S) - C)$, which is nothing else than $0 \in \mathrm{qi}(g(S) + C)$.

(b) From (a) it follows that $0 \in \mathrm{qri}(g(S) + C)$. By applying Proposition 2.5(vii) we get $0 \in \mathrm{qri}(\mathrm{qri}(g(S) + C))$, which is nothing else than $\mathrm{cl}\,\mathrm{cone}[\mathrm{qri}(g(S) + C)]$ is a linear subspace of Y. □

In order to get strong duality between (P_L) and (D_L) in Theorem 2.2 in [4] the authors ask that the following hypotheses are fulfilled: $\mathrm{cl}(C - C) = Y$, there exists $\overline{x} \in S$ such that $g(\overline{x}) \in -\mathrm{qri}\, C$, $\mathrm{qri}(g(S) + C) \neq \emptyset$, and $\mathrm{cl}\,\mathrm{cone}[\mathrm{qri}(g(S) + C)]$ is not a linear subspace of Y. The previous lemma proves that these assumptions are in contradiction.

Next we prove some Lagrange duality results written in terms of the quasi interior and quasi-relative interior, respectively. As in the previous section, we may suppose that $v(P_L)$ is a real number.

Consider the following convex set:

$$\mathcal{E}_{v(P_L)} = \{(f(x) + \alpha - v(P_L), g(x) + y) : x \in S, \alpha \geq 0, y \in C\} \subseteq \mathbb{R} \times Y.$$

Let us notice that the set $-\mathcal{E}_{v(P_L)}$ is in analogy with the *conic extension*, a notion used by Giannessi in the theory of image space analysis (see [7]). One can easily prove that the primal problem (P_L) has an optimal solution if and only if $(0,0) \in \mathcal{E}_{v(P_L)}$. Let us introduce the functions $f_1, f_2 : \mathbb{R} \times Y \to \overline{\mathbb{R}}$,

$$f_1(y_0, y_1) = \begin{cases} y_0 & \text{if } (y_0, y_1) \in \mathcal{E}_{v(P_L)} + (v(P_L), 0), \\ +\infty & \text{otherwise,} \end{cases}$$

and $f_2 = \delta_{\mathbb{R}\times(-C)}$. It holds that

$$\operatorname{dom}(f_1) - \operatorname{dom}(f_2) = \mathbb{R} \times (g(S) + C). \tag{13}$$

Moreover, as pointed out by Magnanti (cf. [13]), we have

$$\inf_{(y_0,y_1)\in\mathbb{R}\times Y} \{f_1(y_0, y_1) + f_2(y_0, y_1)\} = \inf_{\substack{x\in S \\ g(x)\in -C}} f(x) = v(P_L) \tag{14}$$

and

$$\sup_{(y_0^*,y_1^*)\in\mathbb{R}\times Y^*} \{-f_1^*(-y_0^*, -y_1^*) - f_2^*(y_0^*, y_1^*)\} = \sup_{\lambda\in C^*} \inf_{x\in S} [f(x) + \langle \lambda, g(x)\rangle] = v(D_L). \tag{15}$$

With this approach, we can derive from the strong duality results given for Fenchel duality corresponding strong duality results for Lagrange duality.

THEOREM 4.2. *Suppose that* $0 \in \operatorname{qi}[(g(S) + C) - (g(S) + C)]$, $0 \in \operatorname{qri}(g(S) + C)$, *and* $(0,0) \notin \operatorname{qri}\operatorname{co}(\mathcal{E}_{v(P_L)} \cup \{(0,0)\})$. *Then* $v(P_L) = v(D_L)$, *and* (D_L) *has an optimal solution.*

Proof. The hypotheses of the theorem and (13) imply that the conditions $(0,0) \in \operatorname{qi}[(\operatorname{dom}(f_1) - \operatorname{dom}(f_2)) - (\operatorname{dom}(f_1) - \operatorname{dom}(f_2))]$ and $(0,0) \in \operatorname{qri}(\operatorname{dom}(f_1) - \operatorname{dom}(f_2))$ are fulfilled. Further, $\operatorname{epi}(f_1) = \{(y_0, y_1, r) \in \mathbb{R}\times Y\times\mathbb{R} : (y_0, y_1) \in \mathcal{E}_{v(P_L)} + (v(P_L), 0), y_0 \le r\} = \{(f(x)+\alpha, g(x)+y, r) : x \in S, \alpha \ge 0, y \in C, f(x)+\alpha \le r\}$, and $\widehat{\operatorname{epi}}(f_2 - v(P_L)) = \{(y_0, y_1, r) \in \mathbb{R} \times Y \times \mathbb{R} : r \le -f_2(y_0, y_1) + v(P_L)\} = \{(y_0, y_1, r) \in \mathbb{R} \times Y \times \mathbb{R} : y_0 \in \mathbb{R}, y_1 \in -C, r \le v(P_L)\} = \mathbb{R} \times (-C) \times (-\infty, v(P_L)]$. Thus $\operatorname{epi}(f_1) - \widehat{\operatorname{epi}}(f_2 - v(P_L)) = \operatorname{epi}(f_1) + \mathbb{R} \times C \times [-v(P_L), +\infty) = \{(f(x) + \alpha + a, g(x) + y, r - v(P_L) + \varepsilon) : x \in S, \alpha \ge 0, a \in \mathbb{R}, y \in C, \varepsilon \ge 0, f(x) + \alpha \le r\} = \{(f(x) + \alpha + a, g(x) + y, f(x) + \alpha + \varepsilon - v(P_L)) : x \in S, \alpha \ge 0, a \in \mathbb{R}, y \in C, \varepsilon \ge 0\}$, and this means that

$$\operatorname{epi}(f_1) - \widehat{\operatorname{epi}}(f_2 - v(P_L)) = \mathbb{R} \times \{(g(x) + y, f(x) + \alpha - v(P_L)) : x \in S, \alpha \ge 0, y \in C\}.$$

Thus $(0,0,0) \in \operatorname{qri}\operatorname{co}[(\operatorname{epi}(f_1) - \widehat{\operatorname{epi}}(f_2 - v(P_L))) \cup \{(0,0,0)\}]$ if and only if $(0,0) \in \operatorname{qri}\operatorname{co}(\mathcal{E}_{v(P_L)} \cup \{(0,0)\})$. Now we can apply Theorem 3.5 for f_1 and f_2, and we obtain

$$\inf_{(y_0,y_1)\in\mathbb{R}\times Y} \{f_1(y_0, y_1) + f_2(y_0, y_1)\} = \max_{(y_0^*,y_1^*)\in\mathbb{R}\times Y^*} \{-f_1^*(-y_0^*, -y_1^*) - f_2^*(y_0^*, y_1^*)\}.$$

By (14) and (15) the conclusion follows. □

COROLLARY 4.3. *Suppose that the primal problem* (P_L) *has an optimal solution,* $0 \in \operatorname{qi}[(g(S) + C) - (g(S) + C)]$, $0 \in \operatorname{qri}(g(S) + C)$, *and* $(0,0) \notin \operatorname{qri}\mathcal{E}_{v(P_L)}$. *Then* $v(P_L) = v(D_L)$, *and* (D_L) *has an optimal solution.*

Further, like for Fenchel duality, other Lagrange duality results can be stated.

THEOREM 4.4. *Suppose that* $0 \in \operatorname{qi}(g(S)+C)$ *and* $(0,0) \notin \operatorname{qri}\operatorname{co}(\mathcal{E}_{v(P_L)}\cup\{(0,0)\})$. *Then* $v(P_L) = v(D_L)$, *and* (D_L) *has an optimal solution.*

Proof. This is a direct consequence of the previous theorem since $g(S)+C \subseteq (g(S)+C)-(g(S)+C)$, and so the condition $0 \in \operatorname{qi}(g(S)+C)$ implies that $0 \in \operatorname{qi}[(g(S)+C)-(g(S)+C)]$ and $0 \in \operatorname{qri}(g(S)+C)$. □

COROLLARY 4.5. *Suppose that the primal problem (P_L) has an optimal solution, $0 \in \operatorname{qi}(g(S)+C)$, and $(0,0) \notin \operatorname{qri}\mathcal{E}_{v(P_L)}$. Then $v(P_L)=v(D_L)$, and (D_L) has an optimal solution.*

THEOREM 4.6. *Suppose that $\operatorname{cl}(C-C)=Y$ and there exists $\overline{x} \in S$ such that $g(\overline{x}) \in -\operatorname{qri} C$. If $(0,0) \notin \operatorname{qri}\operatorname{co}(\mathcal{E}_{v(P_L)} \cup \{(0,0)\})$, then $v(P_L)=v(D_L)$, and (D_L) has an optimal solution.*

Proof. The condition $(0,0) \notin \operatorname{qri}\operatorname{co}(\mathcal{E}_{v(P_L)} \cup \{(0,0)\})$ implies that $(0,0,0) \notin \operatorname{qri}\operatorname{co}[(\operatorname{epi}(f_1) - \widehat{\operatorname{epi}}(f_2 - v(P_L))) \cup \{(0,0,0)\}]$ (cf. the proof of Theorem 4.2). Further, we have $\operatorname{dom}(f_1) \cap \operatorname{qri}\operatorname{dom}(f_2) = [\mathcal{E}_{v(P_L)} + (v(P_L),0)] \cap \operatorname{qri}(\mathbb{R} \times (-C)) = [\mathcal{E}_{v(P_L)} + (v(P_L),0)] \cap [\mathbb{R} \times (-\operatorname{qri} C)]$. From the Slater-type condition we get that $(f(\overline{x}), g(\overline{x})) \in [\mathcal{E}_{v(P_L)} + (v(P_L),0)] \cap [\mathbb{R} \times (-\operatorname{qri} C)]$, and hence $\operatorname{dom}(f_1) \cap \operatorname{qri}\operatorname{dom}(f_2) \neq \emptyset$. Moreover, $\operatorname{cl}\operatorname{cone}(\operatorname{dom}(f_2) - \operatorname{dom}(f_2)) = \operatorname{cl}\operatorname{cone}[\mathbb{R} \times (C-C)] = \mathbb{R} \times \operatorname{cl}(C-C) = \mathbb{R} \times Y$, and hence $(0,0) \in \operatorname{qi}(\operatorname{dom}(f_2) - \operatorname{dom}(f_2))$. By Theorem 3.10 for f_1 and f_2 we obtain

$$\inf_{(y_0,y_1) \in \mathbb{R} \times Y} \{f_1(y_0,y_1) + f_2(y_0,y_1)\} = \max_{(y_0^*,y_1^*) \in \mathbb{R} \times Y^*} \{-f_1^*(-y_0^*,-y_1^*) - f_2^*(y_0^*,y_1^*)\},$$

and by using again (14) and (15) the conclusion follows. □

COROLLARY 4.7. *Suppose that the primal problem (P_L) has an optimal solution, $\operatorname{cl}(C-C)=Y$, and there exists $\overline{x} \in S$ such that $g(\overline{x}) \in -\operatorname{qri} C$. If $(0,0) \notin \operatorname{qri}\mathcal{E}_{v(P_L)}$, then $v(P_L)=v(D_L)$, and (D_L) has an optimal solution.*

Remark 4.8. Let us notice that from the above results one can derive duality theorems for the case when, in the set of constraints, one has also equalities defined by affine functions. Indeed, consider the optimization problem

$$(P_L^{aff}) \quad \inf_{\substack{x \in S \\ g(x) \in -C \\ h(x)=0}} f(x),$$

where $h: X \to Z$ is an affine mapping and Z is a real normed space (the hypotheses regarding the functions f and g remain the same as in the beginning of this section). The Lagrange dual problem associated to (P_L^{aff}) is

$$(D_L^{aff}) \quad \sup_{\substack{\lambda \in C^* \\ \mu \in Z^*}} \inf_{x \in S} [f(x) + \langle \lambda, g(x) \rangle + \langle \mu, h(x) \rangle],$$

where Z^* is the topological dual space of Z.

By using Theorems 4.2 and 4.4 one can formulate Lagrange duality theorems for (P_L^{aff}) and (D_L^{aff}) by noticing that the primal problem can be reformulated as

$$\inf_{\substack{x \in S \\ g(x) \in -C \\ h(x)=0}} f(x) = \inf_{\substack{x \in S \\ u(x) \in -(C \times \{0\})}} f(x),$$

where $u: S \to Y \times Z$, $u(x) = (g(x), h(x))$. For the optimization problem with equality and cone constraints some regularity conditions have been given in [5] by using the notion of the quasi-relative interior. Along them in the strong duality theorem (Theorem 3.1 in [5]) a "separation assumption," called by the authors *Assumption S*, is imposed. Unfortunately, this assumption is not only a sufficient condition for having

strong duality, as claimed in the paper, but actually an equivalent formulation of this situation (this makes the other regularity conditions inoperative). More than that, in the proof of Theorem 3.1 in [5] a mistake occurred, namely, in the relation after inequality (8) when trying to prove the "nonverticality" of the separating hyperplane.

The approach we propose above offers a viable alternative for dealing with Lagrange duality for this class of optimization problems.

Acknowledgments. The authors are thankful to two anonymous reviewers for comments and remarks which have improved the quality of the paper.

REFERENCES

[1] J. M. Borwein and R. Goebel, *Notions of relative interior in Banach spaces*, J. Math. Sci. (N. Y.), 115 (2003), pp. 2542–2553.
[2] J. M. Borwein, V. Jeyakumar, A. S. Lewis, and H. Wolkowicz, *Constrained Approximation via Convex Programming*, preprint, University of Waterloo, 1988.
[3] J. M. Borwein and A. S. Lewis, *Partially finite convex programming, part* I: *Quasi relative interiors and duality theory*, Math. Program., 57 (1992), pp. 15–48.
[4] F. Cammaroto and B. Di Bella, *Separation theorem based on the quasirelative interior and application to duality theory*, J. Optim. Theory Appl., 125 (2005), pp. 223–229.
[5] P. Daniele and S. Giuffrè, *General infinite dimensional duality theory and applications to evolutionary network equilibrium problems*, Optim. Lett., 1 (2007), pp. 227–243.
[6] P. Daniele, S. Giuffrè, G. Idone, and A. Maugeri, *Infinite dimensional duality and applications*, Math. Ann., 339 (2007), pp. 221–239.
[7] F. Giannessi, *Constrained Optimization and Image Space Analysis, Vol.* 1. *Separation of Sets and Optimality Conditions*, Math. Concepts Methods Sci. Engrg. 49, Springer, New York, 2005.
[8] M. S. Gowda and M. Teboulle, *A comparison of constraint qualifications in infinite-dimensional convex programming*, SIAM J. Control Optim., 28 (1990), pp. 925–935.
[9] R. B. Holmes, *Geometric Functional Analysis*, Springer, Berlin, 1975.
[10] J. Jahn, *Introduction to the Theory of Nonlinear Optimization*, Springer, Berlin, 1996.
[11] V. Jeyakumar and H. Wolkowicz, *Generalizations of Slater's constraint qualification for infinite convex programs*, Math. Program., 57 (1992), pp. 85–101.
[12] M. A. Limber and R.K. Goodrich, *Quasi interiors, Lagrange multipliers, and L^p spectral estimation with lattice bounds*, J. Optim. Theory Appl., 78 (1993), pp. 143–161.
[13] T. L. Magnanti, *Fenchel and Lagrange duality are equivalent*, Math. Program., 7 (1974), pp. 253–258.
[14] R. T. Rockafellar, *Conjugate Duality and Optimization*, Conference Board of the Mathematical Sciences Regional Conference Series in Applied Mathematics 16, Society for Industrial and Applied Mathematics, Philadelphia, 1974.
[15] B. Rodrigues, *The Fenchel duality theorem in Fréchet spaces*, Optimization, 21 (1990), pp. 13–22.
[16] C. Zălinescu, *A comparison of constraint qualifications in infinite-dimensional convex programming revisited*, J. Aust. Math. Soc. Ser. B, 40 (1999), pp. 353–378.
[17] C. Zălinescu, *Convex Analysis in General Vector Spaces*, World Scientific, Singapore, 2002.

SIAM J. OPTIM.
Vol. 19, No. 1, pp. 234–239

A SHORT ALGEBRAIC PROOF OF THE FARKAS LEMMA*

DAVID BARTL†

Abstract. The purpose of this paper is to present a generalization of the Farkas lemma with a short algebraic proof. The generalization lies in the fact that we formulate the Farkas lemma in the setting of two vector spaces over a common linearly ordered field where one of the vector spaces is also linearly ordered. At the end of the paper, we mention the key theorem and two theorems of the alternative, namely Motzkin's theorem and Tucker's theorem.

Key words. Farkas lemma, systems of linear inequalities, theorems of the alternative, linearly ordered vector spaces, Motzkin's theorem, Tucker's theorem, key theorem

AMS subject classifications. 15A39, 06F20, 12J15

DOI. 10.1137/06067438

1. Introduction.

There are many proofs of the Farkas lemma [9], which is usually formulated in the setting of a finite-dimensional real vector space $\mathbb{R}^n$. Broyden [3] distinguishes three classes of proofs: algebraic, algorithmic, and geometric. The Farkas lemma is often the starting point when proving the duality theorem for linear programming (see, e.g., [18]) or when proving some other theorems of the alternative. Overviews of such theorems can be found in [21, 10, 12, 11, 4, 5, 3, 2]; see also [8, 19]. Out of the mentioned works, only the paper of Fan [10], the book of Chernikov [4], and the paper of Bartl [2] deal with the topic of linear inequalities in the setting of a general vector space, whose dimension may be infinite and no topology or any additional structure is assumed on it. While Fan [10] works within a real vector space, Chernikov [4] and Bartl [2] assume the more general setting of a vector space over a linearly ordered field.

Confining our attention to the Farkas lemma and speaking in broad terms, we need a "base" vector space, a linearly ordered field, and the additive group of the field with its ordering to formulate the Farkas lemma given by Chernikov [4, Lemma 2.4, p. 119]. But Bartl [2] showed that it is possible to substitute any linearly ordered vector space instead of the additive group of the field, thus generalizing the result. The resulting duality theorem for linear programming is also presented in [2].

Though the author's proof of the Farkas lemma [2, Lemma 4.1] is elementary and its main idea is easy to comprehend, the proof is rather long. In this paper, we give another proof—which is based on a different idea and, moreover, is significantly shorter—of the same result.

In addition to the algebraic approach, which we shall use in the following sections to prove the result, and as already mentioned above, other approaches are also possible. See, e.g., [6, 18] or [22, 20, 1, 15] for the algorithmic or geometric approach, respectively. Though the algebraic approach allows us to obtain quite a general version of the Farkas lemma, the geometric one permits a very illustrative exposition—see [7], where the geometric approach is applied within a finite-dimensional space. This is a certain "trade-off" between the approaches that has to be considered.

*Received by the editors August 14, 2006; accepted for publication (in revised form) November 14, 2007; published electronically March 5, 2008.
http://www.siam.org/journals/siopt/19-1/66743.html
†Department of Mathematics, Faculty of Science, University of Ostrava, 30. dubna 22, 701 03 Ostrava, Czech Republic (bartl@osu.cz).

2. Notation. Throughout this paper, the symbol F denotes a linearly ordered field; we do *not* assume the commutativity of the field so that F may even be a skew field. The symbol W denotes a vector space over the field F. The dimension of the vector space W may be finite or infinite and we do not assume any other structure (such as topology) on it; it plays the role of the underlying "base" vector space. The symbol V stands for a linearly ordered vector space over the linearly ordered field F. In what follows, we shall see it plays the role of the vector space of "objective values." The ordering of the (possibly skew) field F and that of the space V will be denoted by the symbols "$\leq$" and "$\preceq$," respectively.

Recall that the following five statements must hold true for all $u, v \in V$, and any scalar $\lambda \in F$ so that V is a linearly ordered vector space over the linearly ordered field F: first, $u \preceq v$ if and only if $u - v \preceq 0$; second, $u \succeq 0$ or $u \preceq 0$; third, if $u \succeq 0$ and $u \preceq 0$, then $u = 0$; fourth, if $u \succeq 0$ and $v \succeq 0$, then $u + v \succeq 0$; fifth, if $\lambda \geq 0$ and $u \succeq 0$, then $\lambda u \succeq 0$. It is possible to substitute the additive group of the field F with its ordering "$\leq$" for the space V and its ordering "$\preceq$." Hence, five analogous statements must also hold true when assuming $u, v \in F$, and writing "$\leq$" instead of "$\preceq$."

We shall use the symbol γ to denote a given linear mapping $\gamma\colon W \to V$. It assigns an "objective value" $\gamma(x)$ to any point x of the "base" vector space W.

Now, let m be a nonnegative natural number. Then F^m is the (left) vector space over the (possibly skew) field F of all m-component column vectors whose entries come from the field F. Analogously, the V^m is the space of all m-component columns with entries from the space V. If $\boldsymbol{\lambda} \in F^m$ or $\boldsymbol{u} \in V^m$, then the ith component of the respective column is λ_i or u_i for $i = 1, \ldots, m$. We symbolically write $\boldsymbol{\lambda} = (\lambda_i)_{i=1}^m$ and $\boldsymbol{u} = (u_i)_{i=1}^m$. The transposition of a column is indicated by the letter T in the superscript so that $\boldsymbol{\lambda}^T$ and $\boldsymbol{u}^T$ are the respective rows. The symbol $\boldsymbol{o}$ denotes the origin of the space F^m or V^m (depending on the context) and inequalities like $\boldsymbol{u} \succeq \boldsymbol{o}$, $\boldsymbol{\lambda} \geq \boldsymbol{o}$, $\boldsymbol{\lambda} > \boldsymbol{o}$, etc. are to be understood componentwise.

The symbol A stands for a linear mapping $A\colon W \to F^m$. If $x \in W$, then Ax is a column vector and its ith component is $\alpha_i(x)$ for $i = 1, \ldots, m$. It follows that $\alpha_1, \ldots, \alpha_m$ are the respective linear forms on the vector space W and we have $A = (\alpha_i)_{i=1}^m$.

Let $u \in V$ be a given vector. The linear mapping assigning the scalar multiple of the vector u to any scalar $\mu \in F$ is denoted by writing the Greek letter ι before u so that we have the linear mapping $\iota u\colon F \to V$ and $\iota u(\mu) = \mu u$ for all $\mu \in F$. Analogously, if ι is written before a given scalar $\lambda \in F$, then we have the linear mapping $\iota\lambda\colon F \to F$ and $\iota\lambda(\mu) = \mu\lambda$ for all $\mu \in F$. (As the field F may be skew, it is essential that the given scalar λ is multiplied by the scalar μ from the left.)

Having a vector $u \in V$ and a linear form α defined on the vector space W, then $\iota u\alpha$ denotes the composition of the mappings $\alpha\colon W \to F$ and $\iota u\colon F \to V$. Similarly, given yet a scalar $\lambda \in F$, then $\iota u \iota\lambda\alpha$ is the composition of the form α with $\iota\lambda\colon F \to F$ and ιu.

If the letter ι stands before a row $\boldsymbol{u}^T$, where $\boldsymbol{u} \in V^m$, then it is to be inserted into the row and put before each of its entries. The row can be multiplied by a column vector (from the right) in the usual manner. We obtain the linear mapping $\iota\boldsymbol{u}^T\colon F^m \to V$ and we have $\iota\boldsymbol{u}^T\boldsymbol{\mu} = \mu_1 u_1 + \cdots + \mu_m u_m$ for all $\boldsymbol{\mu} \in F^m$. The same applies when ι stands before a row $\boldsymbol{\lambda}^T$, where $\boldsymbol{\lambda} \in F^m$; we then obtain the linear mapping $\iota\boldsymbol{\lambda}^T\colon F^m \to F$.

Consequently, given a linear mapping $A\colon W \to F^m$, then $\iota\boldsymbol{u}^T A\colon W \to V$ or $\iota\boldsymbol{\lambda}^T A\colon W \to F$, where $\boldsymbol{u} \in V^m$ and $\boldsymbol{\lambda} \in F^m$, is the respective composed linear map-

ping. We have $\iota\boldsymbol{u}^T A = \iota u_1\alpha_1 + \cdots + \iota u_m\alpha_m$ and $\iota\boldsymbol{u}^T Ax = (\alpha_1(x))u_1 + \cdots + (\alpha_m(x))u_m$ for any $x \in W$. Analogous formulas could be written for the mapping $\iota\boldsymbol{\lambda}^T A$.

3. The Farkas lemma. We state the generalized version of the Farkas lemma [2, Lemma 4.1] now. Putting $V = F$, i.e., substituting the additive group of the field F with its ordering for the linearly ordered vector space V, we obtain the formulation of the Farkas lemma given by Chernikov [4, Lemma 2.4, p. 119]. Substituting the field of reals $\mathbb{R}$ with the standard ordering for the linearly ordered field F, still assuming $V = F$, we obtain the Farkas lemma formulated in the setting of a real (possibly infinite-dimensional) vector space. In this setting, Fan [10, Theorem 4] proved Haar's theorem (cf. [13, 14]) which—still within this setting, but not in general—is a strengthening of the Farkas lemma. When $V = F = \mathbb{R}$ and the "base" vector space is finite-dimensional, $W = \mathbb{R}^n$, we then obtain the classical formulation of the Farkas lemma [9, section IV]. Substituting the vector space $\mathbb{R}^N$ with the lexicographic ordering for V, while the dimension of W may be finite or infinite, we then obtain the lexicographic version of the Farkas lemma [2, Theorem 1.5].

LEMMA 1 (Farkas lemma). *Let $A\colon W \to F^m$ and $\gamma\colon W \to V$ be linear mappings. Then the implication $Ax \preceq \boldsymbol{o} \Rightarrow \gamma(x) \preceq 0$ or*

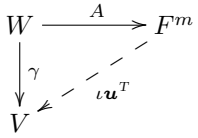
$$\alpha_1(x) \le 0 \wedge \cdots \wedge \alpha_m(x) \le 0 \implies \gamma(x) \preceq 0 \tag{1}$$

holds for all $x \in W$ if and only if

$$\exists u_1, \dots, u_m \succeq 0\colon\ \iota u_1\alpha_1 + \cdots + \iota u_m\alpha_m = \gamma\,, \tag{2}$$

i.e., $\iota\boldsymbol{u}^T A = \gamma$ for some $\boldsymbol{u} \in V^m$ satisfying $\boldsymbol{u} \succeq \boldsymbol{o}$ componentwise.

While the implication $\Leftarrow$ of Lemma 1 is trivial, the implication $\Rightarrow$ yields the following commutative diagram:



In words, if the given linear mappings A and γ satisfy implication (1) for all $x \in W$, then there exists a linear mapping $\iota\boldsymbol{u}^T\colon F^m \to V$ that makes the diagram commute; in addition, we have $\boldsymbol{u} \succeq \boldsymbol{o}$ componentwise.

We prove an alternative formulation of the Farkas lemma, which is given below. It can be easily seen that both formulations, i.e., Lemmas 1 and 2, are logically equivalent.

LEMMA 2 (Farkas lemma). *Let $A\colon W \to F^m$ and $\gamma\colon W \to V$ be linear mappings. Then either* (A) *there exists an $x \in W$ such that $\alpha_1(x) \ge 0, \dots, \alpha_m(x) \ge 0$, and $\gamma(x) \prec 0$, or* (B) *there exist nonnegative vectors $u_1, \dots, u_m \in V$ such that $\gamma = \iota u_1\alpha_1 + \cdots + \iota u_m\alpha_m$. The alternatives* (A) *and* (B) *exclude each other.*

Proof. We proceed by induction. The statement is trivial for $m = 0$ because either (A) there exists an $x \in W$ such that $\gamma(x) \ne 0$, in which case we may assume w.l.o.g. that $\gamma(x) \prec 0$, or (B) it holds that $\gamma(x) = 0$ for all $x \in W$.

Let us assume that the statement is valid for a nonnegative natural number m. We prove the assertion for $m + 1$. We have to prove that either (A) there exists an $x \in W$ such that $\alpha_1(x) \ge 0, \dots, \alpha_m(x) \ge 0$, $\alpha_{m+1}(x) \ge 0$, and $\gamma(x) \prec 0$, or (B) there exist vectors $u_1, \dots, u_m, u_{m+1} \succeq 0$ such that $\gamma = \iota u_1\alpha_1 + \cdots + \iota u_m\alpha_m + \iota u_{m+1}\alpha_{m+1}$.

By the induction hypothesis, either (a) there exists an $x \in W$ such that $\alpha_1(x) \geq 0$, $\ldots$, $\alpha_m(x) \geq 0$, and $\gamma(x) \prec 0$, or (b) there exist vectors $u_1, \ldots, u_m \succeq 0$ such that $\gamma = \iota u_1 \alpha_1 + \cdots + \iota u_m \alpha_m$.

We are finished in the case (b) for it suffices to put $u_{m+1} = 0$ so that the case (B) arises. In the case (a), there are two cases: either (aa) it holds that $\alpha_{m+1}(x) \geq 0$, or (ab) it holds that $\alpha_{m+1}(x) < 0$. We are finished again in the case (aa) because the case (A) arises. It remains to consider the case (ab) in the rest of the proof.

Having an $\bar{x} := x$ such that $\alpha_1(\bar{x}) \geq 0, \ldots, \alpha_m(\bar{x}) \geq 0$, $\alpha_{m+1}(\bar{x}) < 0$, and $\gamma(\bar{x}) \prec 0$, it is easy to see that the system

$$\alpha_1(x) \geq 0 \wedge \cdots \wedge \alpha_m(x) \geq 0 \wedge \alpha_{m+1}(x) \geq 0 \quad \text{and} \quad \gamma(x) \prec 0 \tag{3}$$

has a solution—therefore, the case (A) arises—if and only if the system

$$\alpha_1(x) \geq 0 \wedge \cdots \wedge \alpha_m(x) \geq 0 \wedge \alpha_{m+1}(x) = 0 \quad \text{and} \quad \gamma(x) \prec 0 \tag{4}$$

has a solution.

We find $\lambda_1, \ldots, \lambda_m \in F$ and a $v \in V$ so that $\alpha_i(\bar{x}) - \iota\lambda_i\alpha_{m+1}(\bar{x}) = 0$ and $\gamma(\bar{x}) - \iota v\alpha_{m+1}(\bar{x}) = 0$. We have $\lambda_i = (\alpha_{m+1}(\bar{x}))^{-1}\alpha_i(\bar{x})$ and $v = (\alpha_{m+1}(\bar{x}))^{-1}\gamma(\bar{x})$. As $\alpha_{m+1}(\bar{x}) < 0$, further $\alpha_i(\bar{x}) \geq 0$ and $\gamma(\bar{x}) \prec 0$, it is obvious that $\lambda_i \leq 0$ for $i = 1, \ldots, m$ and that $v \succ 0$.

Let us denote $\alpha'_i = \alpha_i - \iota\lambda_i\alpha_{m+1}$ and $\gamma' = \gamma - \iota v\alpha_{m+1}$. Hence $\alpha'_i(\bar{x}) = 0$ and $\gamma'(\bar{x}) = 0$, where $i = 1, \ldots, m$. It is very easy to see that the last system (4) has a solution—equivalently, the case (A) arises—if and only if the system

$$\alpha'_1(x) \geq 0 \wedge \cdots \wedge \alpha'_m(x) \geq 0 \wedge \alpha_{m+1}(x) = 0 \quad \text{and} \quad \gamma'(x) \prec 0 \tag{5}$$

has a solution.

Using the induction hypothesis again, we obtain that either (aba) there exists an $x' \in W$ such that $\alpha'_1(x') \geq 0, \ldots, \alpha'_m(x') \geq 0$, and $\gamma(x') \prec 0$, or (abb) there exist vectors $u'_1, \ldots, u'_m \succeq 0$ such that $\gamma' = \iota u'_1\alpha'_1 + \cdots + \iota u'_m\alpha'_m$.

Assume the case (aba) first. To solve system (5), we are looking for a point $x \in W$ such that $\alpha_{m+1}(x) = 0$ and $\alpha'_i(x) = \alpha'_i(x')$ for $i = 1, \ldots, m$ and $\gamma'(x) = \gamma'(x')$. We utilize the fact that $\alpha'_i(\bar{x}) = 0$ for $i = 1, \ldots, m$ and $\gamma'(\bar{x}) = 0$. Considering the point $x := x' - \mu\bar{x}$, where $\mu = (\alpha_{m+1}(x'))(\alpha_{m+1}(\bar{x}))^{-1}$, we can see that the case (A) arises.

In the remaining case (abb), we have

$$\begin{aligned} \gamma' &= \iota u'_1\alpha'_1 + \cdots + \iota u'_m\alpha'_m\,, \\ \gamma - \iota v\alpha_{m+1} &= \iota u'_1\alpha_1 - \iota u'_1\iota\lambda_1\alpha_{m+1} + \cdots + \iota u'_m\alpha_m - \iota u'_m\iota\lambda_m\alpha_{m+1}\,, \\ \gamma &= \iota u'_1\alpha_1 + \cdots + \iota u'_m\alpha_m + \iota(v - \lambda_1 u'_1 - \cdots - \lambda_m u'_m)\alpha_{m+1}\,. \end{aligned}$$

Hence, putting $u_i := u'_i$ for $i = 1, \ldots, m$ and $u_{m+1} = v - \lambda_1 u'_1 - \cdots - \lambda_m u'_m$, it is obvious that $u_{m+1} \succ 0$, and we can see that the case (B) arises. □

The presented proof of the Farkas lemma is a modified proof of a lemma due to Tucker [21, Lemma, p. 5]. Tucker himself notes that he uses an argument adapted from a certain unpublished proof by Gale. Indeed, the proof is quite similar to Gale's later published proof of the Farkas lemma [11, Theorem 2.6, p. 44]. Though Tucker [21] notes that he could also work with any linearly ordered field, both Tucker [21] and Gale [11] assume the setting of a finite-dimensional vector space only.

Close to Tucker's original lemma [21, Lemma, p. 5] is the following result. We prove it as a consequence of Lemma 2.

LEMMA 3. *Let $A\colon W \to F^m$ be a linear mapping. Then for any $i = 1, \dots, m$ there exist a componentwise nonnegative vector $\boldsymbol{\lambda} \in F^m$ and a point $x \in W$ such that*

$$\iota\boldsymbol{\lambda}^T A = o\,, \qquad Ax \geq \boldsymbol{o}, \qquad \text{and} \qquad \lambda_i + \alpha_i(x) > 0\,,$$

where o is the zero linear form $o\colon W \to F$.

Proof. Choose an $i = 1, \dots, m$. We use Lemma 2, putting $V = F$, i.e., substituting the additive group of the field F with its ordering for the linearly ordered vector space V, and $\gamma = -\alpha_i$ in it. Then either there exists an $x \in W$ such that $\alpha_1(x) \geq 0, \dots, \alpha_{i-1}(x) \geq 0$, $\alpha_{i+1}(x) \geq 0, \dots, \alpha_m(x) \geq 0$, and $\alpha_i(x) > 0$, or there exist $\lambda_1, \dots, \lambda_{i-1}, \lambda_{i+1}, \dots, \lambda_m \geq 0$ such that $\iota\lambda_1\alpha_1 + \dots + \iota\lambda_{i-1}\alpha_{i-1} + \iota\lambda_{i+1}\alpha_{i+1} + \dots + \iota\lambda_m\alpha_m = -\alpha_i$. Put $\boldsymbol{\lambda} = \boldsymbol{o}$ in the first case, and put $x = 0$ and $\lambda_i = 1$ in the latter one. ∎

4. Key theorem, Motzkin's theorem, and Tucker's theorem. In this section, we briefly mention the key theorem [21, Theorem 1], [12, section 4], [3, Theorem 1.2], [2, Theorem 5.3] and two of its consequences: Motzkin's theorem [16], [17, Theorem D6, p. 60], [21, Corollary 2A part (ii)], [2, Theorem 5.1] and Tucker's theorem [21, Corollary 2A part (i)], [2, Theorem 5.2].

THEOREM 4 (key theorem). *Let $A\colon W \to F^m$ be a linear mapping. Then there exist a componentwise nonnegative column vector $\boldsymbol{\lambda} \in F^m$ and a point $x \in W$ so that*

$$\iota\boldsymbol{\lambda}^T A = o\,, \qquad Ax \geq \boldsymbol{o}, \qquad \text{and} \qquad \boldsymbol{\lambda} + Ax > \boldsymbol{o}\,,$$

where o is the zero linear form $o\colon W \to F$.

Proof. By Lemma 3, there exist points $x_1, \dots, x_m \in W$ and nonnegative columns $\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_m \in F^m$ such that

$$\iota\boldsymbol{\lambda}_i^T A = o\,, \qquad Ax_i \geq \boldsymbol{o}, \qquad \text{and} \qquad \lambda_{ii} + \alpha_i(x_i) > 0$$

for $i = 1, \dots, m$. So it suffices to put $x = x_1 + \dots + x_m$ and $\boldsymbol{\lambda} = \boldsymbol{\lambda}_1 + \dots + \boldsymbol{\lambda}_m$. ∎

The two subsequent theorems involve two linear mappings, $A\colon W \to F^m$ and $B\colon W \to F^n$, where m and n are nonnegative natural numbers. Hence, either of the mappings A or B may be null. The alternatives (A) and (B) exclude each other in both theorems, and both Theorems 5 and 6 share the initial part of their proofs.

THEOREM 5 (Motzkin's theorem). *Either* (A) *there exists an $x \in W$ such that $Ax > \boldsymbol{o}$, $Bx \geq \boldsymbol{o}$, or* (B) *there exist $\boldsymbol{\lambda} \in F^m$ and $\boldsymbol{\mu} \in F^n$ satisfying $\boldsymbol{\lambda} \geq \boldsymbol{o}$, $\boldsymbol{\lambda} \neq \boldsymbol{o}$, $\boldsymbol{\mu} \geq \boldsymbol{o}$, and such that $\iota\boldsymbol{\lambda}^T A + \iota\boldsymbol{\mu}^T B = o$.*

THEOREM 6 (Tucker's theorem). *Either* (A) *there exists an $x \in W$ such that $Ax \geq \boldsymbol{o}$, $Ax \neq \boldsymbol{o}$, $Bx \geq \boldsymbol{o}$, or* (B) *there exist $\boldsymbol{\lambda} \in F^m$ and $\boldsymbol{\mu} \in F^n$ satisfying $\boldsymbol{\lambda} > \boldsymbol{o}$, $\boldsymbol{\mu} \geq \boldsymbol{o}$, and such that $\iota\boldsymbol{\lambda}^T A + \iota\boldsymbol{\mu}^T B = o$.*

Proof. By Theorem 4, there exist an $x \in W$ and componentwise nonnegative $\boldsymbol{\lambda} \in F^m$ and $\boldsymbol{\mu} \in F^n$ such that

$$\begin{matrix} Ax \geq \boldsymbol{o}\,, \\ Bx \geq \boldsymbol{o}\,, \end{matrix} \qquad \iota\boldsymbol{\lambda}^T A + \iota\boldsymbol{\mu}^T B = o, \qquad \text{and} \qquad \begin{matrix} \boldsymbol{\lambda} + Ax > \boldsymbol{o}\,, \\ \boldsymbol{\mu} + Bx > \boldsymbol{o}\,. \end{matrix}$$

If $\boldsymbol{\lambda} = \boldsymbol{o}$, then $Ax > \boldsymbol{o}$; we have thus obtained Theorem 5. If $Ax = \boldsymbol{o}$, then $\boldsymbol{\lambda} > \boldsymbol{o}$; we have then obtained Theorem 6. ∎

As in the previous section, both of the presented proofs are modifications of Tucker's proofs [21, Theorem 1, Corollary 2A]. To conclude, it is interesting that the ideas of the proofs by Gale [11] and Tucker [21] can be used to obtain the currently presented more general results.

Acknowledgments. The author acknowledges that the idea to apply Tucker's original proof to prove Lemma 2 arose during his discussion about the main results of Tucker's paper [21] with Ms. Kateřina Lokajová. The author is indebted to an anonymous referee for comments that helped him to improve the paper and for pointing out some references. Last but not least, the author is very grateful to Professor R. G. Bland who suggested simplifications of the proof of Lemma 2 and whose ideas are incorporated in the proof presented here.

REFERENCES

[1] E. J. Anderson and P. Nash, *Linear Programming in Infinite-Dimensional Spaces*, Wiley, Chichester, UK, 1987.
[2] D. Bartl, *Farkas' Lemma, other theorems of the alternative, and linear programming in infinite-dimensional spaces: A purely linear-algebraic approach*, Linear Multilinear Algebra, 55 (2007), pp. 327–353.
[3] C. G. Broyden, *A simple algebraic proof of Farkas's lemma and related theorems*, Optim. Methods Softw., 8 (1998), pp. 185–199.
[4] S. N. Chernikov, *Linear Inequalities*, Nauka, Moscow, 1968 (in Russian).
[5] A. Dax, *The relationship between theorems of the alternative, least norm problems, steepest descent directions, and degeneracy: A review*, Ann. Oper. Res, 46 (1993), pp. 11–60.
[6] A. Dax, *An elementary proof of Farkas' lemma*, SIAM Rev., 39 (1997), pp. 503–507.
[7] A. Dax, *The distance between two convex sets*, Linear Algebra Appl., 416 (2006), pp. 184–213.
[8] A. Dax and V. P. Sreedharan, *Theorems of the alternative and duality*, J. Optim. Theory Appl., 94 (1997), pp. 561–590.
[9] J. Farkas, *Theorie der einfachen Ungleichungen*, J. Reine Angew. Math., 124 (1902), pp. 1–27.
[10] K. Fan, *On systems of linear inequalities*, in Linear Inequalities and Related Systems, H. W. Kuhn and A. W. Tucker, eds., Annals of Mathematics Studies 38, Princeton University Press, Princeton, NJ, 1956, pp. 99–156.
[11] D. Gale, *The Theory of Linear Economic Models*, McGraw-Hill, New York, 1960.
[12] R. A. Good, *Systems of linear relations*, SIAM Rev., 1 (1959), pp. 1–31.
[13] A. Haar, *A lineáris egyenlőtlenségekről (On linear inequalities)*, Mathematikai és Természettudományi Értesitő, 36 (1918), pp. 279–296 (in Hungarian).
[14] A. Haar, *Über lineare Ungleichungen*, Acta Math. Szeged, 2 (1924), pp. 1–14.
[15] M. M. Marsh, *A note on generalized Farkas alternatives*, Topology Proc., 28 (2004), pp. 153–162.
[16] T. S. Motzkin, *Beiträge zur Theorie der linearen Ungleichungen*, Doctoral dissertation at the University of Basel, Basel, 1934; Azriel, Jerusalem, 1936.
[17] T. S. Motzkin, *Contributions to the theory of linear inequalities*, RAND Corporation Translation 22, RAND Corporation, Santa Monica, CA, 1952. Reprint in Theodore S. Motzkin: Selected Papers, D. Cantor, B. Gordon, and B. Rothschild, eds., Birkhäuser, Boston, 1983, pp. 1–80.
[18] M. Padberg, *Linear Optimization and Extensions*, 2nd ed., Springer-Verlag, Berlin, 1999.
[19] C. Roos and T. Terlaky, *Note on a paper of Broyden*, Oper. Res. Lett., 25 (1999), pp. 183–186.
[20] W. Schirotzek, *On Farkas type theorems*, Comment. Math. Univ. Carolin., 22 (1981), pp. 1–14.
[21] A. W. Tucker, *Dual systems of homogeneous linear relations*, in Linear Inequalities and Related Systems, H. W. Kuhn and A. W. Tucker, eds., Annals of Mathematics Studies 38, Princeton University Press, Princeton, NJ, 1956, pp. 3–18.
[22] C. Zălinescu, *A generalization of the Farkas lemma and applications to convex programming*, J. Math. Anal. Appl., 66 (1978), pp. 651–678.

SIAM J. OPTIM.
Vol. 19, No. 1, pp. 240–260

A CLASS OF INEXACT VARIABLE METRIC PROXIMAL POINT ALGORITHMS*

L. A. PARENTE†, P. A. LOTITO‡, AND M. V. SOLODOV§

Abstract. For the problem of solving maximal monotone inclusions, we present a rather general class of algorithms, which contains hybrid inexact proximal point methods as a special case and allows for the use of a variable metric in subproblems. The global convergence and local linear rate of convergence are established under standard assumptions. We demonstrate the advantage of variable metric implementation in the case of solving systems of smooth monotone equations by the proximal Newton method.

Key words. proximal point methods, variable metric, maximal monotone operators, approximation

AMS subject classifications. 90C30, 90C33

DOI. 10.1137/070688146

1. Introduction.

Given a maximal monotone operator $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, we consider the classical problem of finding a zero of T, i.e., $z \in \mathbb{R}^n$ such that

$$0 \in T(z). \tag{1.1}$$

As is well known, many important problems can be cast in this framework. Some examples are convex optimization, min-max problems, and monotone variational inequalities over convex sets; see, e.g., [23].

Given some $z^k \in \mathbb{R}^n$, the current approximation to a solution of (1.1), the proximal point iteration [19, 22] generates z^{k+1} as the solution of the regularized subproblem

$$0 \in c_k T(z) + z - z^k, \tag{1.2}$$

where $c_k > 0$ is the regularization parameter. As is well known, the proximal point method serves as a basis for developing and analyzing various useful computational techniques, such as splitting methods for variational problems (e.g., [18, 31, 13, 33, 34, 24]), the methods of multipliers (e.g., [21, 15]), and bundle methods for nonsmooth optimization (see, e.g., [16, 3]), to name a few. In computational context, it is important to handle approximate solutions of subproblems; this will be discussed a little further. Also, it is attractive to allow for the use of a variable metric (or preconditioning). Motivated by the latter issue, we shall consider the following *generalized* proximal subproblem:

$$0 \in c_k M_k T(z) + z - z^k, \tag{1.3}$$

*Received by the editors April 12, 2007; accepted for publication (in revised form) October 5, 2007; published electronically March 19, 2008.
http://www.siam.org/journals/siopt/19-1/68814.html
†CONICET, Universidad Nacional de Rosario, Argentina (lparente@fceia.unr.edu.ar).
‡CONICET, Universidad Nacional del Centro de la Provincia de Buenos Aires, Argentina (plotito@exa.unicen.edu.ar).
§IMPA – Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Jardim Botânico, Rio de Janeiro, RJ 22460-320, Brazil (solodov@impa.br). The research of this author is supported in part by CNPq grants 301508/2005-4, 490200/2005-2, and 550317/2005-8, by PRONEX–Optimization, and by FAPERJ.

where M_k is a symmetric positive definite matrix. The case of the classical (exact) iteration (1.2) corresponds to taking $M_k = I$ (the identity matrix) in (1.3). Given the presence of the matrix M_k, we could in principle dispense with the scalar parameter c_k in (1.3). We prefer, however, to keep it because this appears convenient in some parts of the convergence analysis and in our application to solving systems of monotone equations, discussed in section 5.

To handle approximate solutions, we shall use an extension to the variable metric setting of the rules proposed in [27, 26] and unified in [30]. In those algorithms, the *relative* error in the approximation needs only to be bounded (above, by one), which is a numerically sound requirement, and inexact values of the operator T are allowed, which is useful in various applications [29, 28, 24]. Specifically, to solve (1.3) approximately, the task would be to compute a triplet $(\hat{z}^k, \hat{v}^k, \varepsilon_k) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}_+$ such that

$$\begin{cases} \hat{v}^k \in T^{\varepsilon_k}(\hat{z}^k), \\ c_k M_k \hat{v}^k + \hat{z}^k - z^k = \delta^k, \end{cases}$$
$$\|\delta^k\|^2_{M_k^{-1}} + 2c_k\varepsilon_k \le \sigma_k^2 \left(\|c_k M_k \hat{v}^k\|^2_{M_k^{-1}} + \|\hat{z}^k - z^k\|^2_{M_k^{-1}} \right),$$

where $\sigma_k \in [0, 1)$ is the error tolerance (relaxation) parameter, by $\|\cdot\|_M$ we denote the norm induced by a symmetric positive definite matrix $M \in \mathcal{M}^n_{++}$, i.e.,

$$\|z\|_M = \sqrt{\langle z, Mz \rangle},$$

and $T^\varepsilon : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is the ε-enlargement of a maximal monotone operator T [5, 6], defined as

$$T^\varepsilon(z) := \{v \in \mathbb{R}^n \mid \langle w - v, y - z \rangle \ge -\varepsilon \ \forall y \in \mathbb{R}^n, \forall w \in T(y)\}, \ \varepsilon \ge 0.$$

We note that, to check the above criterion, one does not need to invert the matrix M_k, as will be explained in what follows. The presented approximation rule is constructive and has advantages in some situations, when compared to the original [22] (and its variations, e.g., [32, 11, 7]), where essentially one has $\varepsilon_k = 0$ and $\sum_{k=0}^\infty \|\delta^k\| < \infty$ (in the setting of $M_k = I$). We refer the reader to [26, 29, 28, 24] for some applications where the relative-error criterion appears useful. It will also play a central role in the method discussed in section 5.

Most proximal-related schemes in the literature that use variable metrics typically deal only with the special case of optimization, i.e., the case where the operator T is the subdifferential of a convex function [2, 20, 17, 10]. To our knowledge, the exception is [7] and some of the subsequent results [8, 9]. We note that our use of a variable metric is different from [7], where (exact) iteration is of the form

$$z^{k+1} = z^k + M_k((I + c_k T)^{-1} - I)z^k.$$

The exact iteration of solving (1.3) can be written as

$$z^{k+1} = (I + c_k M_k T)^{-1} z^k,$$

and the two are the same only when $M_k = I$. It should be noted, however, that [7] does not require M_k to be symmetric, and in this respect our development can be more restrictive for some applications. On the other hand, global convergence of the method of [7] requires a rather technical assumption about the matrices M_k.

Specifically, the assumption of [7, Hypothesis (H2)] is that there exists a nonempty bounded subset Ω of $T^{-1}(0)$ such that

$$\|(M_k - I)D_k(z^k)\| \leq \gamma_k \|D_k(z^k)\| \quad \text{for all } k,$$

where

$$D_k = (I + c_k T)^{-1} - I,$$

$$\gamma_k = \frac{\|D_k(z^k)\|}{2t_k + 3\|D_k(z^k)\|}, \quad \text{with } t_k = \sup_{z \in \Omega} \|z - z^k\|.$$

This assumption essentially means that matrices M_k should not deviate from the identity too much, in the given sense, and it is in general unverifiable (unless one takes $M_k = I$) and globally quite restrictive. The only assumption we make in our global convergence analysis is, by comparison, rather mild:

$$\frac{1}{1+\eta_k} M_k \preceq M_{k+1}, \ \eta_k > 0 \quad \text{for all } k, \quad \sum_{k=0}^{\infty} \eta_k < \infty, \tag{1.4}$$

where, for $A, B \in \mathcal{M}^n_{++}$, by $A \preceq B$ we mean that $B - A$ is a positive semidefinite matrix. This condition does not introduce any essential restrictions on the choice of the matrix M_{k+1} for a given k (for a particular k, the choice of η_k is rather flexible), and it is always satisfied if we take $M_k \preceq M_{k+1}$. Also, [7] does not allow approximations of the operator T and requires error terms to be summable, basically following [22]. In the aspect of inexact solution of subproblems, our conditions (already mentioned above) are more flexible and constructive.

A few more words about our notation are in order. By $\mathcal{M}^n_{++}$ we denote the space of symmetric positive definite matrices. For $M \in \mathcal{M}^n_{++}$, $\lambda_{min}(M)$ and $\lambda_{max}(M)$ stand for the minimal and the maximal eigenvalues of M, respectively. For any $A \preceq B$, it holds that $\|z\|_A \leq \|z\|_B$. In particular, if

$$0 < \lambda_l \leq \lambda_{min}(M) \leq \lambda_{max}(M) \leq \lambda_u,$$

then for any $x \in \mathbb{R}^n$ it holds that

$$\lambda_l \|x\|^2 \leq \|x\|^2_M \leq \lambda_u \|x\|^2, \quad \frac{1}{\lambda_u}\|x\|^2 \leq \|x\|^2_{M^{-1}} \leq \frac{1}{\lambda_l}\|x\|^2. \tag{1.5}$$

By $\langle x, y \rangle$ we denote the usual inner product between $x, y \in \mathbb{R}^n$. For a matrix $M \in \mathcal{M}^n_{++}$, we denote $\langle x, y \rangle_M = \langle Mx, y \rangle$. For a closed convex set $\Omega \subseteq \mathbb{R}^n$ and a matrix $M \in \mathcal{M}^n_{++}$, the "skewed" projection operator onto Ω under the matrix M is given by

$$P_{\Omega,M}(z) = \arg\min_{x \in \Omega} \frac{1}{2}\langle x - z, M(x - z)\rangle = \arg\min_{x \in \Omega} \frac{1}{2}\|x - z\|^2_M;$$

i.e., it is the projection operator with respect to the norm $\|\cdot\|_M$. Then the associated distance from $z \in \mathbb{R}^n$ to Ω is defined as $\operatorname{dist}(z, \Omega)_M = \|z - P_{\Omega,M}(z)\|$.

2. Approximate solutions of the generalized proximal subproblem. Given a maximal monotone operator $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, $z \in \mathbb{R}^n$, $c > 0$, and $M \in \mathcal{M}^n_{++}$, consider the generalized proximal point subproblem

$$0 \in cMT(y) + y - z, \tag{2.1}$$

with respect to $y \in \mathbb{R}^n$. This is clearly equivalent to

$$0 \in cT(y) + M^{-1}(y - z),$$

and the fact that the inclusion above has a solution follows, e.g., from [4, Proposition 3].

We next define the notion of approximate solutions of generalized proximal subproblems. Consider the system

$$\begin{cases} v \in T(y), \\ 0 = cMv + y - z, \end{cases} \tag{2.2}$$

which is equivalent to (2.1).

DEFINITION 2.1. *We say that a triplet $(y, v, \varepsilon) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}_+$ is an approximate solution of the proximal system* (2.2) *with error tolerance $\sigma \in [0, 1)$ if*

$$v \in T^{\varepsilon}(y)$$

and

$$\|cMv + y - z\|^2_{M^{-1}} + 2c\varepsilon \leq \sigma^2(\|cMv\|^2_{M^{-1}} + \|y - z\|^2_{M^{-1}}). \tag{2.3}$$

Note that the exact solution of (2.2) corresponds to taking $\varepsilon = 0 = \sigma$ in the definition above. We next establish some properties of approximate solutions of generalized proximal systems.

LEMMA 2.2. *Let $z \in \mathbb{R}^n, c > 0$, and $M \in \mathcal{M}^n_{++}$. A triplet $(y, v, \varepsilon) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}_+$ being an approximate solution of the proximal system* (2.2) *with error tolerance $\sigma \in [0, 1)$ is equivalent to the conditions*

$$v \in T^{\varepsilon}(y), \quad \langle v, z - y\rangle - \varepsilon \geq \frac{1 - \sigma^2}{2c} \left(\|cMv\|^2_{M^{-1}} + \|y - z\|^2_{M^{-1}}\right). \tag{2.4}$$

In addition, it holds that

$$\frac{c(1 - \rho)}{1 - \sigma^2}\|Mv\|_{M^{-1}} \leq \|y - z\|_{M^{-1}} \leq \frac{c(1 + \rho)}{1 - \sigma^2}\|Mv\|_{M^{-1}}, \tag{2.5}$$

where $\rho = \sqrt{1 - (1 - \sigma^2)^2}$.

Furthermore, the three conditions

1. $0 \in T(z)$,
2. $v = 0$,
3. $y = z$

are equivalent and imply that $\varepsilon = 0$.

Proof. Rearranging terms in (2.3), we have

$$\begin{aligned} \sigma^2(\|cMv\|^2_{M^{-1}} + \|y - z\|^2_{M^{-1}}) &\geq 2c\varepsilon + \|cMv\|^2_{M^{-1}} + \|y - z\|^2_{M^{-1}} + 2\langle cMv, y - z\rangle_{M^{-1}} \\ &= 2c\varepsilon + \|cMv\|^2_{M^{-1}} + \|y - z\|^2_{M^{-1}} - 2c\langle v, z - y\rangle, \end{aligned}$$

which gives the inequality in (2.4).

By using $\varepsilon \geq 0$ and the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned} \frac{1 - \sigma^2}{2c}(\|cMv\|^2_{M^{-1}} + \|y - z\|^2_{M^{-1}}) &\leq \langle v, z - y\rangle - \varepsilon \\ &\leq \langle Mv, z - y\rangle_{M^{-1}} \leq \|Mv\|_{M^{-1}}\|y - z\|_{M^{-1}}. \end{aligned}$$

By denoting $t = \|y - z\|_{M^{-1}}$ and resolving the quadratic inequality

$$t^2 - \frac{2\|cMv\|_{M^{-1}}}{1-\sigma^2}t + \|cMv\|^2_{M^{-1}} \leq 0$$

with respect to t, we obtain (2.5).

Finally, suppose that $0 \in T(z)$. Since $v \in T^\varepsilon(y)$, we have

$$\langle v - 0, y - z\rangle \geq -\varepsilon \quad \Rightarrow \quad \langle v, z - y\rangle - \varepsilon \leq 0.$$

By using now (2.4), we have $cMv = 0$ (so that $v = 0$) and $y - z = 0$.

If we assume that $v = 0$, then (2.4) implies that $y = z$ and vice versa. In either case, $0 \in T(z)$. From (2.4) it is also clear that all of these conditions imply that $\varepsilon = 0$. □

The next result shows how to make progress towards a solution of the original problem (1.1), by using the obtained approximate solution of the generalized proximal subproblem.

LEMMA 2.3. *Let $z \in \mathbb{R}^n, y \in \mathbb{R}^n, \varepsilon \geq 0$, and $v \in T^\varepsilon(y)$. Suppose that*

$$\langle v, z - y\rangle - \varepsilon > 0.$$

Then, for any $z^* \in T^{-1}(0)$, any $M \in \mathcal{M}^n_{++}$, and any $\tau \geq 0$, it holds that

$$\|z^* - z^+\|^2_{M^{-1}} \leq \|z^* - z\|^2_{M^{-1}} - (1 - (1-\tau)^2)a^2\|Mv\|^2_{M^{-1}},$$

where

$$z^+ := z - \tau a M v$$

and

$$a := \frac{\langle v, z - y\rangle - \varepsilon}{\|Mv\|^2_{M^{-1}}}.$$

Proof. Define the closed half-space

$$H = \{w \in \mathbb{R}^n \mid \langle v, w - y\rangle - \varepsilon \leq 0\}.$$

By the assumption, $z \notin H$. Let $\bar{z}$ be the skewed projection of z onto H, under the matrix M^{-1}. As is easily seen,

$$\bar{z} = P_{H,M^{-1}}(z) = z - aMv.$$

For any $x \in H$, it holds that

$$\langle x - \bar{z}, v\rangle = \langle x - z + aMv, v\rangle = \langle x - z, v\rangle + \frac{\langle v, z - y\rangle - \varepsilon}{\|Mv\|^2_{M^{-1}}}\langle Mv, v\rangle = \langle x - y, v\rangle - \varepsilon \leq 0.$$

Hence,

$$\langle x - \bar{z}, z^+ - z\rangle_{M^{-1}} = \langle x - \bar{z}, M^{-1}(-\tau a M v)\rangle = -\tau a\langle x - \bar{z}, v\rangle \geq 0.$$

Observe that $\bar{z} - z^+ = (\tau - 1)aMv$. We then obtain

$$\begin{aligned}
\|x - z\|^2_{M^{-1}} &= \|x - z^+\|^2_{M^{-1}} + \|z^+ - z\|^2_{M^{-1}} + 2\langle x - z^+, z^+ - z\rangle_{M^{-1}} \\
&= \|x - z^+\|^2_{M^{-1}} + \|z^+ - z\|^2_{M^{-1}} + 2\langle \bar{z} - z^+, z^+ - z\rangle_{M^{-1}} \\
&\quad +2\langle x - \bar{z}, z^+ - z\rangle_{M^{-1}} \\
&\geq \|x - z^+\|^2_{M^{-1}} + \|z^+ - z\|^2_{M^{-1}} + 2\langle \bar{z} - z^+, z^+ - z\rangle_{M^{-1}} \\
&= \|x - z^+\|^2_{M^{-1}} + (\tau a)^2\|Mv\|^2_{M^{-1}} + 2(\tau - 1)a(-\tau a)\|Mv\|^2_{M^{-1}} \\
&= \|x - z^+\|^2_{M^{-1}} + (1 - (1 - \tau)^2)a^2\|Mv\|^2_{M^{-1}}.
\end{aligned}$$

Suppose that $z^* \in T^{-1}(0)$. Since $v \in T^{\varepsilon}(y)$, we have

$$\langle v - 0, y - z^*\rangle \geq -\varepsilon.$$

This shows that $z^* \in H$. We can then set $x = z^*$ in the chain of inequalities above to complete the proof. □

3. The algorithm. Lemma 2.3 shows that, with a proper choice of parameters, a step in the direction obtained from an approximate solution of the generalized proximal system, scaled by the chosen metric, brings us closer to the solution set of the original problem. This suggests the following scheme, which we shall call the variable metric hybrid inexact proximal point method.

ALGORITHM 3.1. **Initialization:** *Choose* $z^0 \in \mathbb{R}^n$, $c > 0$, $\bar{\sigma} \in (0, 1)$, $\theta \in (0, 1)$, *and* $0 < \lambda_l < \lambda_u$. *Set* $k := 0$.

Inexact proximal step: *Choose* $c_k \geq c$, *a symmetric positive definite matrix* M_k *satisfying* $\lambda_l \leq \lambda_{min}(M_k) \leq \lambda_{max}(M_k) \leq \lambda_u$, *and the error tolerance parameter* $\sigma_k \in [0, \bar{\sigma})$. *Find* $\hat{z}^k \in \mathbb{R}^n, \hat{v}^k \in \mathbb{R}^n$, *and* $\varepsilon_k \geq 0$ *such that*

$$\begin{cases} \hat{v}^k \in T^{\varepsilon_k}(\hat{z}^k), \\ \delta^k = c_k M_k \hat{v}^k + \hat{z}^k - z^k \end{cases} \tag{3.1}$$

and

$$\|\delta^k\|^2_{M_k^{-1}} + 2c_k\varepsilon_k \leq \sigma_k^2 \left(\|c_k M_k \hat{v}^k\|^2_{M_k^{-1}} + \|\hat{z}^k - z^k\|^2_{M_k^{-1}}\right). \tag{3.2}$$

Iterates update: *If* $\hat{z}^k = z^k$, *stop. Otherwise, choose* $\tau_k \in [1 - \theta, 1 + \theta]$, *and set*

$$z^{k+1} = z^k - \tau_k a_k M_k \hat{v}^k, \quad a_k = \frac{\langle \hat{v}^k, z^k - \hat{z}^k\rangle - \varepsilon_k}{\|M_k \hat{v}^k\|^2_{M_k^{-1}}}.$$

Set $k := k + 1$, *and go to the inexact proximal step.*

We note that it is not necessary to calculate the inverse of M_k in order to implement Algorithm 3.1 (in particular, for checking the condition (3.2) and computing a_k). Indeed, by (1.5), the condition (3.2) is satisfied if

$$\|\delta^k\|^2 + 2\lambda_u c_k \varepsilon_k \leq \frac{\lambda_u \sigma_k^2}{\lambda_l}\left(\|c_k M_k \hat{v}^k\|^2 + \|\hat{z}^k - z^k\|^2\right).$$

Alternatively, in the latter relation, instead of λ_l and λ_u one can use any other (in particular, tighter) lower and upper bounds for the eigenvalues of M_k. Also, the scalar a_k can be calculated as

$$a_k = \frac{\langle \hat{v}^k, z^k - \hat{z}^k\rangle - \varepsilon_k}{\langle M_k \hat{v}^k, \hat{v}^k\rangle}.$$

The next result shows that some specific realizations of Algorithm 3.1 allow for the simple update

$$z^{k+1} = z^k - c_k M_k \hat{v}^k.$$

This is the update that we shall use for our application in section 5. Specifically, we have the following.

PROPOSITION 3.1. *If the inequality in* (3.2) *is replaced by the stronger condition* $\|\delta^k\|^2_{M_k^{-1}} + 2c_k\varepsilon_k \le \sigma_k^2\|\hat{z}^k - z^k\|^2_{M_k^{-1}}$, *and we choose* $\sigma_k \le \theta$, *then there exists* $\tau_k \in (1-\sigma_k, 1+\sigma_k) \subset (0,2)$ *such that* $\tau_k a_k = c_k$.

Proof. In the case of interest, $\hat{v}^k \neq 0$ and $\hat{z}^k \neq z^k$. By using the triangle inequality, from $\|\delta^k\|_{M_k^{-1}} \le \sigma_k\|\hat{z}^k - z^k\|_{M_k^{-1}}$ we obtain

$$\|\hat{z}^k - z^k\|_{M_k^{-1}} - c_k\|M_k\hat{v}^k\|_{M_k^{-1}} \le \sigma_k\|\hat{z}^k - z^k\|_{M_k^{-1}}$$

and

$$c_k\|M_k\hat{v}^k\|_{M_k^{-1}} - \|\hat{z}^k - z^k\|_{M_k^{-1}} \le \sigma_k\|\hat{z}^k - z^k\|_{M_k^{-1}},$$

implying that

$$(1-\sigma_k)\frac{\|\hat{z}^k - z^k\|_{M_k^{-1}}}{\|M_k\hat{v}^k\|_{M_k^{-1}}} \le c_k \le (1+\sigma_k)\frac{\|\hat{z}^k - z^k\|_{M_k^{-1}}}{\|M_k\hat{v}^k\|_{M_k^{-1}}}. \tag{3.3}$$

Furthermore, by the Cauchy–Schwarz inequality, since $\varepsilon_k \ge 0$ we have

$$a_k = \frac{\langle \hat{v}^k, z^k - \hat{z}^k\rangle - \varepsilon_k}{\|M_k\hat{v}^k\|^2_{M_k^{-1}}} \le \frac{\langle M_k\hat{v}^k, z^k - \hat{z}^k\rangle_{M_k^{-1}}}{\|M_k\hat{v}^k\|^2_{M_k^{-1}}} \le \frac{\|\hat{z}^k - z^k\|_{M_k^{-1}}}{\|M_k\hat{v}^k\|_{M_k^{-1}}}.$$

Finally, since

$$\begin{aligned}\langle \hat{v}^k, \hat{z}^k - z^k\rangle &= \langle M_k\hat{v}^k, \hat{z}^k - z^k\rangle_{M_k^{-1}} \\ &= \frac{\|c_kM_k\hat{v}^k + \hat{z}^k - z^k\|^2_{M_k^{-1}} - \|\hat{z}^k - z^k\|^2_{M_k^{-1}} - \|c_kM_k\hat{v}^k\|^2_{M_k^{-1}}}{2c_k},\end{aligned}$$

by using (2.5) and (3.3), we obtain

$$\begin{aligned}a_k &= \frac{\|\hat{z}^k - z^k\|^2_{M_k^{-1}} + \|c_kM_k\hat{v}^k\|^2_{M_k^{-1}} - \left(\|c_kM_k\hat{v}^k + \hat{z}^k - z^k\|^2_{M_k^{-1}} + 2c_k\varepsilon_k\right)}{2c_k\|M_k\hat{v}^k\|^2_{M_k^{-1}}} \\ &\ge \frac{c_k}{2} + (1-\sigma_k^2)\frac{\|\hat{z}^k - z^k\|^2_{M_k^{-1}}}{\|c_kM_k\hat{v}^k\|^2_{M_k^{-1}}} \ge \frac{c_k}{2}\left(1 + \frac{1-\sigma_k^2}{(1+\sigma_k)^2}\right) = \frac{c_k}{1+\sigma_k}.\end{aligned}$$

Hence,

$$(1-\sigma_k)a_k \le c_k \le (1+\sigma_k)a_k,$$

which establishes the claim. ∎

4. Convergence analysis. If Algorithm 3.1 terminates at some iteration k, then $z^k = \hat{z}^k$, and, by Lemma 2.2, z^k is a solution. We next consider the case when infinite sequences $\{z^k\}, \{\hat{z}^k\}, \{\hat{v}^k\}$, and $\{\varepsilon_k\}$ are generated. For any k, we have $\hat{v}^k \neq 0$, $\hat{z}^k \neq z^k$, and by Lemma 2.2,

$$\langle \hat{v}^k, z^k - \hat{z}^k \rangle - \varepsilon_k \geq \frac{1-\sigma_k^2}{2c_k} \left(\|c_k M_k \hat{v}^k\|^2_{M_k^{-1}} + \|\hat{z}^k - z^k\|^2_{M_k^{-1}} \right) > 0.$$

By the definition of a_k, we then conclude that

$$a_k \geq \frac{1-\sigma_k^2}{2c_k} \left(\frac{\|c_k M_k \hat{v}^k\|^2_{M_k^{-1}} + \|\hat{z}^k - z^k\|^2_{M_k^{-1}}}{\|M_k \hat{v}^k\|^2_{M_k^{-1}}} \right). \tag{4.1}$$

By the Cauchy–Schwarz inequality,

$$\|c_k M_k \hat{v}^k\|^2_{M_k^{-1}} + \|\hat{z}^k - z^k\|^2_{M_k^{-1}} \geq 2c_k \|M_k \hat{v}^k\|_{M_k^{-1}} \|\hat{z}^k - z^k\|_{M_k^{-1}}.$$

By combining this relation with (4.1), we obtain

$$a_k \|M_k \hat{v}^k\|_{M_k^{-1}} \geq (1-\sigma_k^2) \|\hat{z}^k - z^k\|_{M_k^{-1}}. \tag{4.2}$$

Combining (4.1) and (2.5), and using the definition of ρ_k, gives the following lower bound for a_k:

$$\begin{aligned} a_k &\geq \frac{(1-\sigma_k^2)c_k}{2} \left(1 + \frac{\|\hat{z}^k - z^k\|^2_{M_k^{-1}}}{\|c_k M_k \hat{v}^k\|^2_{M_k^{-1}}} \right) \\ &\geq \frac{(1-\sigma_k^2)c_k}{2} \left(1 + \left(\frac{1-\rho_k}{1-\sigma_k^2} \right)^2 \right) \\ &= \frac{c_k \left(\left(1-\sigma_k^2\right)^2 + \left(1 - \sqrt{1-(1-\sigma_k^2)^2} \right)^2 \right)}{2\left(1-\sigma_k^2\right)} \\ &= \frac{c_k \left(1 - \sqrt{1-(1-\sigma_k^2)^2} \right)}{1-\sigma_k^2} \\ &= \frac{(1-\sigma_k^2)c_k}{1+\sqrt{1-(1-\sigma_k^2)^2}}. \end{aligned} \tag{4.3}$$

Hence, the parameter a_k is bounded away from zero:

$$a_k \geq \frac{(1-\bar{\sigma}^2)c}{1+\sqrt{1-(1-\bar{\sigma}^2)^2}} > 0. \tag{4.4}$$

We proceed to establish the global convergence of Algorithm 3.1.

PROPOSITION 4.1. *Suppose that $T^{-1}(0) \neq \emptyset$ and condition (1.4) holds. Then any sequences generated by Algorithm 3.1 have the following properties:*

1. *$\{z^k\}$ is bounded.*
2. *$\sum_{k=0}^{\infty} \|a_k M_k \hat{v}^k\|^2 < \infty$.*
3. *$\lim_{k\to\infty} \|\hat{z}^k - z^k\| = \lim_{k\to\infty} \|\hat{v}^k\| = \lim_{k\to\infty} \|\varepsilon_k\| = 0$.*

Proof. By condition (1.4), it holds that

$$\prod_{k=0}^{\infty}(1+\eta_k) = p < \infty,$$

and, for all k,

$$M_{k+1}^{-1} \preceq (1+\eta_k)M_k^{-1}.$$

By (1.5), for all k we have

$$\lambda_u^{-1}\|z\|^2 \leq \lambda_{min}(M_k^{-1})\|z\|^2 \leq \|z\|^2_{M_k^{-1}} \leq \lambda_{max}(M_k^{-1})\|z\|^2 \leq \lambda_l^{-1}\|z\|^2 \quad \forall\, z \in \mathbb{R}^n.$$

By using (4.1) and Lemma 2.3, we have that for any $z^* \in T^{-1}(0)$ it holds that

$$\begin{aligned}\|z^* - z^{k+1}\|^2_{M_k^{-1}} &\leq \|z^* - z^k\|^2_{M_k^{-1}} - (1-(1-\tau_k)^2)a_k^2\|M_k\hat{v}^k\|^2_{M_k^{-1}}\\ &\leq \|z^* - z^k\|^2_{M_k^{-1}} - (1-\theta^2)\|a_k M_k\hat{v}^k\|^2_{M_k^{-1}}.\end{aligned}$$

Hence,

$$\begin{aligned}\lambda_u^{-1}\|z^* - z^{k+1}\|^2 &\leq \|z^* - z^{k+1}\|^2_{M_{k+1}^{-1}}\\ &\leq (1+\eta_k)\|z^* - z^{k+1}\|^2_{M_k^{-1}}\\ &\leq (1+\eta_k)\left(\|z^* - z^k\|^2_{M_k^{-1}} - (1-\theta^2)\|a_k M_k\hat{v}^k\|^2_{M_k^{-1}}\right)\\ &\leq (1+\eta_k)\|z^* - z^k\|^2_{M_k^{-1}} - (1-\theta^2)\|a_k M_k\hat{v}^k\|^2_{M_k^{-1}}.\end{aligned}$$

By applying this inequality consecutively, we obtain

$$\lambda_u^{-1}\|z^* - z^{k+1}\|^2 \leq \prod_{i=0}^{k}(1+\eta_i)\|z^* - z^0\|^2_{M_0^{-1}} - (1-\theta^2)\sum_{i=0}^{k}\|a_i M_i\hat{v}^i\|^2_{M_i^{-1}}. \tag{4.5}$$

We therefore have, for any k,

$$\|z^* - z^k\|^2 \leq \lambda_u \prod_{i=0}^{k-1}(1+\eta_i)\|z^* - z^0\|^2_{M_0^{-1}} \leq \frac{p\lambda_u}{\lambda_l}\|z^* - z^0\|^2, \tag{4.6}$$

which shows that the sequence $\{z^k\}$ is bounded. From (4.5), we also have

$$(1-\theta^2)\sum_{i=0}^{k}\|a_i M_i\hat{v}^i\|^2_{M_i^{-1}} \leq \prod_{i=0}^{k}(1+\eta_i)\|z^* - z^0\|^2_{M_0^{-1}}.$$

By passing onto the limit when $k \to \infty$ in this relation, we obtain

$$\sum_{k=0}^{\infty}\|a_k M_k\hat{v}^k\|^2 \leq \lambda_u \sum_{k=0}^{\infty}\|a_k M_k\hat{v}^k\|^2_{M_k^{-1}} \leq \frac{p\lambda_u}{1-\theta^2}\|z^* - z^0\|^2_{M_0^{-1}} < \infty.$$

This proves the second item in the assertion and, as a consequence, that

$$\lim_{k\to\infty}\|a_k M_k\hat{v}^k\| = 0.$$

From (4.2) and Lemma 2.2, we then conclude that

$$\lim_{k\to\infty} \|M_k \hat{v}^k\| = 0 \quad \text{and} \quad \lim_{k\to\infty} \|\hat{z}^k - z^k\| = 0.$$

Since the matrices M_k are uniformly positive definite, we also have $\lim_{k\to\infty} \hat{v}^k = 0$. Also, since $\varepsilon_k \le \langle \hat{v}^k, z^k - \hat{z}^k \rangle$, it follows that $\lim_{k\to\infty} \varepsilon_k = 0$. □

We are now in a position to complete the proof of global convergence of Algorithm 3.1. Given the properties established in Proposition 4.1, the argument is close to standard; we include it mainly for completeness.

THEOREM 4.2. *Suppose that* $T^{-1}(0) \neq \emptyset$ *and condition* (1.4) *holds. Then any sequence* $\{z^k\}$ *generated by Algorithm* 3.1 *converges to an element of* $T^{-1}(0)$.

Proof. By Proposition 4.1, the sequence $\{z^k\}$ is bounded. Therefore, it has some accumulation point, say, $\bar{z} \in \mathbb{R}^n$. Let $\{z^{k_j}\}$ be any subsequence converging to this accumulation point: $\lim_{j\to\infty} z^{k_j} = \bar{z}$. Since $\lim_{k\to\infty} \|\hat{z}^k - z^k\| = 0$, we have $\hat{z}^{k_j} \to \bar{z}$. For any $z \in \mathbb{R}^n$ and any $u \in T(z)$, $\langle u - v^{k_j}, z - \hat{z}^{k_j} \rangle \ge -\varepsilon_{k_j}$. Hence,

$$\langle u - 0, z - \hat{z}^{k_j} \rangle \ge \langle v^{k_j}, z - \hat{z}^{k_j} \rangle - \varepsilon_{k_j}.$$

Since $v^{k_j} \to 0$, $\varepsilon_{k_j} \to 0$, and $\hat{z}^{k_j} \to \bar{z}$, by passing onto the limit when $j \to \infty$ we obtain

$$\langle u - 0, z - \bar{z} \rangle \ge 0.$$

As $z \in \mathbb{R}^n$ and $u \in T(z)$ were arbitrarily chosen, and T is maximal monotone, the above relation shows that $0 \in T(\bar{z})$; i.e., $\bar{z}$ is a solution.

Suppose that there exists another subsequence $\{z^{t_i}\}$ converging to $\tilde{z} \neq \bar{z}$. Fix some $d \in (0, \|\tilde{z} - \bar{z}\|)$. Since $\tilde{z}$ and $\bar{z}$ are limits of corresponding subsequences, there exists an index i_0 such that for all $i \ge i_0$

$$\|z^{t_i} - \tilde{z}\| < \frac{d}{2}\sqrt{\frac{\lambda_l}{p\lambda_u}},$$

where $p = \prod_{k=0}^{\infty}(1 + \eta_k)$, and there exists an index j_0 such that for all $j \ge j_0$

$$k_j > t_{i_0} \text{ and } \|z^{k_j} - \bar{z}\| < \frac{d}{2}.$$

Therefore,

$$\|z^{k_j} - \tilde{z}\| > \frac{d}{2} \quad \forall j \ge j_0.$$

Since, as already established above, $\tilde{z} \in T^{-1}(0)$, the same reasoning used to obtain (4.6) gives, for any $j \ge j_0$,

$$\frac{d}{2} < \|z^{k_j} - \tilde{z}\| \le \sqrt{\frac{p\lambda_u}{\lambda_l}}\|z^{t_{i_0}} - \tilde{z}\| < \frac{d}{2},$$

which is a contradiction.

Hence, $\{z^k\}$ has the unique accumulation point, which is a solution. □

We proceed with a convergence rate analysis of Algorithm 3.1. To this end, we first establish an *error bound* for the exact solution of the generalized proximal system

$$\begin{cases} v \in T(y), \\ 0 = cMv + y - z. \end{cases} \tag{4.7}$$

We note that the obtained bound is for the distance both in terms of y and in terms of v, and it does not involve any unknown constants. Specifically, we have the following.

LEMMA 4.3. *Let y^*, v^* be the exact solution of the proximal system* (4.7)*, with some $c > 0$, $z \in \mathbb{R}^n$, and $M \in \mathcal{M}^n_{++}$. Then for any $y \in \mathbb{R}^n$ and any $v \in T^\varepsilon(y)$, it holds that*

$$\|y - y^*\|^2_{M^{-1}} + c^2\|Mv - Mv^*\|^2_{M^{-1}} \le \|cMv + y - z\|^2_{M^{-1}} + 2c\varepsilon.$$

Proof. By using $cMv^* + y^* - z = 0$, we obtain

$$\begin{aligned}\|cMv + y - z\|^2_{M^{-1}} &= \|cMv + y - z - (cMv^* + y^* - z)\|^2_{M^{-1}}\\ &= \|cMv - cMv^* + y - y^*\|^2_{M^{-1}}\\ &= c^2\|Mv - Mv^*\|^2_{M^{-1}} + \|y - y^*\|^2_{M^{-1}} + 2c\langle v - v^*, y - y^*\rangle\\ &\ge c^2\|Mv - Mv^*\|^2_{M^{-1}} + \|y - y^*\|^2_{M^{-1}} - 2c\varepsilon. \qquad \square\end{aligned}$$

We shall show linear convergence of Algorithm 3.1 under the assumption that T^{-1} has the following Lipschitzian property at zero: There exist some $L_1 > 0$ and $L_2 > 0$ such that

$$T^{-1}(v) \subset T^{-1}(0) + L_1\|v\|\mathcal{B} \quad \forall v \in L_2\mathcal{B},$$

where $\mathcal{B} = \{x \in \mathbb{R}^n \mid \|x\| \le 1\}$. Note that this condition does not imply that the solution set $T^{-1}(0)$ is a singleton. The equivalent form of this local Lipschitzian property, used below, is

$$\operatorname{dist}(z, T^{-1}(0)) \le L_1 \min_{v \in T(z)} \|v\| \quad \forall\, z \in \{z' \in \operatorname{dom} T \mid \min_{v \in T(z')} \|v\| \le L_2\}. \tag{4.8}$$

We shall prove the linear convergence rate under one of the following two alternative assumptions on algorithm parameters. One is that $\bar\sigma$ is sufficiently small, while c is sufficiently large (note that those are user-chosen parameters). The other is that

$$\frac{1}{1+\eta_k} M_k \preceq M_{k+1} \preceq (1+\eta_k)M_k,\ \eta_k > 0\ \forall\, k, \quad \sum_{k=0}^{\infty} \eta_k < \infty, \tag{4.9}$$

which is a strengthening of the condition (1.4) used for global convergence. Asymptotically, (4.9) means that the matrices should not differ too much on subsequent iterations (a natural requirement in a neighborhood of a solution).

THEOREM 4.4. *In addition to the assumptions of Theorem* 4.2*, suppose that condition* (4.8) *is satisfied.*

Then, for sufficiently small choices of σ_k and sufficiently large choices of c_k, the sequence $\{z^k\}$ generated by Algorithm 3.1 *converges to an element of $T^{-1}(0)$ at a linear rate. If $c_k \to \infty$ and $\sigma_k \to 0$, the rate of convergence is superlinear.*

If condition (4.9) *holds, then for any choice of parameters $\bar\sigma$ and c, there exists $k_0 \in \mathbb{N}$ such that the sequence $\{z^k\}$ converges at the linear rate in the norm induced by $M^{-1}_{k_0}$.*

Proof. For each k, let τ_k, a_k, z^k be as defined in Algorithm 3.1, and let $x^k, w^k \in T(x^k)$ be the exact solution of the proximal system

$$\begin{cases} w \in T(x), \\ 0 = b_k M_k w + x - z^k, \end{cases}$$

where $b_k = \tau_k a_k$. Since $\hat{v}^k \in T^{\varepsilon_k}(\hat{z}^k)$, by Lemma 4.3 and the definition of a_k, it follows that

$$\begin{aligned}
&\|x^k - \hat{z}^k\|^2_{M_k^{-1}} + b_k^2 \|M_k \hat{v}^k - M_k w^k\|^2_{M_k^{-1}} \\
&\leq \|b_k M_k \hat{v}^k + \hat{z}^k - z^k\|^2_{M_k^{-1}} + 2b_k \varepsilon_k \\
&= \|b_k M_k \hat{v}^k + \hat{z}^k - z^k\|^2_{M_k^{-1}} \\
&\quad -2b_k \left(a_k \|M_k \hat{v}^k\|^2_{M_k^{-1}} + \langle M_k \hat{v}^k, z^k - \hat{z}^k \rangle_{M_k^{-1}} \right) \\
&= \|\hat{z}^k - z^k\|^2_{M_k^{-1}} + (\tau_k^2 - 2\tau_k) \|a_k M_k \hat{v}^k\|^2_{M_k^{-1}}.
\end{aligned}$$

By using (4.2), we then obtain

(4.10)
$$\|x^k - \hat{z}^k\|^2_{M_k^{-1}} + b_k^2 \|M_k \hat{v}^k - M_k w^k\|^2_{M_k^{-1}} \leq \left(\tau_k^2 - 2\tau_k + \frac{1}{(1-\sigma_k^2)^2} \right) \|a_k M_k \hat{v}^k\|^2_{M_k^{-1}}.$$

By using further the definitions of w^k and $\hat{v}^k$, this gives

$$\|x^k - \hat{z}^k\|^2_{M_k^{-1}} + \|x^k - z^{k+1}\|^2_{M_k^{-1}} \leq \left(\tau_k^2 - 2\tau_k + \frac{1}{(1-\sigma_k^2)^2} \right) \|a_k M_k \hat{v}^k\|^2_{M_k^{-1}}. \tag{4.11}$$

From (4.10), we also have

$$\|M_k \hat{v}^k - M_k w^k\|^2_{M_k^{-1}} \leq \left(1 - \frac{2}{\tau_k} + \frac{1}{\tau_k^2 (1-\sigma_k^2)^2} \right) \|M_k \hat{v}^k\|^2_{M_k^{-1}}.$$

Since $\hat{v}^k \to 0$ (see Proposition 4.1), the last relation implies that $w^k \to 0$. Hence, there exists $k_1 \in \mathbb{N}$ such that $\|w^k\| < L_2$ for all $k > k_1$. By (4.8), we then have

$$\operatorname{dist}(x^k, T^{-1}(0)) \leq L_1 \|w^k\| \quad \forall k > k_1.$$

Therefore, for $k > k_1$,

$$\begin{aligned}
\operatorname{dist}(x^k, T^{-1}(0))^2_{M_k^{-1}} &\leq \frac{1}{\lambda_l} \operatorname{dist}(x^k, T^{-1}(0))^2 \leq \frac{L_1^2}{\lambda_l} \|w^k\|^2 \\
&\leq \frac{L_1^2}{\lambda_l^2} \|w^k\|^2_{M_k} = \frac{L_1^2}{\lambda_l^2} \|M_k w^k\|^2_{M_k^{-1}} \\
&= \frac{L_1^2}{\lambda_l^2 b_k^2} \|z^k - x^k\|^2_{M_k^{-1}}. \qquad (4.12)
\end{aligned}$$

Let $\bar{x}^k$ be the skewed projection of x^k onto $T^{-1}(0)$ under the norm induced by M_k^{-1}, i.e.,

$$\bar{x}^k := P_{T^{-1}(0), M_k^{-1}}(x^k).$$

Then, for $k > k_1$, we have

$$\begin{aligned}
\operatorname{dist}(z^{k+1}, T^{-1}(0))_{M_k^{-1}} &\leq \|z^{k+1} - \bar{x}^k\|_{M_k^{-1}} \\
&\leq \|z^{k+1} - x^k\|_{M_k^{-1}} + \operatorname{dist}(x^k, T^{-1}(0))_{M_k^{-1}} \\
&\leq \|z^{k+1} - x^k\|_{M_k^{-1}} + \frac{L_1}{\lambda_l b_k} \|z^k - x^k\|_{M_k^{-1}} \\
&\leq \|z^{k+1} - x^k\|_{M_k^{-1}} + \frac{L_1}{\lambda_l b_k} \|x^k - \hat{z}^k\|_{M_k^{-1}} + \frac{L_1}{\lambda_l b_k} \|\hat{z}^k - z^k\|_{M_k^{-1}},
\end{aligned}$$

where the third inequality is by (4.12). By the Cauchy–Schwarz inequality, it holds that

$$\begin{aligned}
&\frac{L_1}{\lambda_l b_k}\|x^k-\hat{z}^k\|_{M_k^{-1}}+\|x^k-z^{k+1}\|_{M_k^{-1}}\\
&\le\sqrt{1+\frac{L_1^2}{\lambda_l^2 b_k^2}}\,\sqrt{\|x^k-\hat{z}^k\|^2_{M_k^{-1}}+\|x^k-z^{k+1}\|^2_{M_k^{-1}}}\\
&\le\sqrt{\left(1+\frac{L_1^2}{\lambda_l^2 b_k^2}\right)\left(\tau_k^2-2\tau_k+\frac{1}{(1-\sigma_k^2)^2}\right)}\,\|a_k M_k\hat{v}^k\|_{M_k^{-1}},
\end{aligned}$$

where the second inequality follows from (4.11). By combining the latter relation with (4.13) and using also (4.2), we obtain

$$\begin{aligned}
&\operatorname{dist}(z^{k+1},T^{-1}(0))_{M_k^{-1}}\\
&\le\left(\sqrt{\left(1+\frac{L_1^2}{\lambda_l^2 b_k^2}\right)\left(\tau_k^2-2\tau_k+\frac{1}{(1-\sigma_k^2)^2}\right)}+\frac{L_1}{\lambda_l b_k(1-\sigma_k^2)}\right)\|a_k M_k\hat{v}^k\|_{M_k^{-1}}.
\end{aligned}\tag{4.13}$$

Define

$$\mu_k:=\sqrt{\alpha_k^2+1}\sqrt{\beta_k^2-1}+\alpha_k\beta_k,\tag{4.14}$$

where

$$\alpha_k:=\frac{L_1\left(1+\sqrt{1-(1-\sigma_k^2)^2}\right)}{\lambda_l c_k(1-\sigma_k^2)(1-\theta)}\le\frac{L_1\left(1+\sqrt{1-(1-\bar{\sigma}^2)^2}\right)}{\lambda_l c(1-\bar{\sigma}^2)(1-\theta)}=:\alpha,\tag{4.15}$$

$$\beta_k:=\frac{1}{1-\sigma_k^2}\le\frac{1}{1-\bar{\sigma}^2}=:\beta.\tag{4.16}$$

With those definitions, by using (4.13) and (4.3), we conclude that

$$\operatorname{dist}(z^{k+1},T^{-1}(0))_{M_k^{-1}}\le\mu_k\|a_k M_k\hat{v}^k\|_{M_k^{-1}}.\tag{4.17}$$

Let $\bar{z}^k:=P_{T^{-1}(0),M_k^{-1}}(z^k)$. By Lemma 2.3, it holds that

$$\operatorname{dist}(z^k,T^{-1}(0))^2_{M_k^{-1}}\ge\|\bar{z}^k-z^{k+1}\|^2_{M_k^{-1}}+(1-(1-\tau_k)^2)a_k^2\|M_k\hat{v}^k\|^2_{M_k^{-1}}\tag{4.18}$$

$$\ge\operatorname{dist}(z^{k+1},T^{-1}(0))^2_{M_k^{-1}}+(1-\theta^2)a_k^2\|M_k\hat{v}^k\|^2_{M_k^{-1}}.\tag{4.19}$$

By using (4.17), we then conclude that

$$\operatorname{dist}(z^k,T^{-1}(0))^2_{M_k^{-1}}\ge\left(1+\frac{1-\theta^2}{\mu_k^2}\right)\operatorname{dist}(z^{k+1},T^{-1}(0))^2_{M_k^{-1}}.\tag{4.20}$$

Therefore,

$$\operatorname{dist}(z^{k+1},T^{-1}(0))\le\frac{\mu_k\sqrt{\lambda_u}}{\sqrt{\lambda_l(\mu_k^2+1-\theta^2)}}\operatorname{dist}(z^k,T^{-1}(0)).\tag{4.21}$$

By the definitions (4.15) and (4.16), by taking c_k sufficiently large we can make α_k arbitrarily small, and by taking σ_k sufficiently small we can make β_k arbitrarily close

to one. By the definition (4.14), this means that we can make μ_k arbitrarily small, so that the scalar in the right-hand side of (4.21) is less than one. Then (4.21) shows that the sequence $\{\operatorname{dist}(z^k, T^{-1}(0))\}$ converges linearly to zero. Also, the inequality (4.19) shows that this sequence is Fejér-monotone with respect to the set $T^{-1}(0)$ (for the given norm). For Fejér-monotone sequences, linear convergence of $\{\operatorname{dist}(z^k, T^{-1}(0))\}$ is equivalent to the linear convergence rate of $\{z^k\}$ to its limit (see, e.g., [1]).

By the same argument as above, if $c_k \to \infty$ and $\sigma_k \to 0$, then $\mu_k \to 0$, and (4.21) shows a superlinear convergence rate.

Assume now that the condition (4.9) holds. Then

$$\begin{aligned}\frac{1}{(1+\eta_k)}\operatorname{dist}(z, T^{-1}(0))^2_{M_k^{-1}} &= \inf_{y\in T^{-1}(0)} \frac{1}{(1+\eta_k)}\|z-y\|^2_{M_k^{-1}} \\ &\le \inf_{y\in T^{-1}(0)} \|z-y\|^2_{M_{k+1}^{-1}} \\ &= (1+\eta_k)\operatorname{dist}(z, T^{-1}(0))^2_{M_{k+1}^{-1}} \\ &\le \inf_{y\in T^{-1}(0)} (1+\eta_k)\|z-y\|^2_{M_k^{-1}} \\ &= (1+\eta_k)\operatorname{dist}(z, T^{-1}(0))^2_{M_k^{-1}}. \end{aligned} \tag{4.22}$$

Define

$$\mu = \sqrt{\alpha^2+1}\sqrt{\beta^2-1} + \alpha\beta,$$

with α and β given by (4.15) and (4.16), respectively. Note that $\mu > \mu_k$ for all k.

Since $\prod_{i=0}^{\infty}(1+\eta_i) < \infty$, there exists $k_2 \in \mathbb{N}$ such that

$$\prod_{i=k_2}^{\infty}(1+\eta_i) < \frac{\sqrt{\mu^2+1-\theta^2}}{\mu}.$$

From (4.20), by applying (4.22) consecutively, for any $k \ge k_0 := \max\{k_1, k_2\}$, we have

$$\begin{aligned}&\left(\prod_{i=k_0}^{\infty}\frac{1}{(1+\eta_i)}\right)\operatorname{dist}(z^{k+1}, T^{-1}(0))^2_{M_{k_0}^{-1}} \\ &\le \left(\prod_{i=k_0}^{k-1}\frac{1}{(1+\eta_i)}\right)\operatorname{dist}(z^{k+1}, T^{-1}(0))^2_{M_{k_0}^{-1}} \\ &\le \operatorname{dist}(z^{k+1}, T^{-1}(0))^2_{M_k^{-1}} \\ &\le \frac{\mu^2}{\mu^2+1-\theta^2}\operatorname{dist}(z^k, T^{-1}(0))^2_{M_k^{-1}} \\ &\le \left(\prod_{i=k_0}^{k-1}(1+\eta_i)\right)\frac{\mu^2}{\mu^2+1-\theta^2}\operatorname{dist}(z^k, T^{-1}(0))^2_{M_{k_0}^{-1}} \\ &\le \left(\prod_{i=k_0}^{\infty}(1+\eta_i)\right)\frac{\mu^2}{\mu^2+1-\theta^2}\operatorname{dist}(z^k, T^{-1}(0))^2_{M_{k_0}^{-1}}. \end{aligned}$$

In particular,

$$\operatorname{dist}(z^{k+1}, T^{-1}(0))_{M_{k_0}^{-1}} \le \nu \operatorname{dist}(z^k, T^{-1}(0))_{M_{k_0}^{-1}},$$

where

$$\nu := \frac{\mu}{\sqrt{\mu^2 + 1 - \theta^2}} \prod_{i=k_0}^{\infty} (1 + \eta_i) < 1,$$

as claimed. □

5. A variable metric proximal Newton method. In this section, we show how the proposed variable metric approach can be used to obtain a computational advantage when solving a system of monotone differentiable equations

$$F(x) = 0, \tag{5.1}$$

where $F : \mathbb{R}^n \to \mathbb{R}^n$. Problems of this type appear, for example, in smooth multiplier methods for monotone complementarity problems [14]. We start with describing the method and giving its theoretical justification and then report on our numerical experiments.

5.1. Description and justification of the method. In [25, 29], it has been shown that hybrid inexact proximal point schemes (with a fixed metric) can be used to construct Newton methods for monotone problems with a very attractive combination of global and local convergence properties. In particular, global convergence from any starting point to a solution is guaranteed, regardless of any degeneracy along the trajectory, which is not true in the case of more standard merit function-based globalizations that can get stuck at stationary points of the function that are not global minimizers. Fast local convergence for nondegenerate problems is also preserved, in a natural way. We refer the reader to [25, 29] for more detailed discussion.

When the Newton step is computed for the proximal subproblem (with the fixed metric $M_k = I$)

$$c_k F(z) + (z - z^k) = 0,$$

as in [25], one needs to solve the system of linear equations

$$c_k F(z^k) + (c_k \nabla F(z^k) + I)d = 0, \tag{5.2}$$

with respect to $d \in \mathbb{R}^n$. The crucial point is that, under natural assumptions, this single Newton step is enough to obtain an acceptable approximate solution of the proximal subproblem. Note that the above system is, in general, *asymmetric*. For future comparison, note that to compute LU factorization of the matrix $c_k \nabla F(z^k) + I$ and then the solution d^k, the number of arithmetic operations required is $2(n^3/3 + n^2)$. If to solve the linear system one uses instead of matrix factorization the conjugate gradient method, calculation of $(\nabla F(z^k))^\top \nabla F(z^k)$ is needed. Apart from extra computational cost (which is not negligeable when n is large), the latter is in general a dense matrix even when $\nabla F(z^k)$ is sparse. In what follows, we show how choosing a special variable metric can reduce the number of calculations in the case of using matrix factorizations and can preserve sparsity if the conjugate gradient method is used.

The idea is to choose a metric in such a way that, instead of solving a general asymmetric linear system, we will have to solve one triangular system and one symmetric system (with a positive definite matrix). As we shall see, this has a number of advantages.

Consider the proximal subproblem

$$0 = c_k M_k F(z) + (z - z^k). \tag{5.3}$$

We shall compute the Newton step for its equivalent formulation

$$0 = c_k F(z) + A_k(z - z^k),$$

where A_k plays the role of the inverse of M_k (no matrices are actually inverted, of course; we simply choose A_k and work with it throughout, as explained next). The Newton step for the latter equation is given by

$$c_k F(z^k) + (c_k \nabla F(z^k) + A_k) d^k = 0. \tag{5.4}$$

In what follows, we shall show that, with proper choices of parameters, the point

$$y^k = z^k + d^k \tag{5.5}$$

is an acceptable approximate solution of (5.3), in the sense of Algorithm 3.1 (even more specifically, in the sense of Proposition 3.1). Then the next iterate is given by

$$z^{k+1} = z^k - c_k M_k F(y^k),$$

which can be implemented as solving the system of linear equations

$$c_k F(y^k) + A_k s = 0, \tag{5.6}$$

with respect to $s \in \mathbb{R}^n$, and setting

$$z^{k+1} = z^k + s^k. \tag{5.7}$$

As A_k we suggest to take the symmetrization of the upper triangular part of the matrix $-c_k \nabla F(z^k)$ with appropriate diagonal elements, so that it is positive definite. One choice is

$$(A_k)_{i,j} := \begin{cases} -c_k(\nabla F(z^k))_{i,j} & \text{for} \quad i < j, \\ (A_k)_{j,i} & \text{for} \quad i > j, \\ 1 + \sum_{i \neq j} |(A_k)_{i,j}| & \text{for} \quad i = j. \end{cases} \tag{5.8}$$

Since A_k is symmetric and strictly diagonally dominant, it is positive definite by the Gerschgorin theorem [12, Theorem 3.5.9], and

$$\lambda_{min}(A_k) \geq 1. \tag{5.9}$$

The proposed implementation, therefore, consists of solving the linear system (5.4) with the triangular matrix $c_k \nabla F(z^k) + A_k$ and the linear system (5.6) with the symmetric positive definite matrix A_k. If the Cholesky factorization is used for the latter, the total cost of the iteration is $n^3/3 + 2n^2 + n^2/2$ arithmetic operations. The savings compared to the fixed metric (asymmetric) implementation discussed above amounts to $n^2(n/3 - 1/2)$, which is significant for large n. If instead of matrix factorization the conjugate gradient method is used to solve (5.6), it is important that it works directly with the symmetric matrix A_k, which is sparse if $\nabla F(z^k)$ is also. Recall that, in the case of solving the asymmetric system, the method has to work with $(c_k \nabla F(z^k) + I)^\top (c_k \nabla F(z^k) + I)$, which is in general dense even when $\nabla F(z^k)$ is sparse.

To validate our proposal, it remains to show that the single Newton step defined by (5.4) produces a point acceptable by the approximation criterion of Algorithm 3.1 and that this strategy does not increase too much the overall number of iterations of the method as compared to the asymmetric fixed metric implementation. We deal with the first issue next and then present some numerical experiments to address the second.

Let $M_k = A_k^{-1}$. By (5.9), we have

$$\lambda_{max}(M_k) \leq 1. \tag{5.10}$$

In particular, we can take $\lambda_u = 1$ in Algorithm 3.1. Since d^k is the solution of the linear system (5.4), we have

$$d^k = y^k - z^k = -c_k M_k F(z^k) - c_k M_k \nabla F(z^k) d^k. \tag{5.11}$$

To prove the claim that this Newton step is sufficient to solve the proximal subproblem within the required tolerance, we have to show that

$$\|c_k M_k F(y^k) + d^k\|^2_{M_k^{-1}} \leq \sigma_k^2 \left(\|c_k M_k F(y^k)\|^2_{M_k^{-1}} + \|d^k\|^2_{M_k^{-1}} \right). \tag{5.12}$$

Let ∇F be Lipschitz-continuous with modulus $\gamma > 0$ (on the bounded set containing the sequences $\{z^k\}$ and $\{y^k\}$, whose boundedness has been already established). It then holds that

$$\|F(y^k) - F(z^k) - \nabla F(z^k) d^k\| \leq \frac{\gamma}{2} \|d^k\|^2.$$

Since it follows from (5.11) that

$$-c_k F(z^k) - c_k \nabla F(z^k) d^k = M_k^{-1} d^k,$$

we obtain

$$\|c_k F(y^k) + M_k^{-1} d^k\| \leq \frac{\gamma c_k}{2} \|d^k\|^2. \tag{5.13}$$

Furthermore, by recalling (5.10), we have

$$\|c_k F(y^k) + M_k^{-1} d^k\|^2 \geq \|c_k F(y^k) + M_k^{-1} d^k\|^2_{M_k} = \|c_k M_k F(y^k) + d^k\|^2_{M_k^{-1}}. \tag{5.14}$$

Also, by using (5.10) and (5.11), we obtain

$$\begin{aligned} \|d^k\|^2 &\leq \|d^k\|^2_{M_k^{-1}} \\ &= \langle d^k, M_k^{-1}(-c_k M_k F(z^k) - c_k M_k \nabla F(z^k) d^k) \rangle \\ &= -c_k \langle d^k, F(z^k) \rangle - c_k \langle d^k, \nabla F(z^k) d^k \rangle \\ &\leq c_k \|d^k\| \|F(z^k)\|, \end{aligned}$$

where we have used the fact that $\nabla F(z^k)$ is positive semidefinite (by the monotonicity of F). Hence,

$$\|d^k\| \leq c_k \|F(z^k)\|.$$

By combining this relation with (5.13) and (5.14), we conclude that

$$\|c_k M_k F(y^k) + d^k\|_{M_k^{-1}} \leq \frac{\gamma c_k^2 \|F(z^k)\|}{2} \|d^k\| \leq \frac{\gamma c_k^2 \|F(z^k)\|}{2} \|d^k\|_{M_k^{-1}},$$

where (5.10) was also taken into account. Therefore, by choosing the regularization parameter

$$0 < c_k \leq \frac{\sqrt{2}\sigma_k}{\sqrt{\gamma \|F(z^k)\|}},$$

we obtain (5.12). This analysis also shows that we are in the setting of Proposition 3.1, so that the step $z^{k+1} = z^k + c_k M_k F(y^k)$ is admissible (implemented above as solving the linear system (5.6)).

If an estimate for the Lipschitz constant γ of ∇F is not available, c_k can be obtained by an Armijo-type line-search procedure. Alternatively, instead of making one Newton step for each subproblem, we can make several, until the relative error approximation criterion is satisfied. In our computational experience, however, one Newton step was always enough. Moreover, by assuming the nonsingularity of ∇F at the solution, for k large enough one can take $c_k = \frac{\sqrt{2}\sigma_k}{\sqrt{\|F(z^k)\|}}$, without any line search, and make a single Newton step. The superlinear rate of convergence can be established by analysis analogous to [25].

5.2. Numerical experiments. We have compared the proximal Newton methods, with a fixed metric and a variable metric, on the following examples.

Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be given by

$$F(z) = \tilde{F}(z) + Hz,$$

where

$$\tilde{F}_i(z) = \frac{1 + (-1)^{i+1}}{2} f(z_i),$$

$f : \mathbb{R} \to \mathbb{R}$ is a monotone function with a Lipschitz-continuous derivative, and H is the $n \times n$ matrix given by

$$(H)_{i,j} = \begin{cases} n/2 & \text{for } i = j = 1, \\ 5n & \text{for } i = 1 \text{ and } j = n, \\ -5n & \text{for } i = n \text{ and } j = 1, \\ n + i - 1 & \text{for } i = j \text{ and } i \notin \{1, n\}, \\ 1 & \text{for } j = n \text{ and } i \notin \{1, n\}, \\ 1 & \text{for } j < i \text{ and } i \neq n, \\ -1 & \text{for } i = n \text{ and } j \notin \{1, n\}, \\ 0 & \text{elsewhere.} \end{cases}$$

It can be seen that H is positive semidefinite (because $(H + H^\top)/2$ is diagonally dominant), but it is not positive definite (because $e_n^\top H e_n = 0$, where e_n is the nth vector of the canonical basis). This fact and the monotonicity of f imply that F is monotone. Note that, for $n = 2k$ with $k \in \mathbb{N}$, F is not strictly monotone, even if f is strictly monotone.

It can be seen that its Jacobian $\nabla F(z)$ is Lipschitz-continuous with the same Lipschitz constant as f', and, for any $z \in \mathbb{R}^n$, $\nabla F(z)$ is a nonsymmetric matrix, with a sparse upper triangular part.

We have coded both the Newton proximal method (NPM) and the variable metric Newton proximal method (VMNPM) by using Scilab 4.0 (INRIA-ENPC, see www.scilab.org). An iteration of NPM consists of solving the system of equations (5.2), while VMNPM is the procedure given by (5.4)–(5.7), with A_k defined in (5.8). For both methods, the regularization parameter is taken as $c_k = \sqrt{2/\|F(z^k)\|}$.

In the case of solving linear systems by matrix factorization, the comparison is exactly as predicted by the arithmetic operations counts, mentioned above. The variable metric approach requires more iterations, but already for moderate dimensions

TABLE 5.1

$f(x) = x + \exp(-x^2)$

Dim	NPM			VMNPM			T_1/T_2
	Iter	T_1	$\|F\|$	Iter	T_2	$\|F\|$	
100	4	0.16	3.98e-008	20	0.20	8.12e-008	0.77
300	4	1.34	3.40e-008	22	1.11	3.57e-008	1.21
500	4	4.16	4.04e-008	22	3.59	4.13e-008	1.16
700	4	9.06	4.32e-008	22	7.28	6.74e-008	1.24
900	4	16.22	4.88e-008	23	12.72	4.62e-008	1.28
1100	4	26.16	5.37e-008	23	19.06	5.10e-008	1.37
1300	4	39.00	6.81e-008	23	26.38	5.05e-008	1.48
1500	4	55.45	6.94e-008	23	35.13	4.65e-008	1.58
1700	4	75.94	9.14e-008	23	44.84	4.39e-008	1.69
1900	4	100.70	9.59e-008	23	55.91	5.12e-008	1.80

$f(x) = 2\arctan(x + 1)$

Dim	NPM			VMNPM			T_1/T_2
	Iter	T_1	$\|F\|$	Iter	T_2	$\|F\|$	
100	4	0.13	7.42e-008	20	0.17	6.63e-008	0.73
300	4	1.36	4.80e-008	22	1.16	6.50e-008	1.18
500	4	4.38	5.77e-008	23	3.78	1.52e-008	1.16
700	4	9.22	6.42e-008	23	7.91	2.25e-008	1.17
900	4	16.45	6.51e-008	23	13.22	5.09e-008	1.24
1100	4	26.38	6.71e-008	23	19.66	9.95e-008	1.34
1300	4	39.27	7.14e-008	24	28.55	3.95e-008	1.38
1500	4	55.78	8.37e-008	24	37.89	3.18e-008	1.47
1700	4	76.92	9.02e-008	24	49.11	2.79e-008	1.57
1900	4	101.33	1.15e-007	24	60.64	3.97e-008	1.67

$f(x) = \frac{1}{2}x\sqrt{x^2+5} + \frac{5}{2}\ln(x + \sqrt{x^2+5})$

Dim	NPM			VMNPM			T_1/T_2
	Iter	T_1	$\|F\|$	Iter	T_2	$\|F\|$	
100	4	0.14	8.04e-008	20	0.20	9.22e-008	0.69
300	4	1.36	5.71e-008	23	1.19	1.78e-008	1.14
500	4	4.22	6.93e-008	23	3.80	5.05e-008	1.11
700	4	9.13	7.99e-008	24	8.08	5.35e-008	1.13
900	4	16.38	8.47e-008	24	13.30	4.05e-008	1.23
1100	4	26.31	8.51e-008	24	19.78	4.18e-008	1.33
1300	4	39.25	8.98e-008	24	27.94	9.05e-008	1.40
1500	4	55.73	9.75e-008	25	38.22	5.00e-008	1.46
1700	4	76.27	1.08e-007	25	48.89	3.49e-008	1.56
1900	4	101.08	1.18e-007	25	60.77	2.64e-008	1.66

(say, $n = 500$) the cheaper iteration cost starts to pay off, with the advantage growing with n, as predicted by the operations counts. We shall not report this comparison here, for the sake of brevity.

Instead, we shall report results for solving the linear systems by the conjugate gradient method. The Scilab `sparse` utility is used to take advantage of structure. As already pointed out, the matrix $(c_k\nabla F(z^k) + I)^\top(c_k\nabla F(z^k) + I)$ in the fixed metric approach is essentially dense, while the matrix A_k in the variable metric approach preserves structure.

The comparison of the respective performances, for three different choices of f, on a 1.66 GHz, 512 MB RAM Intel Centrino processor PC is shown in Table 5.1. The first column shows the dimension, then the number of iterations, the computation time in seconds, and the norm of the residual at termination. The last column shows the ratio between the computational times.

Figure 5.1 compares the computational time evolution for both methods. The performance of the NPM is almost the same for the three functions involved, and

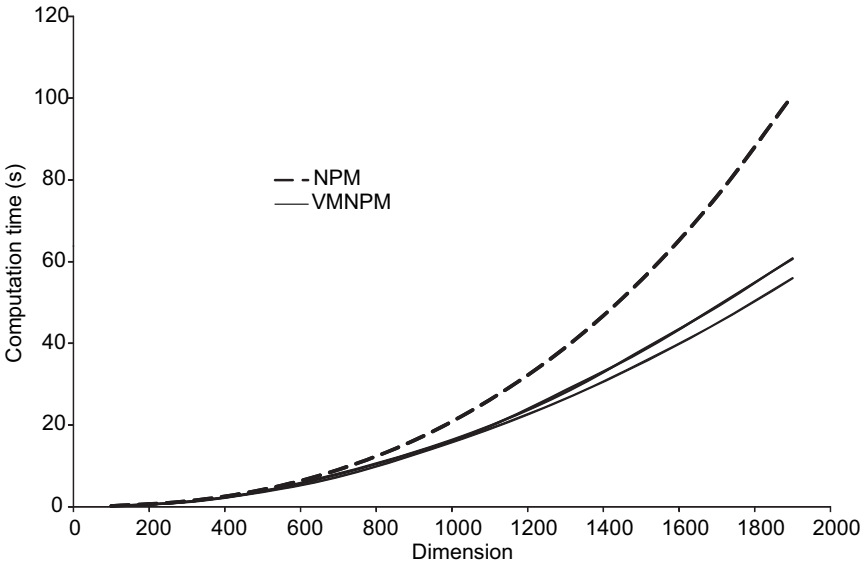


FIG. 5.1. *Performance comparison.*

it is not distinguishable in the graphic scale, while the performance of the VMNPM presents little variations for the three examples. As we have anticipated, the variable metric proximal Newton method outperforms the Newton proximal method already for moderate dimensions, with the advantage becoming more and more significant as n grows.

REFERENCES

[1] H. H. BAUSCHKE, *Projection algorithms: Results and open problems*, in Inherently Parallel Algorithms in Feasibility and Optimization and their Applications, Stud. Comput. Math. 8, D. Butnariu, Y. Censor, and S. Reich, eds., Elsevier Science B. V., Amsterdam, 2001, pp. 11–22.

[2] J. F. BONNANS, J. CH. GILBERT, C. LEMARÉCHAL, AND C. SAGASTIZÁBAL, *A family of variable metric proximal point methods*, Math. Program., 68 (1995), pp. 15–47.

[3] J. F. BONNANS, J. CH. GILBERT, C. LEMARÉCHAL, AND C. SAGASTIZÁBAL, *Numerical Optimization: Theoretical and Practical Aspects*, Springer-Verlag, Berlin, 2003.

[4] R. S. BURACHIK AND A. N. IUSEM, *A generalized proximal point algorithm for the variational inequality problem in a Hilbert space*, SIAM J. Optim., 8 (1998), pp. 197–216.

[5] R. S. BURACHIK, A. N. IUSEM, AND B. F. SVAITER, *Enlargement of monotone operators with applications to variational inequalities*, Set-Valued Anal., 5 (1997), pp. 159–180.

[6] R. S. BURACHIK, C. A. SAGASTIZÁBAL, AND B. F. SVAITER, *ε-Enlargements of maximal monotone operators: Theory and applications*, in Reformulation - Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, M. Fukushima and L. Qi, eds., Kluwer Academic Publishers, Norwell, MA, 1999, pp. 25–44.

[7] J. V. BURKE AND M. QIAN, *A variable metric proximal point algorithm for monotone operators*, SIAM J. Control Optim., 37 (1999), pp. 353–375.

[8] J. V. BURKE AND M. QIAN, *On the local super-linear convergence of a matrix secant implementation of the variable metric proximal point algorithm for monotone operators*, in Reformulation - Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, M. Fukushima and L. Qi, eds., Kluwer Academic Publishers, Norwell, MA, 1999, pp. 317–334.

[9] J. V. BURKE AND M. QIAN, *On the superlinear convergence of the variable metric proximal point algorithm using Broyden and BFGS matrix secant updating*, Math. Program., 88 (2000), pp. 157–181.

[10] X. Chen and M. Fukushima, *Proximal quasi-Newton methods for nondifferentiable convex optimization*, Math. Program., 85 (1999), pp. 313–334.
[11] R. Cominetti, *Coupling the proximal point algorithm with approximation methods*, J. Optim. Theory Appl., 95 (1997), pp. 581–600.
[12] J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
[13] J. Eckstein and D. P. Bertsekas, *On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators*, Math. Program., 55 (1992), pp. 293–318.
[14] J. Eckstein and M. C. Ferris, *Smooth methods of multipliers for complementarity problems*, Math. Program., 86 (1999), pp. 65–90.
[15] D. Gabay, *Applications of the method of multipliers to variational inequalities*, in Augmented Lagrangian Methods: Applications to the Solution of Boundary-Value Problems, M. Fortin and R. Glowinski, eds., North–Holland, Amsterdam, 1983, pp. 299–331.
[16] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms*, Springer-Verlag, Berlin, 1993.
[17] C. Lemaréchal and C. Sagastizábal, *Variable metric bundle methods: From conceptual to implementable forms*, Math. Program., 76 (1997), pp. 393–410.
[18] P. L. Lions and B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM J. Numer. Anal., 16 (1979), pp. 964–979.
[19] B. Martinet, *Regularisation d'inequations variationelles par approximations successives*, Revue Française d'Informatique et de Recherche Opérationelle, 4 (1970), pp. 154–159.
[20] L. Qi and X. Chen, *A preconditioning proximal Newton's method for nondifferentiable convex optimization*, Math. Program., 76 (1995), pp. 411–430.
[21] R. T. Rockafellar, *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Math. Oper. Res., 1 (1976), pp. 97–116.
[22] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.
[23] R. T. Rockafellar and J.-B. Wets, *Variational Analysis*, Springer-Verlag, New York, 1997.
[24] M. V. Solodov, *A class of decomposition methods for convex optimization and monotone variational inclusions via the hybrid inexact proximal point framework*, Optim. Methods Softw., 19 (2004), pp. 557–575.
[25] M. V. Solodov and B. F. Svaiter, *A globally convergent inexact Newton method for systems of monotone equations*, in Reformulation - Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, M. Fukushima and L. Qi, eds., Kluwer Academic Publishers, Norwell, MA, 1999, pp. 355–369.
[26] M. V. Solodov and B. F. Svaiter, *A hybrid approximate extragradient–proximal point algorithm using the enlargement of a maximal monotone operator*, Set-Valued Anal., 7 (1999), pp. 323–345.
[27] M. V. Solodov and B. F. Svaiter, *A hybrid projection–proximal point algorithm*, J. Convex Anal., 6 (1999), pp. 59–70.
[28] M. V. Solodov and B. F. Svaiter, *Error bounds for proximal point subproblems and associated inexact proximal point algorithms*, Math. Program., 88 (2000), pp. 371–389.
[29] M. V. Solodov and B. F. Svaiter, *A truly globally convergent Newton-type method for the monotone nonlinear complementarity problem*, SIAM J. Optim., 10 (2000), pp. 605–625.
[30] M. V. Solodov and B. F. Svaiter, *A unified framework for some inexact proximal point algorithms*, Numer. Funct. Anal. Optim., 22 (2001), pp. 1013–1035.
[31] J. E. Spingarn, *Applications of the method of partial inverses to convex programming*, Math. Program., 32 (1985), pp. 199–223.
[32] P. Tossings, *The perturbed proximal point algorithm and some of its applications*, Appl. Math. Optim., 29 (1994), pp. 125–159.
[33] P. Tseng, *Applications of a splitting algorithm to decomposition in convex programming and variational inequalities*, SIAM J. Control Optim., 29 (1991), pp. 119–138.
[34] P. Tseng, *A modified forward-backward splitting method for maximal monotone mappings*, SIAM J. Control Optim., 38 (2000), pp. 431–446.

SIAM J. OPTIM.
Vol. 19, No. 1, pp. 261–280

A CONDITION NUMBER FOR MULTIFOLD CONIC SYSTEMS*

DENNIS CHEUNG†, FELIPE CUCKER‡, AND JAVIER PEÑA§

Abstract. Let $A : Y \to X$ be a linear map and $K \subseteq X$ be a regular closed convex cone. Consider the problem of finding a nontrivial solution to the conic feasibility problem $Ay \in K$. Condition numbers for this problem (as well as for related ones) are studied to quantify various issues concerning properties of the conic feasibility problem. Some issues especially relevant are the behavior of the problem under data perturbations, the geometry of the set of solutions, and the complexity analyses of algorithms that solve the problem. In this paper we define and characterize a condition number that exploits the possible factorization of K as a product of simpler cones. This condition number extends both Renegar's condition number and the one we defined in [*Math. Program.*, 91 (2001), pp. 163–174] for polyhedral conic systems. We see these results as a step in developing a theory of conditioning that takes into account the structure of the problem.

Key words. condition numbers, conic systems

AMS subject classifications. Primary, 90C25; Secondary, 90C31

DOI. 10.1137/060665427

1. Introduction.

1.1. Multifold conic systems and condition. Let X, Y be real finite-dimensional vector spaces (not necessarily of the same dimension) endowed with norms $\| \ \|_X$ and $\| \ \|_Y$, and let $K \subseteq X$ be a regular closed convex cone (a precise definition is in section 2 below). Denote by $L(Y, X)$ the space of linear maps from Y to X endowed with the operator norm. Given $A \in L(Y, X)$, consider the feasibility problem: decide whether there exists a nontrivial $y \in Y$ satisfying

$$Ay \in K. \tag{1.1}$$

This format encompasses, after homogenization, a large class of feasibility problems. For example, the linear programming feasibility problem corresponds to $K = \mathbb{R}^n_+$, the nonnegative orthant in $\mathbb{R}^n$, and semidefinite programming corresponds to $K = \mathbf{S}^n_+$, the set of $n \times n$ positive semidefinite matrices. Consider also the *alternative* feasibility problem

$$A^*x^* = 0, \ x^* \in K^*, \tag{1.2}$$

where X^*, Y^* are the dual spaces of X, Y, respectively, $A^* \in L(X^*, Y^*)$ is the adjoint of A, and $K^* \subseteq X^*$ is the dual cone of K.

The problem (1.1) is strictly feasible if there exists $y \in Y$ such that $Ay \in \mathsf{int}(K)$. Let $\mathcal{D}$ denote the set of instances $A \in L(Y, X)$ for which (1.1) is strictly feasible.

*Received by the editors July 19, 2006; accepted for publication (in revised form) November 12, 2007; published electronically March 19, 2008.
http://www.siam.org/journals/siopt/19-1/66542.html
†United International College, Tang Jia Wan, Zhuhai, Guandong Province, People's Republic of China (dennisc@uic.edu.hk).
‡Department of Mathematics, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong (macucker@math.cityu.edu.hk). This author has been partially funded by SRG grant 7001860.
§Tepper School of Business, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213-3890 (jfp@andrew.cmu.edu).

Observe that $A \in \mathcal{D}$ if and only if $AY - K = X$, i.e., if and only if the conic system

$$Ay - c \in K, \tag{1.3}$$

is feasible for every $c \in X$.

Likewise, the problem (1.2) is strictly feasible if there exists $x^* \in \mathsf{int}(K^*)$ such that $A^*x^* = 0$. Let $\mathcal{P}$ denote the set of instances $A \in L(Y, X)$ such that A^* is surjective and (1.2) is strictly feasible. Observe that $A \in \mathcal{P}$ if and only if $A^*K^* = Y^*$, i.e., if and only if the conic system

$$A^*x^* = b^*,\ x^* \in K^*, \tag{1.4}$$

is feasible for every $b^* \in Y^*$.

It is easy to see that the sets $\mathcal{D}$ and $\mathcal{P}$ are open, and $\mathcal{P} = \overline{D}^c$. The sets $\mathcal{D}$ and $\mathcal{P}$ are the set of "well-posed" feasible instances for problems (1.1) and (1.2), respectively. The boundary $\Sigma := \partial\mathcal{D} = \partial\mathcal{P}$ is the set of "ill-posed" instances. Given $A \in \Sigma$, arbitrarily small perturbations of A may yield instances in both $\mathcal{D}$ and $\mathcal{P}$.

The feasibility problems (1.1) and (1.2) can be solved via algorithms (such as interior-point or ellipsoid methods). The theoretical running time of such algorithms grows as A approaches Σ. Consequently, a complexity analysis of these algorithms has been carried out in terms of a measure capturing this distance. Similar remarks hold as well, with natural modifications, for linear conic optimization problems with constraints of the form (1.3) or (1.4). A general analysis of such a type for interior-point methods is due to Renegar [30, 31], who introduced the condition number

$$C(A) = \frac{\|A\|}{\mathsf{dist}(A, \Sigma)} = \frac{\|A\|}{\min_{\widetilde{A} \in \Sigma} \|\widetilde{A} - A\|}. \tag{1.5}$$

Renegar's condition number is thus the normalized inverse of the distance to ill-posedness. The condition number $C(A)$ can also be used in the complexity analysis of the ellipsoid method [17] and in the round-off analysis of interior-point algorithms [10]. For the linear programming feasibility problem ($K = \mathbb{R}^n_+$), the quantities $C(A)$ and $\ln C(A)$ have also been analyzed as random variables when A is random [7, 13]. Bounds for the expected value of $C(A)$ (or for that of $\ln C(A)$) yield average case bounds for the algorithms mentioned above.

It is often the case that a feasibility problem of the form (1.1) is actually the coupling of a number of similar feasibility problems. More precisely, if $X = X_1 \times \cdots \times X_r$ and $K = K_1 \times \cdots \times K_r$, where each $K_j \subseteq X_j$ is a regular closed convex cone, then (1.1) can be written as

$$\begin{aligned} A_1 y &\in K_1 \\ &\vdots \\ A_r y &\in K_r, \end{aligned} \tag{1.6}$$

where each $A_j \in L(Y, X_j)$ is the projection of $A \in L(Y, X)$ onto $L(Y, X_j)$. In this *multifold* case, it may well be the case that $C(A)$ is large, but a natural preconditioning, such as component normalization, could remove the seemingly bad conditioning. This limitation of $C(A)$ may yield a nonessential overestimate on the conditioning of the problem (1.6). The latter in turn often leads to results concerning the geometry of the set of feasible solutions, as well as complexity estimates of algorithms that are overly

conservative. Consequently, some condition-based analyses such as those in [10, 27] are presented in terms of the condition number of a problem after performing some appropriate preprocessing steps on the data. In the case of linear programming (i.e., when $X = \mathbb{R}^n$, $K = \mathbb{R}^n_+$, $r = n$, and $K_j = \mathbb{R}_+$) another condition number $\mathscr{C}(A)$ was introduced in [6] (extending ideas in [19]) exploiting the multifold structure of $\mathbb{R}^n_+$. This condition number is close in spirit to $C(A)$ but is invariant under row scaling and is defined (in the feasible case) in terms of a *best conditioned* solution to (1.6). This condition number can also be used in the analysis of algorithms (e.g., the analysis in [10] carries over to $\mathscr{C}(A)$) and has also been studied as a random variable [8, 11, 20].

In this paper we show that the definition and key characterization of $\mathscr{C}(A)$ extend to the general multifold conic system (1.6) for a particular class of norms in X. An immediate consequence of our results is a maxmin characterization of $C(A)$, which emphasizes the close relationship between $C(A)$ and $\mathscr{C}(A)$. In the special case when the multifold structure of (1.6) is ignored, our work is closely related to previous characterizations of the distance to ill-posedness of the system $Ay \in K$ by Freund and Vera [18] and by Cánovas et al. [5]. More precisely, Theorem 1.1 without scaling on the components of (1.6) follows from [18, Thms. 7 and 10]. Furthermore, Theorem 1.1 without scaling also holds for an infinite family of linear inequalities as was shown in [5, Thm. 7]. On the other hand, an extension of the condition number $\mathscr{C}(A)$ to general conic systems was proposed by Lara and Tunçel [22]. However, that condition number conveys only information about the geometry of the set of feasible solutions of (1.6) and does not have a direct relationship to the distance to ill-posedness of $Ay \in K$.

The central results in our paper, namely, Theorems 1.1 and 1.2, can be seen as steps in the development of a theory of *structured condition numbers* in the spirit introduced by Peña [28, 29] and Lewis [23].

1.2. Statement of the main results. Given a triple (X, K, e), with X a finite-dimensional normed space, $K \subseteq X$ a regular closed convex cone, and $e \in \mathsf{int}(K)$ a given point, define $\lambda_{\min} : X \to \mathbb{R}$ as

$$x \mapsto \max\{t \in \mathbb{R} : x - te \in K\}. \tag{1.7}$$

Notice that $\lambda_{\min}$ is *positively homogeneous*, i.e., it satisfies

$$\lambda_{\min}(sx) = s\lambda_{\min}(x) \quad \text{for all} \quad s \geq 0 \quad \text{and} \quad x \in X,$$

and *superlinear*, i.e., it satisfies

$$\lambda_{\min}(x + u) \geq \lambda_{\min}(x) + \lambda_{\min}(u) \quad \text{for all} \quad x, u \in X.$$

Notice also that $x \in K \Leftrightarrow \lambda_{\min}(x) \geq 0$ and $x \in \mathsf{int}(K) \Leftrightarrow \lambda_{\min}(x) > 0$.

We next proceed with conditioning.

For $j = 1, \ldots, r$, let $e_j \in \mathsf{int}(K_j)$ be fixed and $\lambda^j_{\min} : X_j \to \mathbb{R}$ denote the function (1.7) corresponding to the triple (X_j, K_j, e_j).

Let $\mathbb{R}_{++} = (0, +\infty)$ and $\alpha \in \mathbb{R}^r_{++}$ be given. Define the *condition value* of a point $y \in Y \setminus \{0\}$ to be

$$v_{A,\alpha}(y) := \min_{j=1,\ldots,r} \frac{\lambda^j_{\min}(A_j y)}{\alpha_j \|y\|}. \tag{1.8}$$

Observe that y is a strict solution to (1.6) if and only if $v_{A,\alpha}(y) > 0$. Define the best conditioned value $\overline{v}_{A,\alpha}$ to be

$$\overline{v}_{A,\alpha} := \max_{y \neq 0} v_{A,\alpha}(y).$$

Notice that $A \in \mathcal{D}$ if and only if $\overline{v}_{A,\alpha} > 0$, $A \in \Sigma$ if and only if $\overline{v}_{A,\alpha} = 0$, and $A \in \mathcal{P}$ if and only if $\overline{v}_{A,\alpha} < 0$. Notice also that this is valid for all $\alpha \in \mathbb{R}^r_{++}$.

For $a \in X$ and $\delta > 0$ let

$$\mathbb{B}_X(a, \delta) := \{x \in X : \|x - a\| \leq \delta\}.$$

We shall say that the triple (X, K, e) satisfies the *norm compatibility condition* if $\|e\| = 1$ and the following condition holds:

$$\mathbb{B}_X(e, 1) \subseteq K. \tag{NC}$$

We will see in section 2 below that a natural, canonical norm can be associated to any triple (X, K, e) such that the triple satisfies the norm compatibility condition for this canonical norm.

We are now ready to state our main results (we delay their proofs to section 3).

THEOREM 1.1. *If each one of the triples (X_j, K_j, e_j), $j = 1, \ldots, r$, satisfies the norm compatibility condition* (NC)*, then*

$$|\overline{v}_{A,\alpha}| = \min_{\widetilde{A} \in \Sigma} \max_{j=1,\ldots,r} \frac{\|A_j - \widetilde{A}_j\|}{\alpha_j}, \tag{1.9}$$

where the norms in the right-hand side are the operator norms induced by the norms in Y and X_j.

It is customary to define condition numbers either as a relativized distance to ill-posedness or as the condition of a best conditioned solution. Theorem 1.1 shows that both choices lead to the same notion by taking the *condition number with respect to α* to be

$$C_\alpha(A) := \frac{1}{|\overline{v}_{A,\alpha}|}$$

and requiring the norm in X to satisfy $\|(x_1, \ldots, x_r)\| = \max_{j=1,\ldots,r} \|x_j\|$. Note that the distance to ill-posedness in the right-hand side in Theorem 1.1 is relativized by the vector α.

In the previous development we have assumed that $\alpha_j > 0$ for $j = 1, \ldots, r$. From a perturbation theory viewpoint this corresponds to assuming that all of the A_j can be perturbed and that the magnitude of these perturbations are weighted (or relativized) by the α_j.

We next consider the case where some blocks are rigid, i.e., cannot be perturbed. This amounts to setting the corresponding α_j to zero. To that end, assume that $B \cup N = \{1, \ldots, r\}$ is a partition of $\{1, \ldots, r\}$, with $B \neq \emptyset$. Let $X_N = \prod_{j \in N} X_j$, $K_N = \prod_{j \in N} K_j$, and $A_N = \prod_{j \in N} A_j$. Write also $\alpha_N = (\alpha_j)_{j \in N}$ and $\alpha_B = (\alpha_j)_{j \in B}$. If we allow only perturbations in the blocks A_j for $j \in B$, then the following extension of Theorem 1.1 holds.

THEOREM 1.2. *Assume that each one of the triples (X_j, K_j, e_j), $j = 1, \ldots, r$, satisfies the norm compatibility condition. Then, for $A \notin \Sigma$,*

$$\left| \max_{\substack{A_N y \in K_N \\ y \neq 0}} \min_{j \in B} \frac{\lambda^j_{\min}(A_j y)}{\alpha_j \|y\|} \right| = \min_{\substack{\widetilde{A} \in \Sigma \\ \widetilde{A}_N = A_N}} \max_{j \in B} \frac{\|A_j - \widetilde{A}_j\|}{\alpha_j} \tag{1.10}$$

with the convention that the left-hand side above is $+\infty$ if

$$\{y : A_N y \in K_N,\ y \neq 0\} = \emptyset$$

and the right-hand side is $+\infty$ *if*

$$\{\widetilde{A} \in \Sigma : \widetilde{A}_N = A_N\} = \emptyset.$$

Furthermore, (1.10) *can be seen as a limit case of* (1.9) *in Theorem* 1.1. *More precisely, for* $A \notin \Sigma$,

$$\left| \max_{\substack{A_N y \in K_N \\ y \neq 0}} \min_{j \in B} \frac{\lambda^j_{\min}(A_j y)}{\alpha_j \|y\|} \right| = \lim_{\substack{\alpha_B \text{ fixed} \\ \alpha_N \downarrow 0}} |\overline{v}_{A,\alpha}|$$

and

$$\min_{\substack{\widetilde{A} \in \Sigma \\ \widetilde{A}_N = A_N}} \max_{j \in B} \frac{\|A_j - \widetilde{A}_j\|}{\alpha_j} = \lim_{\substack{\alpha_B \text{ fixed} \\ \alpha_N \downarrow 0}} \min_{\widetilde{A} \in \Sigma} \max_{j=1,\ldots,r} \frac{\|A_j - \widetilde{A}_j\|}{\alpha_j}.$$

Remark 1.3. The identity (1.10) in Theorem 1.2 does not necessarily hold if $A \in \Sigma$. For instance, consider the example $r = 2$, $B = \{1\}$, $\alpha_1 = 1$, $X_1 = \mathbb{R}$, $X_2 = Y = \mathbb{R}^2$, $K_1 = \mathbb{R}_+$, $K_2 = \mathbb{R}^2_+$, $A_1 = [\,1\ 0\,]$, and $A_2 = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right]$. In this example $A \in \Sigma$, and thus the right-hand side in (1.10) is zero, but a simple calculation shows that the left-hand side is one.

Nevertheless, when $A \in \Sigma$, a modified version of (1.10) holds if the set of ill-posed instances Σ is redefined by taking into account the relationship between the rigid part $A_N Y$ and the cone K_N.

1.3. Geometric interpretation of $\overline{v}_{A,\alpha}$. When $A \in \mathcal{D}$, any point $\overline{y} \in Y$ that satisfies

$$v_{A,\alpha}(\overline{y}) = \overline{v}_{A,\alpha}$$

can be interpreted as a best conditioned point for (1.1). Notice that in this case the best condition value $\overline{v}_{A,\alpha}$ satisfies

$$\overline{v}_{A,\alpha} = \max\{\delta > 0 : \exists y \in Y, \|y\| = 1, \text{ such that (s.t.) } \|x_i\| \leq \delta \alpha_i \Rightarrow Ay + x \in K\}.$$

In particular, $\mathbb{B}_Y(\overline{y}, \rho)$ is contained in the feasible solution set of (1.1) for $\rho = \min_{j=1,\ldots,r} \frac{\overline{v}_{A,\alpha}\alpha_j}{\|A_j\|}$. Furthermore, from Theorem 1.1 and [28, Thm. 2.11], it follows that, for $A \in \mathcal{D}$, the best condition value $\overline{v}_{A,\alpha}$ satisfies

$$\overline{v}_{A,\alpha} = \max\left\{\delta > 0 : \|x_i\| \leq \delta \alpha_i \Rightarrow x \in \{Ay - K : \|y\| \leq 1\}\right\}.$$

In other words, $\overline{v}_{A,\alpha} \prod \mathbb{B}_{X_i}(0, \alpha_i)$ is the largest multiple of $\prod \mathbb{B}_{X_i}(0, \alpha_i)$ contained in the set

$$\{Ay + K : \|y\| \leq 1\}.$$

Likewise, from Theorem 1.1 and [28, Thm. 2.8], it follows that, for $A \in \mathcal{P}$, the best condition value $\overline{v}_{A,\alpha}$ satisfies the following geometric property:

(1.11)
$$|\overline{v}_{A,\alpha}| = \max\left\{\delta > 0 : \|y^*\|^* \leq \delta \Rightarrow y^* \in \left\{A^* x^* : x^* \in K^*, \sum \alpha_i \|x_i^*\|^* \leq 1\right\}\right\}.$$

In other words, $\mathbb{B}_{Y^*}(0, |\overline{v}_{A,\alpha}|)$ is the largest (dual) ball centered at 0 that is contained in the set

$$\left\{A^*x^* : x^* \in K^*, \sum \alpha_i \|x_i^*\|^* \leq 1\right\}.$$

The identity (1.11) yields the existence of *well-conditioned* solutions to (1.2), as is stated more precisely in section 2.4 below.

We also note that the above geometric interpretations of $\overline{v}_{A,\alpha}$ can be extended to the case when some of the components of α are zero. Specifically, if $B \cup N = \{1, \ldots, r\}$ is a partition of $\{1, \ldots, r\}$ such that $\alpha_B > 0$ and $\alpha_N = 0$, then the statements above hold as long as $\overline{v}_{A,\alpha}$ is replaced with

$$\max_{\substack{A_N y \in K_N \\ y \neq 0}} \min_{j \in B} \frac{\lambda^j_{\min}(A_j y)}{\alpha_j \|y\|}.$$

2. Canonical norm and examples. In this section we recall some basic notions, describe various cones, exhibit explicit descriptions of their corresponding functions $\lambda_{\min}$ and canonical norms, and show how Theorems 1.1 and 1.2 apply in a number of situations.

2.1. Cones and norms. A *pointed cone* in $\mathbb{R}^n$ is a set $K \subseteq \mathbb{R}^n$ satisfying
(i) for every $x \in \mathbb{R}^n$, if $x \in K$, then $\lambda x \in K$ for all $\lambda \geq 0$, and
(ii) $K \cap -K = \{0\}$.
A cone is *regular* if it is pointed and has a nonempty interior. In what follows we assume that all cones are closed, convex, and regular.

We first show that the norm compatibility condition for a given triple (X, K, e) can be alternatively stated in terms of other geometric objects. To do so, recall [19] that the *width* of K is given by $\tau_K = \max\{r \in \mathbb{R} \mid \mathbb{B}(x, r) \subseteq K, \|x\| = 1\}$ and the *center* of K is the point $f \in K$ where τ_K is attained. It follows from the regularity of K that $0 < \tau_K \leq 1$.

PROPOSITION 2.1. *Assume that* $\|e\| = 1$. *The following conditions are equivalent:*
(NC) $\mathbb{B}(e, 1) \subseteq K$.
(NC') $|\lambda_{\min}(x) - \lambda_{\min}(u)| \leq \|x - u\|$ *for all* $x, u \in X$.
(W) $\tau_K = 1$ *and* e *is the center of* K.
(L) $\|s\|^* = \langle e, s\rangle$ *for all* $s \in K^*$.

Proof. To prove the equivalence between (NC) and (NC'), first observe that (NC') can be equivalently phrased as

(NC") $\qquad |\lambda_{\min}(x + d) - \lambda_{\min}(x)| \leq 1 \quad$ for all $d \in \mathbb{B}(e, 1)$ and $x \in X$.

Assume that (NC") holds. Then in particular for all $d \in \mathbb{B}(e, 1)$ we have

$$\lambda_{\min}(e + d) - \lambda_{\min}(e) \geq -1.$$

So

$$\lambda_{\min}(e + d) \geq -1 + \lambda_{\min}(e) = 0,$$

and consequently $e + d \in K$. Since this holds for all $d \in \mathbb{B}(e, 1)$, we get (NC).

Conversely, assume that (NC) holds. Let $d \in \mathbb{B}(e, 1)$ and $x \in X$ be given. By the construction of $\lambda_{\min}$ we have

$$x - \lambda_{\min}(x)e \in K,$$

and by (NC) we have

$$d + e \in K.$$

Hence $x + d - (\lambda_{\min}(x) - 1)e \in K$. Consequently $\lambda_{\min}(x + d) \geq \lambda_{\min}(x) - 1$ by the construction of $\lambda_{\min}$. Thus

$$\lambda_{\min}(x + d) - \lambda_{\min}(x) \geq -1. \tag{2.1}$$

On the other hand, again by the construction of $\lambda_{\min}$ and by (NC), we have

$$x + d - \lambda_{\min}(x + d)e \in K$$

and

$$-d + e \in K.$$

Hence $x - (\lambda_{\min}(x + d) - 1)e \in K$. Consequently $\lambda_{\min}(x) \geq \lambda_{\min}(x + d) - 1$, i.e.,

$$\lambda_{\min}(x + d) - \lambda_{\min}(x) \leq 1. \tag{2.2}$$

We thus get (NC") from (2.1) and (2.2).

Condition (NC) amounts to saying that $\tau_K = 1$ and that e is the center of K. Hence the equivalence of (NC) and (W). Finally, the equivalence between (W) and (L) is shown in [17, Proposition 2.1]. □

Remark 2.2. (i) Any triple (X, K, e) can be endowed with the following *canonical norm* $\| \ \|_{\mathbf{c}}$ so that (X, K, e) satisfies the norm compatibility condition:

$$\|x\|_{\mathbf{c}} := \min\{\alpha \geq 0 : x + \alpha e \in K,\ -x + \alpha e \in K\}.$$

This canonical norm plays a central role in primal-dual interior-point methods for self-scaled cones. In such a context, it is generally denoted as $| \ |_e$. See, e.g., [25, 26, 32].

(ii) In case the norms of some X_j do not satisfy (NC), one may extend Theorem 1.1 to obtain inequalities involving the widths τ_{K_j} of the respective cones.

As some of the examples below illustrate, the canonical norm $\|x\|_{\mathbf{c}}$ above coincides with commonly used norms in a number of cases.

Example 1 (cone of squares in Euclidean Jordan algebras). Consider $(X, K, e) = (\mathcal{E}, \mathcal{K}, e)$, where $\mathcal{E}$ is an Euclidean Jordan algebra, $\mathcal{K}$ is the closure of the cone of squares in $\mathcal{E}$, and $e \in \mathcal{K}$ is the identity element [14]. In this case

$$\lambda_{\min}(x) = \min_{j=1,\ldots,q} \lambda_j(x),$$

and the canonical norm is the spectral norm

$$\|x\|_{\mathbf{c}} = \max_{j=1,\ldots,q} |\lambda_j(x)|,$$

where the $\lambda_j(x)$, $j = 1, \ldots, q$, are the Jordan algebra eigenvalues of x, i.e., the eigenvalues of the characteristic polynomial $\det(\lambda e - x)$ for a suitable homogeneous polynomial det [14, Chap. 3].

Examples 2–7 specialize Example 1 above. They provide explicit expressions for $\lambda_{\min}(\ )$ and $\| \ \|_{\mathbf{c}}$ for some specific Jordan algebras. It should be noted that the explicit expressions in Examples 2–4 have been known in optimization for some time (see [16, sect. 2] and [25, sect. 3]).

Example 2 (nonnegative orthant). Consider $(X, K, e) = (\mathbb{R}^n, \mathbb{R}^n_+, (1, \ldots, 1))$. In this case $q = n,\ \lambda_j(x) = x_j,\ j = 1, \ldots, n$. Consequently,

$$\lambda_{\min}(x) = \min_j x_j, \quad \|x\|_{\mathbf{c}} = \|x\|_\infty = \max_j |x_j|.$$

Example 3 (second-order cone). Consider $(X, K, e) = (\mathbb{R}^{n+1}, \mathcal{Q}^n, (1, 0, \ldots, 0))$, where $\mathcal{Q}^n$ is the *second-order cone* defined to be

$$\mathcal{Q}^n := \{x = (x_0, \overline{x}), \overline{x} \in \mathbb{R}^n : x_0 \geq \|\overline{x}\|_2\}.$$

In this case $q = 2,\ \lambda_1(x) = x_0 - \|\overline{x}\|_2,\ \lambda_2(x) = x_0 + \|\overline{x}\|_2$. Consequently,

$$\lambda_{\min}(x) = x_0 - \|\overline{x}\|_2, \quad \|x\|_{\mathbf{c}} = |x_0| + \|\overline{x}\|_2.$$

Example 4 (semidefinite cone). Consider $(X, K, e) = (\mathbf{S}^n, \mathbf{S}^n_+, \mathrm{I})$, where $\mathbf{S}^n$ is the set of $n \times n$ symmetric matrices, $\mathbf{S}^n_+$ is the subset of those which are positive semidefinite, and I is the identity matrix. In this case

$$\lambda_{\min}(x) = \min_{j=1,\ldots,n} \lambda_j(x) \qquad \text{and} \qquad \|x\|_{\mathbf{c}} = \max_{j=1,\ldots,n} |\lambda_j(x)|,$$

where $\lambda_j(x),\ j = 1, \ldots, n$, are the usual eigenvalues of x, i.e., the roots of $p(\lambda) := \det(\lambda \mathrm{I} - x)$.

Example 5 (cones of positive semidefinite Hermitian matrices). Consider (X, K, e), where X is the real vector space $\mathrm{Herm}(n, \mathbb{C})$ of $n \times n$ Hermitian matrices with complex entries, K is the cone of positive semidefinite Hermitian matrices in X, and e is the $n \times n$ identity matrix. In this case $q = n,\ \lambda_j(x),\ j = 1, \ldots, n$, are the usual eigenvalues of x, i.e., the roots of $p(\lambda) := \det(\lambda \mathrm{I} - x)$ which, it is well known, are real.

Example 6 (cones of positive semidefinite Hermitian matrices with quaternions entries). Consider (X, K, e), where X is the real vector space $\mathrm{Herm}(n, \mathbb{H})$ of $n \times n$ Hermitian matrices with quaternion entries, K is the cone of positive semidefinite Hermitian matrices in X, and e is the $n \times n$ identity matrix. In this case $q = n$ and $\lambda_j(x),\ j = 1, \ldots, n$, are the roots (as a univariate polynomial in λ) of the "characteristic polynomial" $\det(\lambda e - x)$ of X. This polynomial is defined as follows [15]. Let J be the $2n \times 2n$ matrix $\left[\begin{smallmatrix} 0 & \mathrm{I}_n \\ -\mathrm{I}_n & 0 \end{smallmatrix}\right]$. Then $\mathrm{Herm}(n, \mathbb{H})$ can be written as $\{z \in \mathrm{Herm}(2n, \mathbb{C}) : \overline{z} J = J z\}$, and, for $z \in \mathrm{Herm}(n, \mathbb{H})$,

$$\det(z) = \mathrm{Pf}(Jz),$$

where $\mathrm{Pf}(Jz)$ is the *Pfaffian* of Jz, i.e., the unique polynomial satisfying $\mathrm{Pf}(J) = 1$ and $\mathrm{Pf}(Jz)^2 = \det(Jz)$. Again, it is well known that the $\lambda_j(x),\ j = 1, \ldots, n$, are real. (For a more detailed discussion on Pfaffians, see, e.g., [21].)

Example 7 (cones of squares in the Albert algebra). Consider (X, K, e), where X is the real vector space of 3×3 Hermitian matrices with octonion entries [2, 9], K is the cone of squares in X, i.e., $K = \{x^2 : x \in X\}$, and e is the 3×3 identity matrix. In this case $q = 3$ and $\lambda_j(x),\ j = 1, 2, 3$, are the roots of the characteristic polynomial

$$p(\lambda) = \det(\lambda e - x) = \lambda^3 - \mathrm{trace}(x)\lambda^2 + \sigma(x)\lambda - \det(x),$$

where $\mathrm{trace}(x), \sigma(x), \det(x)$ are defined as follows [12, 15]. For a, b, c octonions and $p, m, n \in \mathbb{R}$,

$$x = \begin{bmatrix} p & a & \overline{b} \\ \overline{a} & m & c \\ b & \overline{c} & n \end{bmatrix},$$

$$\begin{aligned}\operatorname{trace}(x) &= p + m + n,\\ \sigma(x) &= pm + mn + pn - |a|^2 - |b|^2 - |c|^2,\\ \det(x) &= pmn + b(ac) + \overline{b(ac)} - n|a|^2 - m|b|^2 - p|c|^2.\end{aligned}$$

Note that we write $b(ac)$ to emphasize the order of the multiplications in the nonassociative ring of octonions. Just as in the previous examples, λ_1, λ_2, and λ_3 are real.

Example 8 (hyperbolicity cones). Let $e \in \mathbb{R}^n$ and $p \in \mathbb{R}[X_1, \ldots, X_n]$ be a *complete hyperbolic polynomial* in the direction e, i.e., a homogeneous polynomial satisfying that, for all $x \in \mathbb{R}^n$, the univariate polynomial $\lambda \mapsto p(\lambda e - x)$ has only real roots, and at least one of them is nonzero for $x \neq 0$. (For a detailed exposition on hyperbolic polynomials, see, e.g., [3, 33].) Consider $(X, K, e) = (\mathbb{R}^n, C(p, e), e)$, where $C(p, e)$ is the closure of the *hyperbolicity cone* for p in the direction e; i.e., $C(p, e)$ is the closure of the connected component of the set $\{x : p(x) > 0\}$ that contains e. In this case

$$\lambda_{\min}(x) = \min_{j=1,\ldots,d} \lambda_j(x),$$

where $\lambda_j(x)$, $j = 1, \ldots, d = \deg(p)$, are the roots of the polynomial $\lambda \mapsto p(\lambda e - x)$, and

$$\|x\|_{\mathbf{c}} = \max_{j=1,\ldots,d} |\lambda_j(x)|.$$

Example 9 (nonnegative, finitely spanned, functions on a compact domain). Assume that $d \in \mathbb{N}$, $D \subseteq \mathbb{R}^n$ is a nonempty compact set, and $f_0, \ldots, f_d$ are continuous, real-valued functions defined on D, with $f_0(x) = 1$, for all $x \in D$. Consider the triple (X, K, e), where

$$X = \operatorname{span}\{f_0, \ldots, f_d\},$$

$$K = \{f \in X : f(x) \geq 0 \text{ for all } x \in D\},$$

and $e \in X$ is the constant function f_0. In this case, for $f \in X$,

$$\lambda_{\min}(f) = \min_{x \in D} f(x)$$

and

$$\|f\|_{\mathbf{c}} = \max_{x \in D} |f(x)|.$$

2.2. Cone reducibility. Assume that X is a finite-dimensional inner product space. Then the map $u \mapsto \langle u, \cdot \rangle$ defines an isomorphism between X and its dual space X^*. The *dual* of a cone $K \subseteq X$ is identified via this isomorphism with the cone

$$K^* = \{u \in X : \langle u, x \rangle \geq 0 \text{ for all } x \in K\}.$$

We say that K is *self-dual* if $K^* = K$. We say that K is *homogeneous* if for all $x, u \in \mathsf{int}(K)$ there exists $g \in \operatorname{Aut}(\mathsf{int}(K))$ such that $gx = u$, where $\operatorname{Aut}(\mathsf{int}(K)) = \{g \in \operatorname{GL}(X) : g(\mathsf{int}(K)) = \mathsf{int}(K)\}$. Here $\operatorname{GL}(X)$ is the general linear group over X. A cone is *symmetric* if it is self-dual and homogeneous [14].

Symmetric cones coincide with *self-scaled cones,* a class of cones that plays a central role in interior-point methods [25]. Nesterov and Todd identified the properties

of self-scaled cones as the fundamental building blocks for the development of symmetric primal-dual interior-point algorithms [26, 32] for conic programs over these cones. Symmetric cones have been extensively studied in other areas of mathematics. It can also be shown that they coincide with the cones of squares of Euclidean Jordan algebras (cf. Example 1). Furthermore, they satisfy a unique factorization property; namely, they can be written in a unique way (up to ordering) as a product of cones in the classes described in Examples 3–7 above [14].

Because second-order conic feasibility over a single second-order cone can be solved in closed form, all interesting examples of second-order conic feasiblity problems are written in terms of a nontrivial product of second-order cones (see, e.g., [1, 24]).

Given X and K, one may wonder how the canonical norm and the minimum-eigenvalue constructs depend on different factorizations of (X, K). The following proposition settles this question.

PROPOSITION 2.3. *Let $X = X_1 \times \cdots \times X_r$, $K = K_1 \times \cdots \times K_r$, and $e = (e_1, \ldots, e_r)$, with $e_j \in \mathsf{int}(K_j)$. Then, for all $x = (x_1, \ldots, x_r) \in X$,*

(i)

$$\lambda_{\min}(x) = \min\{\lambda^1_{\min}(x_1), \ldots, \lambda^r_{\min}(x_r)\}.$$

where $\lambda^j_{\min}$ is the minimum eigenvalue associated to (X_j, K_j, e_j) and $\lambda_{\min}$ that associated to (X, K, e).

(ii)

$$\|x\|_{\mathbf{c}} = \max_{j=1,\ldots,r} \|x_j\|_{\mathbf{c},j},$$

where $\| \ \|_{\mathbf{c},j}$ is the canonical norm associated to (X_j, K_j, e_j) and $\| \ \|_{\mathbf{c}}$ that associated to (X, K, e). In particular, the restriction of $\| \ \|_{\mathbf{c}}$ to X_j is $\| \ \|_{\mathbf{c},j}$.

Proof. From (1.7) it follows that, for $x = (x_1, \ldots, x_r) \in X$,

$$\begin{aligned}
\lambda_{\min}(x) &= \max\{t \in \mathbb{R} : x - te \in K\} \\
&= \max\{t \in \mathbb{R} : x_j - te_j \in K_j \ \text{ for } \ j = 1, \ldots, r\} \\
&= \min_{j=1,\ldots,r} \max\{t \in \mathbb{R} : x_j - te_j \in K_j\} \\
&= \min\{\lambda^1_{\min}(x_1), \ldots, \lambda^r_{\min}(x_r)\}.
\end{aligned}$$

This shows part (i). For part (ii) we first claim that

$$B = B_1 \times \cdots \times B_r. \tag{2.3}$$

Indeed, given $d = (d_1, \ldots, d_r) \in X$,

$$\begin{aligned}
d \in B &\Leftrightarrow e + d, e - d \in K \\
&\Leftrightarrow e_j + d_j, e_j - d_j \in K_j \ \text{ for } \ j = 1, \ldots, r \\
&\Leftrightarrow d_j \in B_j \ \text{ for } \ j = 1, \ldots, r \\
&\Leftrightarrow d \in B_1 \times \cdots \times B_r.
\end{aligned}$$

From (2.3) it follows that, for $x = (x_1, \dots, x_r) \in X$,

$$\begin{aligned}\|x\|_{\mathbf{c}} &= \inf\left\{t : \frac{1}{t}x \in B\right\}\\ &= \inf\left\{t : \frac{1}{t}x_j \in B_j \ \text{ for } \ j = 1, \dots, r\right\}\\ &= \max_{j=1,\dots,r} \inf\left\{t : \frac{1}{t}x_j \in B_j\right\}\\ &= \max_{j=1,\dots,r} \|x_j\|_{\mathbf{c},j}. \qquad \square\end{aligned}$$

We have already mentioned that we endow $L(Y, X)$ with the operator norm with respect to the norms in Y and X. Therefore, the canonical norm in X induces a *canonical norm in* $L(Y, X)$. In particular, in the case where $X = X_1 \times \cdots \times X_r$, Proposition 2.3(ii) yields the relation

$$\|A\|_{\mathbf{c}} = \max_{j=1,\dots,r} \|A_j\|_{\mathbf{c},j}$$

for the canonical norms in $L(Y, X)$ and those in $L(Y, X_j)$, $j = 1, \dots, r$.

Remark 2.4. Note that the factorization mentioned above together with Proposition 2.3(ii) and Examples 3–7 yield expressions for the canonical norm for every symmetric cone. If the factorization is explicit, then the expressions for the canonical norm are explicit as well.

2.3. Condition numbers and the choice of $\boldsymbol{\alpha}$. We mentioned in section 1.1 the role of Renegar's condition number in the analysis of algorithms for conic feasibility problems. We also mentioned there that, in the case of polyhedral cones, the condition number $\mathscr{C}(A)$ exploited the reducibility of the cone $\mathbb{R}^n_+$. We next show how these condition numbers are obtained by appropriately selecting α.

Assume a factorization $X = X_1 \times \cdots \times X_r$ and $K = K_1 \times \cdots \times K_r$. Basic choices for α are

(1) $\alpha_j = \|A\|$ for $j = 1, \dots, r$;
(2) $\alpha_j = \|A_j\|$ for $j = 1, \dots, r$.

The first choice leads to Renegar's condition number $C(A)$ for the norm in $L(Y, X)$ defined by

$$\|A\| = \max_{j=1,\dots,r} \|A_j\| \tag{2.4}$$

because in this case

$$C_\alpha(A) = \frac{\|A\|}{\min\limits_{\widetilde{A}\in\Sigma} \max\limits_{j=1,\dots,r} \|A_j - \widetilde{A}_j\|} = \frac{\|A\|}{\min\limits_{\widetilde{A}\in\Sigma} \|A - \widetilde{A}\|} = C(A).$$

Theorem 1.1 then takes the form of a minmax characterization of the distance to ill-posedness (and therefore of $C(A)$). We note that this can also be obtained from [18, Thms. 7 and 10].

The second choice of α above leads to (extensions of) the condition number $\mathscr{C}(A)$ introduced in [7].

The discusion above assumes that $\alpha_j > 0$ for all $j = 1, \dots, r$. If some cones are rigid, say, $K = K_1 \times \cdots \times K_r \times K_N$ with $\alpha_j > 0$ for $j = 1, \dots, r$ and $\alpha_N = 0$, then,

by letting $B = \{1, \dots, r\}$, one defines

$$C(A) = \frac{\|A_B\|}{\min\limits_{\substack{\widetilde{A}\in\Sigma \\ \widetilde{A}_N = A_N}} \max\limits_{j\in B} \|A_j - \widetilde{A}_j\|}$$

and

$$\mathscr{C}(A) = \frac{1}{\left| \max\limits_{\substack{A_N y \in K_N \\ y\neq 0}} \min\limits_{j\in B} \dfrac{\lambda^j_{\min}(A_j y)}{\|A_j\|\,\|y\|} \right|}.$$

The proof of the following proposition is an immediate consequence of the fact that $\|A_j\| \leq \|A\|$ for all $j = 1, \dots, r$.

PROPOSITION 2.5. *For all $A \in L(Y, X)$, $\mathscr{C}(A) \leq C(A)$. This holds as well if some factors of A are rigid.*

We next see how the choices of α above materialize in the case of polyhedral conic systems.

Example 2 (revisited). Recall that $X = \mathbb{R}^n$, $Y = \mathbb{R}^m$, and $K = \mathbb{R}^n_+$.

(i) Consider the case $r = 1$. We do not decompose K. In this case we take $e = (1, \dots, 1)$ and, as we have seen, $\|x\| = \|x\|_\infty$. This induces the canonical norm in $L(Y, X)$ given by

$$\|A\| = \|A\|_{Y\infty} = \max_{\|y\|=1} \|Ay\|_\infty = \max_{\|y\|=1} \max_{j=1,\dots,n} |A_j y|.$$

We now take $\alpha = \|A\|$. Theorem 1.1 and the fact that $\lambda_{\min}(x) = \min_{j=1,\dots,n} x_j$ give then the following characterization of Renegar's condition number:

$$C(A) = \frac{\|A\|}{\mathsf{dist}(A, \Sigma)} = \frac{\|A\|}{\left| \max\limits_{\|y\|=1} \min\limits_{j=1,\dots,n} Ay \right|}. \tag{2.5}$$

(ii) Consider now the case where $r = n$. Here we take $K_j = [0, +\infty)$ and $e_j = 1$ for $j = 1, \dots, r$. We obtain the canonical norm $\|x\| = |x|$ and the minimum-eigenvalue $\lambda_{\min}(x) = x$. The former induces the canonical norm in $L(Y, X)$ given by

$$\|A\| = \max_{j=1,\dots,n} \|A_j\| = \max_{j=1,\dots,n} \max_{\|y\|=1} |A_j y|,$$

that is, as in case (i) above. Again, take $\alpha_j = \|A\|$ for all j. Then, not surprisingly, Theorem 1.1 characterizes $C(A)$ by (2.5) as well.

(iii) We now take r, K_j, and e_j as in (ii) but choose instead $\alpha_j = \|A_j\| = \max_{\|y\|=1} |A_j y|$. In this case Theorem 1.1 gives us the well-known [7] characterization of $\mathscr{C}(A)$:

$$\mathscr{C}(A) := \frac{1}{|\overline{v}_{A,\alpha}|} = \frac{1}{\min\limits_{\widetilde{A}\in\Sigma} \max\limits_{j=1,\dots,n} \frac{\|A_j - \widetilde{A}_j\|}{\|A_j\|}}.$$

(iv) For $M \in \mathbb{R}^{n\times m}$ consider the system

$$\begin{aligned} My &\geq 0, \\ y &\geq 0. \end{aligned}$$

This system can be thought of as a special case of the above with $A = (M, \mathrm{I})$. The identity matrix I, however, should be considered to be rigid (not subject to perturbations) and its corresponding α_{I} then be set to 0.

By taking $r = 2$ (two blocks, corresponding to M and I) and $\alpha_M = \|M\| = \max_{j=1,\dots,n} \max_{\|y\|=1} |M_j y|$, we obtain (in the right-hand side of (1.10)) Renegar's condition number, and Theorem 1.2 shows that

$$C(M) := \frac{\|M\|}{\min\limits_{\widetilde{M}\in\Sigma} \|M - \widetilde{M}\|} = \frac{\|M\|}{\min\limits_{\widetilde{M}\in\Sigma} \max\limits_{j=1,\dots,n} \|M_j - \widetilde{M}_j\|} = \frac{\|M\|}{\left|\max\limits_{y\geq 0} \min\limits_{j=1,\dots,n} \frac{M_j y}{\|y\|}\right|}.$$

Finally, by taking $r = n$ and $\alpha_j = \|M_j\| = \max_{\|y\|=1} |M_j y|$, $j = 1, \dots, n$, we obtain $\mathscr{C}(M)$ in the left-hand side of (1.10), and now Theorem 1.2 shows that

$$\mathscr{C}(M) := \frac{1}{\left|\max\limits_{y\geq 0} \min\limits_{j=1,\dots,n} \frac{M_j y}{\|M_j\|\|y\|}\right|} = \frac{1}{\min\limits_{\widetilde{M}\in\Sigma} \max\limits_{j=1,\dots,n} \frac{\|M_j - \widetilde{M}_j\|}{\|M_j\|}}.$$

We have revisited Example 2 to see how Theorems 1.1 and 1.2, together with appropriate choices of α, yield characterizations of $C(A)$ and $\mathscr{C}(A)$ in the case of polyhedral conic systems, possibly with rigid components. The other examples in section 2.1, and arbitrary products of them, may be similarly dealt with. We will not do so to avoid being repetitious.

2.4. Well-conditioned solutions. As was noted in section 1.3, for $A \in \mathcal{D}$ any point $\overline{y} \in Y$ that satisfies

$$v_{A,\alpha}(\overline{y}) = \overline{v}_{A,\alpha} \tag{2.6}$$

can be interpreted as a best conditioned solution for (1.1). Indeed, from (2.6) it follows that $A\overline{y} \in \mathsf{int}(K)$ and for each $i = 1, \dots, r$

$$\frac{\mathsf{dist}(A_i\overline{y}, \partial K_i)}{\|\overline{y}\|} \geq \overline{v}_{A,\alpha}\, \alpha_i.$$

The following proposition provides an analogous statement for $A \in \mathcal{P}$.

PROPOSITION 2.6. *Assume that each one of the triples (X_j, K_j, e_j), $j = 1, \dots, r$, satisfies the norm compatibility condition* (NC) *and $A \in \mathcal{P}$. Then there exists $\overline{x} \in \mathsf{int}(K^*)$ such that $A^*\overline{x} = 0$ and for each $i = 1, \dots, r$*

$$\frac{\mathsf{dist}(\overline{x}_i, \partial K_i^*)}{\|\overline{x}_i\|} \geq \frac{|\overline{v}_{A,\alpha}|\, \alpha_i\, \tau_{K_i^*}}{r\|A_i\| + |\overline{v}_{A,\alpha}|\alpha_i}. \tag{2.7}$$

In particular, if $\alpha_i = \|A_i\|$, $i = 1, \dots, r$, then there exists $\overline{x} \in K^*$ such that $A^*\overline{x} = 0$ and for each $i = 1, \dots, r$

$$\frac{\mathsf{dist}(\overline{x}_i, \partial K_i^*)}{\|\overline{x}_i\|} \geq \frac{|\overline{v}_{A,\alpha}|\tau_{K_i^*}}{r + |\overline{v}_{A,\alpha}|} \geq \frac{|\overline{v}_{A,\alpha}|\tau_{K_i^*}}{r+1}.$$

Proof. For each $i = 1, \dots, r$, let $f_i^* \in K_i^*$ be the center of K_i^*, i.e., $\|f_i^*\|^* = 1$ and $\mathsf{dist}(f_i^*, \partial K_i^*) = \tau_{K_i^*}$. Define $\widetilde{x} \in K^*$ and $y^* \in Y^*$ as follows:

$$\widetilde{x}_i := \frac{|\overline{v}_{A,\alpha}|}{r\|A_i\|} f_i^*,\ y^* := -A^*\widetilde{x}.$$

It is immediate that $\|y^*\|^* \leq |\overline{v}_{A,\alpha}|$, so by (1.11) there exists $x^* \in K^*$ such that $A^*x^* = y^* = -A^*\widetilde{x}$ and $\sum_{i=1}^r \alpha_i \|x_i^*\|^* \leq 1$. Thus the point $\overline{x} := x^* + \widetilde{x} \in K^*$ satisfies $A^*\overline{x} = 0$. To finish, we next show that $\overline{x}$ satisfies (2.7). Since $\sum_{i=1}^r \alpha_i \|x_i^*\|^* \leq 1$, it follows that $\|x_i^*\|^* \leq 1/\alpha_i$ for each $i = 1, \ldots, r$. Hence

$$\|\overline{x}_i\| \leq \|x_i^*\| + \|\widetilde{x}_i\| \leq \frac{1}{\alpha_i} + \frac{|\overline{v}_{A,\alpha}|}{r\|A_i\|} = \frac{r\|A_i\| + |\overline{v}_{A,\alpha}|\alpha_i}{r\|A_i\|\alpha_i}. \tag{2.8}$$

On the other hand, since $\mathsf{dist}(f_i^*, \partial K_i^*) = \tau_{K_i^*}$, it follows that

$$\mathsf{dist}(\overline{x}_i, \partial K_i^*) \geq \frac{|\overline{v}_{A,\alpha}|\tau_{K_i^*}}{r\|A_i\|}. \tag{2.9}$$

Inequality (2.7) then follows from (2.8) and (2.9). ∎

3. Proof of the main results. The result is trivial when $\overline{v}_{A,\alpha} = 0$. Therefore, we will assume that $\overline{v}_{A,\alpha} \neq 0$. For ease of exposition, we split the proof of Theorem 1.1 into two parts, namely, Propositions 3.1 and 3.2.

PROPOSITION 3.1.

$$|\overline{v}_{A,\alpha}| \leq \min_{\widetilde{A} \in \Sigma} \max_{j=1,\ldots,r} \frac{\|A_j - \widetilde{A}_j\|}{\alpha_j}.$$

Proof. Assume that $\widetilde{A}$ is such that

$$\max_{j=1,\ldots,r} \frac{\|A_j - \widetilde{A}_j\|}{\alpha_j} < |\overline{v}_{A,\alpha}|.$$

We need to prove that $\widetilde{A} \notin \Sigma$, i.e., $\widetilde{A} \in \mathcal{P} \cup \mathcal{D}$.

Let $\overline{y}_A \in Y$ be such that $v_{A,\alpha}(\overline{y}_A) = \overline{v}_{A,\alpha}$. Assume without loss of generality that $\|\overline{y}_A\| = 1$. Because each (X_j, K_j, e_j) satisfies the norm compatibility condition, it follows from Proposition 2.1 that, for all $\widetilde{A}$ and $y \in Y \setminus \{0\}$,

$$\begin{aligned} \frac{|\lambda^j_{\min}(A_j y) - \lambda^j_{\min}(\widetilde{A}_j y)|}{\alpha_j} &\leq \frac{\|A_j - \widetilde{A}_j\|}{\alpha_j} \|y\| \\ &\leq \max_{j=1,\ldots,r} \frac{\|A_j - \widetilde{A}_j\|}{\alpha_j} \|y\| \\ &< |\overline{v}_{A,\alpha}| \|y\|. \end{aligned} \tag{3.1}$$

In particular

$$\frac{|\lambda^j_{\min}(A_j \overline{y}_A) - \lambda^j_{\min}(\widetilde{A}_j \overline{y}_A)|}{\alpha_j} < |\overline{v}_{A,\alpha}|. \tag{3.2}$$

We now consider the cases $\overline{v}_{A,\alpha} < 0$ and $\overline{v}_{A,\alpha} > 0$ separately.

Case 1: $\overline{v}_{A,\alpha} > 0$. In this case $A \in \mathcal{D}$. From (1.8), (3.2), and the equality $v_{A,\alpha}(\overline{y}_A) = \overline{v}_{A,\alpha}$, we get, for $j = 1, \ldots, r$,

$$\begin{aligned} \frac{\lambda^j_{\min}(\widetilde{A}_j \overline{y}_A)}{\alpha_j} &\geq \frac{\lambda^j_{\min}(A_j \overline{y}_A)}{\alpha_j} - \frac{|\lambda^j_{\min}(A_j \overline{y}_A) - \lambda^j_{\min}(\widetilde{A}_j \overline{y}_A)|}{\alpha_j} \\ &> \overline{v}_{A,\alpha} - \overline{v}_{A,\alpha} = 0. \end{aligned}$$

Therefore

$$v_{\widetilde{A}}(\overline{y}_A) = \min_{j=1,\dots,r} \frac{\lambda^j_{\min}(\widetilde{A}_j \overline{y}_A)}{\alpha_j} > 0,$$

which shows that $\overline{y}_A$ is a strict solution for $\widetilde{A}$ and, consequently, that $\widetilde{A} \in \mathcal{D}$.

Case 2: $\overline{v}_{A,\alpha} < 0$. In this case $A \in \mathcal{P}$. Let y be any point in $Y \setminus \{0\}$. Since $\overline{v}_{A,\alpha} < 0$, we must have $v_{A,\alpha}(y) \le \overline{v}_{A,\alpha} < 0$. Let $\overline{j} = \overline{j}(y)$ be such that $\frac{\lambda^{\overline{j}}_{\min}(A_{\overline{j}} y)}{\alpha_{\overline{j}} \|y\|} = v_{A,\alpha}(y)$. We claim that $\lambda^{\overline{j}}_{\min}(\widetilde{A}_{\overline{j}} y) < 0$. Indeed, by (3.1),

$$\begin{aligned}
&\left|\lambda^{\overline{j}}_{\min}(A_{\overline{j}} y) - \lambda^{\overline{j}}_{\min}(\widetilde{A}_{\overline{j}} y)\right| < -\overline{v}_{A,\alpha} \alpha_{\overline{j}} \|y\| \le -v_{A,\alpha}(y) \alpha_{\overline{j}} \|y\| \\
&\Rightarrow \lambda^{\overline{j}}_{\min}(\widetilde{A}_{\overline{j}} y) - \lambda^{\overline{j}}_{\min}(A_{\overline{j}} y) < -v_{A,\alpha}(y) \alpha_{\overline{j}} \|y\| \\
&\Rightarrow \lambda^{\overline{j}}_{\min}(\widetilde{A}_{\overline{j}} y) - v_{A,\alpha}(y) \alpha_{\overline{j}} \|y\| < -v_{A,\alpha}(y) \alpha_{\overline{j}} \|y\| \\
&\Rightarrow \lambda^{\overline{j}}_{\min}(\widetilde{A}_{\overline{j}} y) < 0.
\end{aligned}$$

Hence, for all $y \in Y \setminus \{0\}$ there exists j such that $\lambda^j_{\min}(\widetilde{A}_j y) < 0$. It follows that

$$\overline{v}_{\widetilde{A},\alpha} = \max_{y \ne 0} v_{\widetilde{A},\alpha}(y) = \max_{y \ne 0} \min_{j=1,\dots,r} \frac{\lambda^j_{\min}(\widetilde{A}_j y)}{\alpha_j \|y\|} < 0,$$

that is, $\widetilde{A} \in \mathcal{P}$. ∎

Recall that, given vector spaces X and Y and a linear mapping $A \in L(Y, X)$, its *adjoint* $A^* \in L(X^*, Y^*)$ is the unique linear mapping that satisfies

$$\langle v, Ay \rangle = \langle A^* v, y \rangle \quad \text{for all } v \in X^*,\ y \in Y.$$

Proposition 3.2.

$$|\overline{v}_{A,\alpha}| \ge \min_{\widetilde{A} \in \Sigma} \max_{j=1,\dots,r} \frac{\|A_j - \widetilde{A}_j\|}{\alpha_j}.$$

Proof. We consider the cases $\overline{v}_{A,\alpha} < 0$ and $\overline{v}_{A,\alpha} > 0$ separately.

Case 1: $\overline{v}_{A,\alpha} < 0$. In this case $A \notin \mathcal{D}$, so it suffices to show that for all $\delta > 0$ there exists $\widetilde{A} \in \mathcal{D}$ such that, for all $j = 1, \dots, r$, $\|A_j - \widetilde{A}_j\| \le \alpha_j(|\overline{v}_{A,\alpha}| + \delta)$. Let $\overline{y}_A \in Y$ be such that $v_A(\overline{y}_A) = \overline{v}_{A,\alpha}$. Assume without loss of generality that $\|\overline{y}_A\| = 1$. By the Hahn–Banach theorem [34, Thm. 5.20], there exists $v \in Y^*$ such that $\langle v, \overline{y}_A \rangle = \|\overline{y}_A\| = 1$ and $\|v\|^* = 1$. For $j = 1, \dots, r$, consider $\widetilde{A}_j \in L(Y, X_j)$ given by

$$\widetilde{A}_j = A_j - \alpha_j(\overline{v}_{A,\alpha} - \delta)\langle v, \cdot \rangle e_j.$$

We claim that $\widetilde{A} \in \mathcal{D}$. To see this, first notice that, for all $j = 1, \dots, r$, $A_j \overline{y}_A - \overline{v}_{A,\alpha} e_j \in K_j$ because

$$\overline{v}_{A,\alpha} = v_{A,\alpha}(\overline{y}_A) \le \frac{\lambda^j_{\min}(A_j \overline{y}_A)}{\alpha_j} = \max\{t \mid A_j \overline{y}_A - \alpha_j t e_j \in K_j\}.$$

Therefore,

$$\widetilde{A}_j \overline{y}_A = A_j \overline{y}_A - \alpha_j(\overline{v}_{A,\alpha} - \delta)\langle v, \overline{y}_A \rangle e_j = (A_j \overline{y}_A - \alpha_j \overline{v}_{A,\alpha} e_j) + \alpha_j \delta e_j \in \mathsf{int}(K_j)$$

since K_j is convex and $e_j \in \mathsf{int}(K_j)$. This shows that $\overline{y}_A$ is a strict solution for $\widetilde{A}$. To finish, just observe that $\|A_j - \widetilde{A}_j\| = \alpha_j \|(\overline{v}_{A,\alpha} - \delta)\langle v, \cdot\rangle e_j\| \leq \alpha_j |\overline{v}_{A,\alpha} - \delta| \|v\|^* \|e_j\| = \alpha_j(|\overline{v}_{A,\alpha}| + \delta)$.

Case 2: $\overline{v}_{A,\alpha} > 0$. In this case $A \in \mathcal{D}$, so it suffices to show that there exists $\widetilde{A} \notin \mathcal{D}$ such that, for all $j = 1, \ldots, r$, $\|A_j - \widetilde{A}_j\| \leq \alpha_j \overline{v}_{A,\alpha}$. Let $e = (e_1, \ldots, e_r) \in K = K_1 \times \cdots \times K_r$. Let $B_j = \frac{1}{\alpha_j} A_j$ and $B = [B_1, \ldots, B_r] \in L(Y, X)$. From Proposition 2.3(i) and the positive homogeneity of $\lambda_{\min}$ it follows that for $y \in Y \setminus \{0\}$

$$\begin{aligned} v_{A,\alpha}(y) &= \min_{j=1,\ldots,r} \frac{\lambda^j_{\min}(A_j y)}{\alpha_j \|y\|} = \frac{1}{\|y\|} \min_{j=1,\ldots,r} \lambda^j_{\min}(B_j y) \\ &= \frac{1}{\|y\|} \lambda_{\min}(By) = \frac{1}{\|y\|} \max\{t : By - te \in K\}. \end{aligned}$$

Then, by taking maxima on both sides above,

$$\overline{v}_{A,\alpha} = \max_{y \neq 0} v_{A,\alpha}(y) = \max_{\|y\|=1} \max\{t : By - te \in K\}.$$

Since $\overline{v}_{A,\alpha} > 0$ we may rewrite the above as a maximum over a convex set

$$\overline{v}_{A,\alpha} = \max_{\substack{By - te \in K \\ \|y\| \leq 1}} t. \tag{3.3}$$

Consider the Lagrangian dual (see [4]) of the right-hand side of (3.3):

$$\begin{aligned} \min_{x \in K^*} \max_{\substack{\|y\| \leq 1 \\ t \in \mathbb{R}}} t + \langle x, By - te\rangle &= \min_{x \in K^*} \max_{\substack{\|y\| \leq 1 \\ t \in \mathbb{R}}} t(1 - \langle x, e\rangle) + \langle x, By\rangle \\ &= \min_{\substack{x \in K^* \\ \langle x, e\rangle = 1}} \max_{\|y\| \leq 1} \langle B^* x, y\rangle \\ &= \min_{\substack{x \in K^* \\ \langle x, e\rangle = 1}} \|B^* x\|^*. \end{aligned} \tag{3.4}$$

Since both (3.3) and (3.4) are convex programs and satisfy the Slater condition, by [4, Thm. 4.3.7], they attain the same optimal value $\overline{v}_{A,\alpha}$. Hence there exists $\overline{x} \in K^*$ such that $\|B^*\overline{x}\|^* = \overline{v}_{A,\alpha}$ and $\langle \overline{x}, e\rangle = 1$. Let $\widetilde{A}_j = A_j - \alpha_j \langle B_j^* \overline{x}_j, \cdot\rangle e_j = A_j - \langle A_j^* \overline{x}_j, \cdot\rangle e_j$. We claim that $\widetilde{A} \notin \mathcal{D}$. Indeed, otherwise, there would exist $y \in Y$ and $\epsilon > 0$ such that $\widetilde{A}y - \epsilon e \in K$ and, therefore,

$$\begin{aligned} 0 &\leq \langle \overline{x}, \widetilde{A}y - \epsilon e\rangle \qquad (\text{because } \overline{x} \in K^*) \\ &= \langle \overline{x}, Ay - (\langle A^*\overline{x}, y\rangle + \epsilon)e\rangle \\ &= -\epsilon \qquad (\text{because } \langle \overline{x}, e\rangle = 1) \\ &< 0, \end{aligned}$$

which is a contradiction. Hence $\widetilde{A} \notin \mathcal{D}$. To finish, observe that

$$\frac{\|\widetilde{A}_j - A_j\|}{\alpha_j} = \|\langle B_j^* \overline{x}_j, \cdot\rangle e_j\| = \|B_j^* \overline{x}_j\|^* \leq \|B^* \overline{x}\|^* = \overline{v}_{A,\alpha}. \qquad \square$$

We next prove Theorem 1.2. We will need the following result.

LEMMA 3.3. *Assume that $A \notin \Sigma$. If the system*

$$A_N y \in \partial K_N, \quad A_B y \in \mathsf{int}(K_B)$$

has a nontrivial solution, then so does the system

$$A_N y \in \mathsf{int}(K_N), \quad A_B y \in \mathsf{int}(K_B).$$

Proof. Let $\mathbf{1} = (1, 1, \ldots, 1) \in \mathbb{R}^d$. By hypothesis, $v_{A,\mathbf{1}}(y) = 0$ and so $\overline{v}_{A,\mathbf{1}} \geq 0$. Since $A \notin \Sigma$, $\overline{v}_{A,\mathbf{1}} > 0$, and so there is $y' \neq 0$ such that $v_{A,\mathbf{1}}(y') > 0$. But this implies that $A_N y' \in \mathsf{int}(K_N)$ and $A_B y' \in \mathsf{int}(K_B)$). ∎

Proof of Theorem 1.2. We first show that

$$\{y \neq 0 \mid A_N y \in K_N\} = \emptyset \iff \{\widetilde{A} \in \Sigma \mid \widetilde{A}_N = A_N\} = \emptyset.$$

This will show that the left-hand side in (1.10) is $+\infty$ if and only if so is the right-hand side.

For the only if direction, assume that there exists $\widetilde{A}_B$ such that $\mathscr{A} = (\widetilde{A}_B, A_N) \in \Sigma$. The latter implies that $\overline{v}_{\mathscr{A},\mathbf{1}} = 0$. Hence, there exists $\overline{y} \in S_Y := \{y \in Y \mid \|y\| = 1\}$ such that $\min_{j=1,\ldots,r} \lambda^j_{\min}(\mathscr{A}_j \overline{y}) = 0$ and, therefore, such that $\lambda^j_{\min}(\mathscr{A}_j \overline{y}) \geq 0$ for $j = 1, \ldots, r$. But this implies that $A_N \overline{y} \in K_N$.

For the if direction, assume that there exists $\overline{y} \neq 0$ such that $A_N \overline{y} \in K_N$. Let $\mathcal{A} = (0, A_N)$. Then, for all $y \neq 0$, and since $0y = 0 \in \partial K_B$,

$$v_{\mathcal{A},\mathbf{1}}(y) = \min_{j=1,\ldots,r} \lambda^j_{\min}(\mathcal{A}_j y) \leq 0.$$

This implies that $\overline{v}_{\mathcal{A},\mathbf{1}} \leq 0$. But $v_{\mathcal{A},\mathbf{1}}(\overline{y}) = 0$ since $A_N \overline{y} \in K_N$. Therefore $\overline{v}_{\mathcal{A},\mathbf{1}} = 0$, which implies that $\mathcal{A} \in \Sigma$.

We now assume that the sets above are nonempty and take limits when $\alpha_N \to 0$. We will show that both the left- and right-hand sides of (1.9) tend to the corresponding sides in (1.10) when $\alpha_N \to 0$. Equation (1.10) will therefore hold since Theorem 1.1 does.

Recall that the left-hand side in the equality of Theorem 1.1 is

$$\left| \max_{y \in S_Y} \min_{j=1,\ldots,r} \frac{\lambda^j_{\min}(A_j y)}{\alpha_j} \right|.$$

For any $y \in S_Y$ such that $A_N y \in \mathsf{int}(K_N)$, we have

$$\begin{aligned} & \forall j \in N \ \frac{\lambda^j_{\min}(A_j y)}{\alpha_j} > 0 \\ \implies & \forall j \in N \ \lim_{\alpha_j \to 0} \frac{\lambda^j_{\min}(A_j y)}{\alpha_j} = +\infty \\ \implies & \lim_{\alpha_N \to 0} \min_{j=1,\ldots,r} \frac{\lambda^j_{\min}(A_j y)}{\alpha_j} = \min_{j \in B} \frac{\lambda^j_{\min}(A_j y)}{\alpha_j}. \end{aligned} \tag{3.5}$$

On the other hand, for any $y \in S_Y$ such that $A_N y \notin K_N$,

$$\lim_{\alpha_N \to 0} \min_{j=1,\ldots,r} \frac{\lambda^j_{\min}(A_j y)}{\alpha_j} = -\infty. \tag{3.6}$$

Finally, for any $y \in S_Y$ such that $A_N y \in \partial K_N$, we have

$$\lim_{\alpha_N \to 0} \min_{j=1,\dots,r} \frac{\lambda^j_{\min}(A_j y)}{\alpha_j} = \begin{cases} 0 & \text{if } A_B y \in \mathsf{int}(K_B), \\ \min_{j \in B} \frac{\lambda^j_{\min}(A_j y)}{\alpha_j} & \text{otherwise.} \end{cases} \tag{3.7}$$

By taking the maximum over $y \in S_Y$ on the equalities (3.5), (3.6), and (3.7) and using Lemma 3.3, it follows that

$$\max_{y \in S_Y} \lim_{\alpha_N \to 0} \min_{j=1,\dots,r} \frac{\lambda^j_{\min}(A_j y)}{\alpha_j} = \max_{\substack{y \in S_Y \\ A_N y \in K_N}} \min_{j \in B} \frac{\lambda^j_{\min}(A_j y)}{\alpha_j}.$$

Hence to show that the left-hand side in (1.9) tends to the left-hand side in (1.10) when $\alpha_N \to 0$, we need to show that

$$\max_{y \in S_Y} \lim_{\alpha_N \to 0} \min_{j=1,\dots,r} \frac{\lambda^j_{\min}(A_j y)}{\alpha_j} = \lim_{\alpha_N \to 0} \max_{y \in S_Y} \min_{j=1,\dots,r} \frac{\lambda^j_{\min}(A_j y)}{\alpha_j}. \tag{3.8}$$

If $\{y \neq 0 \mid A_N y \in K_N\} = \emptyset$, then from (3.6) if follows that both sides of (3.8) are $-\infty$. Assume that $\{y \neq 0 \mid A_N y \in K_N\} \neq \emptyset$. From Lemma 3.3 and (3.5) it follows that both sides of (3.8) are finite. Let $\epsilon > 0$ be given. By Lemma 3.3, (3.5), (3.6), and (3.7) there exists $y_\epsilon \in S_Y$ such that $A_N y_\epsilon \in \mathsf{int}(K_N)$ and

$$\max_{y \in S_Y} \lim_{\alpha_N \to 0} \min_{j=1,\dots,r} \frac{\lambda^j_{\min}(A_j y)}{\alpha_j} - \epsilon < \lim_{\alpha_N \to 0} \min_{j=1,\dots,r} \frac{\lambda^j_{\min}(A_j y_\epsilon)}{\alpha_j}.$$

Thus

$$\begin{aligned} \max_{y \in S_Y} \lim_{\alpha_N \to 0} \min_{j=1,\dots,r} \frac{\lambda^j_{\min}(A_j y)}{\alpha_j} - \epsilon &< \lim_{\alpha_N \to 0} \min_{j=1,\dots,r} \frac{\lambda^j_{\min}(A_j y_\epsilon)}{\alpha_j} \\ &\leq \lim_{\alpha_N \to 0} \max_{y \in S_Y} \min_{j=1,\dots,r} \frac{\lambda^j_{\min}(A_j y)}{\alpha_j}. \end{aligned}$$

This shows that the left-hand side in (3.8) is smaller than or equal to the right-hand side. For the reverse inequality let $\epsilon > 0$ be given. By Lemma 3.3, (3.5), (3.6), and (3.7) there exists $y_\epsilon \in S_Y$ such that $A_N y_\epsilon \in \mathsf{int}(K_N)$ and

$$\lim_{\alpha_N \to 0} \max_{y \in S_Y} \min_{j=1,\dots,r} \frac{\lambda^j_{\min}(A_j y)}{\alpha_j} - \epsilon < \min_{j=1,\dots,r} \frac{\lambda^j_{\min}(A_j y_\epsilon)}{\alpha_j}.$$

Hence (3.5) yields

$$\begin{aligned} \lim_{\alpha_N \to 0} \max_{y \in S_Y} \min_{j=1,\dots,r} \frac{\lambda^j_{\min}(A_j y)}{\alpha_j} - \epsilon &< \min_{j=1,\dots,r} \frac{\lambda^j_{\min}(A_j y_\epsilon)}{\alpha_j} \\ &= \lim_{\alpha_N \to 0} \min_{j=1,\dots,r} \frac{\lambda^j_{\min}(A_j y_\epsilon)}{\alpha_j} \\ &\leq \max_{y \in S_Y} \lim_{\alpha_N \to 0} \min_{j=1,\dots,r} \frac{\lambda^j_{\min}(A_j y)}{\alpha_j}. \end{aligned}$$

Therefore the right-hand side of (3.8) is also smaller than or equal to the left-hand side.

Next, we show that the right-hand side of (1.9), namely,

$$\min_{\widetilde{A}\in\Sigma}\max_{j=1,\dots,r}\frac{\|A_j-\widetilde{A}_j\|}{\alpha_j},$$

tends to the right-hand side of (1.10) when $\alpha_N\to 0$. Take $\widetilde{A}\in\Sigma$. Then

$$\begin{aligned}\widetilde{A}_N\neq A_N &\Longrightarrow \exists j\in N\ \|A_j-\widetilde{A}_j\|\neq 0\\ &\Longrightarrow \exists j\in N\ \lim_{\alpha_j\to 0}\frac{\|A_j-\widetilde{A}_j\|}{\alpha_j}=+\infty\\ &\Longrightarrow \lim_{\alpha_N\to 0}\max_{j=1,\dots,r}\frac{\|A_j-\widetilde{A}_j\|}{\alpha_j}=+\infty.\end{aligned}$$

This implies that

$$\lim_{\alpha_N\to 0}\min_{\widetilde{A}\in\Sigma}\max_{j=1,\dots,r}\frac{\|A_j-\widetilde{A}_j\|}{\alpha_j}=\lim_{\alpha_N\to 0}\min_{\substack{\widetilde{A}\in\Sigma\\ \widetilde{A}_N=A_N}}\max_{j=1,\dots,r}\frac{\|A_j-\widetilde{A}_j\|}{\alpha_j}. \tag{3.9}$$

But if $\widetilde{A}_N=A_N$, then $\frac{\|A_j-\widetilde{A}_j\|}{\alpha_j}=0$ for all $j\in N$. Therefore,

$$\lim_{\alpha_N\to 0}\min_{\substack{\widetilde{A}\in\Sigma\\ \widetilde{A}_N=A_N}}\max_{j=1,\dots,r}\frac{\|A_j-\widetilde{A}_j\|}{\alpha_j}=\min_{\substack{\widetilde{A}\in\Sigma\\ \widetilde{A}_N=A_N}}\max_{j\in B}\frac{\|A_j-\widetilde{A}_j\|}{\alpha_j},$$

and the claimed limit follows. ∎

REFERENCES

[1] F. Alizadeh and D. Goldfarb, *Second-order cone programming*, Math. Program., 95 (2003), pp. 3–51.

[2] J. Baez, *The octonions*, Bull. Amer. Math. Soc., 39 (2002), pp. 145–205.

[3] H.H. Bauschke, O. Güler, A.S. Lewis, and H.S. Sendov, *Hyperbolic polynomials and convex analysis*, Canad. J. Math., 53 (2001), pp. 470–488.

[4] J. Borwein and A. Lewis, *Convex Analysis and Nonlinear Optimization: Theory and Examples*, Springer-Verlag, New York, 2000.

[5] M. Cánovas, M. López, J. Parra, and F. Toledo, *Distance to ill-posedness and the consistency of linear semi-infinite inequality systems*, Math. Program., 103 (2005), pp. 95–126.

[6] D. Cheung and F. Cucker, *A new condition number for linear programming*, Math. Program., 91 (2001), pp. 163–174.

[7] D. Cheung and F. Cucker, *Probabilistic analysis of condition numbers for linear programming*, J. Optim. Theory Appl., 114 (2002), pp. 55–67.

[8] D. Cheung, F. Cucker, and R. Hauser, *Tail decay and moment estimates of a condition number for random linear conic systems*, SIAM J. Optim., 15 (2005), pp. 1237–1261.

[9] J. Conway and D. Smith, *On quaternions and octonions: Their geometry, arithmetic, and symmetry*, A K Peters, Wellesley, MA, 2003.

[10] F. Cucker and J. Peña, *A primal-dual algorithm for solving polyhedral conic systems with a finite-precision machine*, SIAM J. Optim., 12 (2002), pp. 522–554.

[11] F. Cucker and M. Wschebor, *On the expected condition number of linear programming problems*, Numer. Math., 94 (2002), pp. 419–478.

[12] T. Dray and C. Manogue, *The exceptional Jordan eigenvalue problem*, Internat. J. Theoret. Phys., 38 (1999), pp. 2901–2916.

[13] J. Dunagan, D.A. Spielman, and S.-H. Teng, *Smoothed Analysis of Renegar's Condition Number for Linear Programming*, preprint available at http://theory.lcs.mit.edu/ spielman, 2003.

[14] J. Faraut and A. Korányi, *Analysis on Symmetric Cones*, Clarendon Press, Oxford, 1994.
[15] L. Faybusovich, *Linear systems in Jordan algebras and primal-dual interior-point algorithms*, J. Comput. Appl. Math., 86 (1997), pp. 149–175.
[16] R. Freund, *On the behavior of the homogeneous self-dual model for conic convex optimization*, Math. Program., 109 (2007), pp. 445–475.
[17] R.M. Freund and J.R. Vera, *Condition-based complexity of convex optimization in conic linear form via the ellipsoid algorithm*, SIAM J. Optim., 10 (1999), pp. 155–176.
[18] R.M. Freund and J.R. Vera, *Some characterizations and properties of the "distance to ill-posedness" and the condition measure of a conic linear system*, Math. Program., 86 (1999), pp. 225–260.
[19] J.-L. Goffin, *The relaxation method for solving systems of linear inequalities*, Math. Oper. Res., 5 (1980), pp. 388–414.
[20] R. Hauser and T. Müller, *Algebraic Tail Decay of Condition Numbers for Random Conic Systems under a General Family of Input Distributions*, preprint available at http://www.optimization-online.org/DB_HTML/2006/02/1336.html, 2006.
[21] S. Lang, *Algebra*, 3rd ed., Addison-Wesley, Reading, MA, 1993.
[22] H. Lara and Tunçel, *Condition and Complexity Measures for Infeasibility Certificates of Systems of Linear Inequalities and Their Sensitivity Analysis*, Research report COOR 2002-10, Department of Combinatorics and Optimization, University of Waterloo, 2002.
[23] A. Lewis, *The structured distance to ill-posedness for conic systems*, Math. Oper. Res., 29 (2005), pp. 776–785.
[24] M. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, *Applications of second-order cone programming*, Linear Algebra Appl., 284 (1998), pp. 193–228.
[25] Yu. E. Nesterov and M.J. Todd, *Self-scaled barriers and interior-point methods for convex programming*, Math. Oper. Res., 22 (1997), pp. 1–42.
[26] Yu. E. Nesterov and M.J. Todd, *Primal-dual interior-point methods for self-scaled cones*, SIAM J. Optim., 8 (1998), pp. 324–364.
[27] F. Ordóñez and R. Freund, *Computational experience and the explanatory value of condition measures for linear optimization*, SIAM J. Optim., 14 (2003), pp. 307–333.
[28] J. Peña, *A characterization of the distance to infeasibility under block-structured perturbations*, Linear Algebra Appl., 370 (2003), pp. 193–216.
[29] J. Peña, *On the block-structured distance to non-surjectivity of sublinear mappings*, Math. Program, 103 (2005), pp. 561–573.
[30] J. Renegar, *Incorporating condition measures into the complexity theory of linear programming*, SIAM J. Optim., 5 (1995), pp. 506–524.
[31] J. Renegar, *Linear programming, complexity theory and elementary functional analysis*, Math. Program., 70 (1995), pp. 279–351.
[32] J. Renegar, *A Mathematical View of Interior-Point Methods in Convex Optimization*, SIAM, Philadelphia, 2000.
[33] J. Renegar, *Hyperbolic programs and their derivative relaxations*, Found. Comput. Math., 6 (2006), pp. 59–79.
[34] W. Rudin, *Real and Complex Analysis*, McGraw-Hill, New York, 1987.

SIAM J. OPTIM.
Vol. 19, No. 1, pp. 281–306

A TRUST REGION SPECTRAL BUNDLE METHOD FOR NONCONVEX EIGENVALUE OPTIMIZATION*

P. APKARIAN†, D. NOLL‡, AND O. PROT‡

Abstract. We present a nonsmooth optimization technique for nonconvex maximum eigenvalue functions and for nonsmooth functions which are infinite maxima of eigenvalue functions. We prove global convergence of our method in the sense that for an arbitrary starting point, every accumulation point of the sequence of iterates is critical. The method is tested on several problems in feedback control synthesis.

Key words. eigenvalue optimization, trust region method, proximity control, spectral bundle, output feedback control, H_∞-synthesis

AMS subject classifications. 90C26, 90C22, 93B36, 49J52

DOI. 10.1137/060665191

1. Introduction.

Eigenvalue optimization has a wide spectrum of applications in physics, engineering, statistics, and finance. This spectrum includes composite materials [15], quantum computational chemistry [60], optimal system design [47, 8], shape optimization [17], pole placement in linear system theory, robotics [44], relaxations of combinatorial optimization problems [26, 39], experimental design [55, 58], and much more. Many of these problems are nonconvex, but even in the realm of convexity, eigenvalue optimization has a prominent place. Semidefinite programming (SDP) is an important class of convex programs, which may be solved by way of eigenvalue optimization [48].

The idea of solving large semidefinite programs via eigenvalue optimization can be traced back to [16, 25, 41]. It results from the insight that interior-point methods are not the appropriate choice when problems are sizable. Due to its importance in practice, eigenvalue optimization has been intensively studied since the 1980s. Early contributions are Wolfe [59], Cullum, Donath, and Wolfe [16], Polak and Wardi [54], and Fletcher [19]. Starting in the late 1980s, Overton contributed a series of papers ([49, 50], and [51] with Womersley), where in particular Newton-type methods are discussed. Oustry [48] presents a synthesis of first and second order methods suited for convex maximum eigenvalue functions.

Here our interest is in nonconvex eigenvalue programs, which arise frequently in automatic control applications and especially in controller synthesis. In particular, solving bilinear matrix inequalities (BMIs) is a prominent application, which may be addressed via nonconvex eigenvalue optimization. In [45, 46] we have shown how to adapt the approach of [41, 48] to handle nonconvex situations. Applications and extensions of these ideas are presented in [4, 1, 57, 10, 5, 6].

*Received by the editors March 22, 2006; accepted for publication (in revised form) December 4, 2007; published electronically March 19, 2008. The authors acknowledge financial support from Agence Nationale de Recherche (ANR) under contract SSIA_NV_6 *Controvert* and contract NT05-1_43040 *Guidage*, and from *Fondation d'Entreprise EADS* under contract *Solving challenging problems in feedback control*.

http://www.siam.org/journals/siopt/19-1/66519.html

†CERT-ONERA, 2, avenue Edouard Belin, 31055 Toulouse, France (pierre.apkarian@cert.fr).

‡Université Paul Sabatier, Institut de Mathématiques, 118, route de Narbonne, 31062 Toulouse, France (noll@mip.ups-tlse.fr, oprot@mip.ups-tlse.fr).

The goal of the present paper is twofold. In the first part we investigate how to expand on the idea of Helmberg and Rendl's spectral bundle method [25] in order to deal with nonconvex eigenvalue programs. Nonconvexity requires a new approximation technique, complementing the convex mechanism used in [25]. We achieve our goal by a trust region approach or, what is equivalent, by a dual approach using proximity control. This method has antecedents in classical bundling, such as in Lemaréchal [36, 37, 38], Lemaréchal, Nemirovskii, and Nesterov [40], and Kiwiel [31, 32, 33]. Extensions of the convex case to include bound constraints are given in [23].

In the second part we extend our method to address more general classes of functions which are infinite suprema of maximum eigenvalue functions. This includes optimization of the H_∞-norm, an important application in feedback control synthesis. Optimization of the H_∞-norm has been pioneered by Polak and coworkers. See, for instance, [42, 43, 52], and the references given there. Our own approach to optimizing the H_∞-norm is developed in [4, 1, 5].

The structure of the paper is as follows. After some preparations in sections 2–5, the algorithm is presented in section 6. Convergence analysis follows in sections 7 and 8. The semi-infinite case, which includes optimization of the H_∞-norm, is presented in section 9. While the main objective of this work is the convergence analysis of our method, we have added several numerical tests for eigenvalue programs in section 10 to validate the algorithm. Numerical tests for the H_∞-norm and for related problems are presented in [7].

Notation. Our terminology follows [28] and [14]. We let $\|\cdot\|$ denote the Euclidean norm on the space $\mathbb{R}^n$ equipped with the scalar product $x^\top y$, while the space $\mathbb{S}^m$ of $m \times m$ symmetric matrices is equipped with the scalar product $X \bullet Y = \mathrm{Tr}(XY)$. The corresponding matrix norm is also denoted by $\|\cdot\|$. For $X \in \mathbb{S}^m$, $X \preceq 0$ means X is negative semidefinite.

2. Elements from nonsmooth analysis. Recall that the maximum eigenvalue function $\lambda_1 : \mathbb{S}^m \to \mathbb{R}$ is convex but generally nonsmooth and defined on the space $\mathbb{S}^m$ of $m \times m$ symmetric or Hermitian matrices. We consider composite functions of the form $f = \lambda_1 \circ F$, where $F : \mathbb{R}^n \to \mathbb{S}^m$ is a class C^2 operator. Notice that f is nonsmooth, due to nonsmoothness of λ_1, and nonconvex unless F is an affine operator. The case where F is affine and therefore f convex has been studied by many authors [16, 25, 41]. Here our interest is focused on handling nonconvex f.

Notice that as a composite function, f has a favorable structure. In particular, the Clarke subdifferential [14] is given by the chain rule

$$\partial f(x) = F'(x)^\star \partial\lambda_1\left(F(x)\right),$$

where $\partial\lambda_1(X)$ is the usual subdifferential of convex analysis at $X \in \mathbb{S}^m$, and where $F'(x)$ is the derivative of F, $F'(x)^\star$ its adjoint, mapping $\mathbb{S}^m$ back into $\mathbb{R}^n$. Recall that λ_1 is itself highly structured, as it is the support function of the convex compact set

$$\mathcal{G} = \{G \in \mathbb{S}^m : G \succeq 0, \mathrm{Tr}(G) = 1\}.$$

That means $\lambda_1(X) = \max\{G \bullet X : G \in \mathcal{G}\}$, and therefore (cf. [28, I, sect. 5.1, p. 275])

$$\partial\lambda_1(X) = \{G \in \mathcal{G} : G \bullet X = \lambda_1(X)\}.$$

3. First local model. Consider the minimization of $f = \lambda_1 \circ F$ over $\mathbb{R}^n$ and suppose $x \in \mathbb{R}^n$ is the current iterate. In order to generate a descent step from x to y, we may consider the following convex model of f around x:

$$\phi(y;x) = \lambda_1 \left(F(x) + F'(x)(y-x)\right), \tag{1}$$

where $y \mapsto F(x) + F'(x)(y-x)$ is the first order affine approximation of $F(y)$ in a neighborhood of x. Clearly $\phi(x;x) = f(x)$, and $f = \phi$ in those cases where F itself is affine (see, e.g., [16, 48, 25]). Taylor's theorem suggests that $\phi(y;x) \approx f(y)$ for y sufficiently close to x. This observation is made precise by the following.

PROPOSITION 1. *For every bounded set $B \subset \mathbb{R}^n$ there exists a constant $L > 0$ such that*

$$|f(y) - \phi(y;x)| \leq L\|y-x\|^2 \tag{2}$$

for all $x, y \in B$.

Proof. Notice that for any given matrices $A, E \in \mathbb{S}^m$, the estimate

$$\lambda_m(E) \leq \lambda_1(A+E) - \lambda_1(A) \leq \lambda_1(E)$$

is satisfied. This is also known as Weyl's theorem. Now as F is of class C^2, expanding at $x \in B$ gives $F(y) = F(x) + F'(x)(y-x) + R(y;x)$ with $\|R(y;x)\| \leq L\|y-x\|^2$ for some constant $L > 0$ and all $x, y \in B$. Using $X = F(x)$, $D = F'(x)(y-x)$, we have $f(y) = \lambda_1 \left(X + D + R(y;x)\right)$. We now apply Weyl's theorem with $A = X + D$, $E = R(y;x)$, which gives

$$\begin{aligned} |f(y) - \phi(y;x)| &= |\lambda_1(X+D) - \lambda_1(X+D+R(y;x))| \\ &\leq \max\left\{|\lambda_1\left(R(y;x)\right)|, |\lambda_m\left(R(y;x)|\right\}\right. \\ &\leq \|R(y;x)\| \leq L\|y-x\|^2 \end{aligned}$$

for all $x, y \in B$. That proves the claim. ∎

4. Second local model. Along with (1) we consider a second local model of f in a neighborhood of the current iterate x, which we update recursively. Notice that

$$\phi(y;x) = \max\left\{G \bullet \left[F(x) + F'(x)(y-x)\right] : G \in \mathcal{G}\right\},$$

where $\mathcal{G} = \{G \in \mathbb{S}^m : \mathrm{Tr}(G) = 1, G \succeq 0\}$ as before. This suggests the following approximation $\phi_k(y;x)$ of f, where $\mathcal{G}$ is replaced by a smaller and easier-to-compute subset $\mathcal{G}_k \subset \mathcal{G}$. We will generate a sequence $\mathcal{G}_k \subset \mathcal{G}$ of such approximations and let

$$\phi_k(y;x) = \max\left\{G \bullet \left[F(x) + F'(x)(y-x)\right] : G \in \mathcal{G}_k\right\}. \tag{3}$$

Clearly $\phi_k \leq \phi$. The idea is that we generate descent steps for the $\phi_k(\cdot;x)$, which will ultimately lead to descent in f at x, as the agreement between f and ϕ_k improves at each step k. The minimal requirement for $\mathcal{G}_k$ is the following and is obvious.

LEMMA 2. *Suppose $\mathcal{G}_k$ contains a subgradient of the form $G = ee^\top \in \mathcal{G}$, where e is a normalized eigenvector associated with $\lambda_1\left(F(x)\right)$. Then $\phi_k(x;x) = f(x)$.*

5. Proximity control. Let x be our current iterate. In order to generate a new trial step, we use the current model $\phi_k(\cdot;x)$ and compute the solution y^{k+1} of the unconstrained optimization program,

$$\text{minimize} \quad \phi_k(y;x) + \tfrac{\tau_k}{2}\|y-x\|^2, \quad y \in \mathbb{R}^n, \tag{4}$$

where $\tau_k \geq 0$ is the proximity parameter, and where the term $\frac{\tau_k}{2}\|y-x\|^2$ is referred

to as the proximity control. It is well known (see, e.g., [28, II, Prop. 2.2.3, p. 291]) that (4) is equivalent to a trust region program of the form

$$\begin{aligned} &\text{minimize} && \phi_k(y;x), \quad y \in \mathbb{R}^n \\ &\text{subject to} && \|y-x\| \le t_k, \end{aligned} \tag{5}$$

where $t_k > 0$ is the trust region radius. Indeed, minima of (5) and minima of (4) are in one-to-one correspondence in the following sense: If y^{k+1} is a minimum of (4) for fixed $\tau_k > 0$, then y^{k+1} also solves (5) with $t_k := \|y^{k+1} - x\|$ and associated multiplier $\tau_k > 0$. Conversely, if y^{k+1} solves (5) and if the associated multiplier τ_k is strictly positive, then y^{k+1} solves (4) with that proximity parameter $\tau_k > 0$. The case $\tau_k = 0$ in (4) obviously corresponds to those cases in (5) where the trust region constraint is inactive.

With the models ϕ and ϕ_k we introduce two levels of approximation of f, so it is not surprising that two mechanisms to adjust the degree of exactness are applied. First, in order to control the agreement between f and ϕ, we need to adjust t_k at each step, which is done indirectly via the management of τ_k. If the agreement between f and ϕ is good, we increase t_k, which corresponds to decreasing τ_k, while we have to reduce t_k when agreement is bad, achieved indirectly by increasing τ_k. Second, we update $\mathcal{G}_k$ into $\mathcal{G}_{k+1}$ after each trial y^{k+1} in order to drive ϕ_k closer to ϕ, and thereby also closer to f. We use the standard terminology in nonsmooth optimization. If the solution y^{k+1} of (4) is not used as the next iterate, we call it a null step. If y^{k+1} is accepted and becomes the next iterate x^+, we speak of a serious step.

In order to test the quality of the trial steps y^{k+1}, we use the quotient

$$\rho_k = \frac{f(x) - f(y^{k+1})}{f(x) - \phi_k(y^{k+1};x)}. \tag{6}$$

Fixing constants $0 < \gamma < \Gamma < 1$, we say that f and $\phi_k(\cdot;x)$ are in good agreement when $\rho_k > \Gamma$, and we say that the agreement is bad if $\rho_k < \gamma$. The bad case includes, in particular, situations where $\rho_k \le 0$. Since always $f(x) - \phi_k(y^{k+1};x) > 0$, unless $0 \in \partial f(x)$, we deduce that $\rho_k \le 0$ corresponds to cases where $f(x) - f(y^{k+1}) \le 0$, that is, where the proposed step y^{k+1} is not even a descent step for f. In our algorithm we use the following rule: y^{k+1} is accepted as soon as $\rho_k \ge \gamma$, i.e., as soon as the step is not bad. The question is then what we shall do when agreement between f and ϕ_k is bad, i.e., when $\rho_k < \gamma$.

Here we compute a second test parameter,

$$\widetilde{\rho}_k := \frac{f(x) - \phi(y^{k+1};x)}{f(x) - \phi_k(y^{k+1};x)},$$

and we compare it to a second control parameter $\widetilde{\gamma}$, where $\gamma < \widetilde{\gamma} < \frac{1}{2}$. We then have two possibilities. If $\rho_k < \gamma$ and also $\widetilde{\rho}_k < \widetilde{\gamma}$, then we do not change τ_k, but improve the approximation $\mathcal{G}_{k+1}$ so that ϕ_{k+1} gets closer to ϕ. On the other hand, if $\rho_k < \gamma$, but $\widetilde{\rho}_k \ge \widetilde{\gamma}$, then ϕ and f are not in good agreement, while ϕ_k is already close to ϕ. Driving ϕ_k even closer to ϕ in that case alone will therefore not improve the situation. Here we have to decrease the trust region radius t_k, or what comes down to the same, increase the proximity control parameter τ_k. While doing this, we still update $\mathcal{G}_k$ to a better $\mathcal{G}_{k+1}$, i.e., we still let ϕ_k approach ϕ, so this process is always applied.

6. Aggregate subgradients. As our convergence analysis will show, the approximations $\mathcal{G}_k \subset \mathcal{G}$ need only satisfy the following three conditions:

(G_1) $G_0 = e_0 e_0^\top \in \mathcal{G}_k$ for some normalized eigenvector e_0 associated with $\lambda_1\,(F(x))$.

(G_2) $G_{k+1} = e_{k+1}e_{k+1}^\top \in \mathcal{G}_{k+1}$ for some normalized eigenvector e_{k+1} associated with $\lambda_1\big(F(x) + F'(x)(y^{k+1} - x)\big)$.

(G_3) $G_k^* \in \mathcal{G}_{k+1}$ for some of the $G_k^* \in \mathcal{G}_k$, where the maximum $\phi_k(y^{k+1}; x)$ is attained, and which satisfies $0 = F'(x)^\star G_k^* + \tau_k(y^{k+1} - x)$.

Below we will discuss practical choices of the sets $\mathcal{G}_k$, combining ideas from [25], [48], [24], and [45]. We let $\mathcal{G}_k$ consist of sets of the form

$$\alpha_k \overline{G}_k + Q_k Y_k Q_k^\top, \tag{7}$$

where $Y_k \in \mathbb{S}^{r_k}$ has $Y_k \succeq 0$, $\alpha_k + \mathrm{Tr}(Y_k) = 1$, $0 \le \alpha_k \le 1$, where Q_k is an $m \times r_k$ matrix whose $r_k \ge 1$ columns form an orthogonal basis of an invariant subspace of $F(x) + F'(x)(y^{k+1} - x)$, and where $\overline{G}_k \in \mathcal{G}$ is the aggregate subgradient. We assume that at least one normalized eigenvector e_k associated with the maximum eigenvalue $\lambda_1\big(F(x) + F'(x)(y^k - x)\big)$ is in the span of the columns of Q_k, and moreover, that e_0 is in the span of the columns of Q_k at all times. The idea is to build the new set $\mathcal{G}_{k+1}$ along the same lines, using an updating strategy, which we now explain.

Let y^{k+1} be the solution of program (4), obtained with the help of $\mathcal{G}_k$, and suppose it is a null step. The necessary optimality condition gives $0 \in \partial\phi_k(y^{k+1}; x) + \tau_k(y^{k+1} - x)$. Due to the structure of ϕ_k and (7), this means there exist $G_k^* \in \mathcal{G}_k$ such that

$$0 = F'(x)^\star G_k^* + \tau_k(y^{k+1} - x), \quad G_k^* = \alpha_k^* \overline{G}_k + Q_k Y_k^* Q_k^\top, \tag{8}$$

where $0 \le \alpha_k^* \le 1$, $Y_k^* \succeq 0$, and $\alpha_k^* + \mathrm{Tr}(Y_k^*) = 1$. Now the simplest method is to let

$$\overline{G}_{k+1} = \alpha_k^* \overline{G}_k + Q_k Y_k^* Q_k^\top, \tag{9}$$

the new aggregate subgradient. Helmberg and Rendl [25] use a refinement of (9), which is suited for large problem size: Let $Y_k^* = PDP^\top$ be a spectral decomposition of the $r_k \times r_k$ matrix Y_k^*. Decompose $P = [P_1 P_2]$ with corresponding spectra D_1 and D_2 so that P_1 contains as columns those eigenvectors associated with the large eigenvalues of Y_k^*, and P_2 are the remaining columns. Now put

$$\overline{G}_{k+1} = \big(\alpha_k^* \overline{G}_k + Q_k P_2 D_2 P_2^\top Q_k^\top\big) \,/\, (\alpha_k^* + \mathrm{Tr}(D_2))\,, \tag{10}$$

the new aggregate subgradient, which is an element of $\mathcal{G}$. In this way only the minor part of Y_k^* is kept in the aggregate subgradient. The dominant part of Y_k^* is retained in the next eigenbasis by letting $Q_k P_1$ be part of Q_{k+1}. Moreover, in view of axiom (G_2), one eigenvector e_{k+1} of the maximum eigenvalue of $F(x) + F'(x)(y^{k+1} - x)$ is computed and included in Q_{k+1}. In order to guarantee axiom (G_1), we also keep at least one normalized eigenvector e_0 associated with the maximum eigenvalue of $F(x)$ in Q_{k+1}.

Altogether, $\mathcal{G}_{k+1}$ consists of all $\alpha\overline{G}_{k+1} + Q_{k+1}Y_{k+1}Q_{k+1}^\top$, where $0 \le \alpha \le 1$, $Y_{k+1} \succeq 0$, and $\alpha + \mathrm{Tr}(Y_{k+1}) = 1$, and where Q_{k+1} has the properties above. For this construction we have the following.

LEMMA 3. *The sets $\mathcal{G}_k$ so defined satisfy the rules (G_1)–(G_3). In particular, $\phi_k(x; x) = f(x)$, $\phi_{k+1}(y^{k+1}; x) = \phi(y^{k+1}; x)$, $\phi_{k+1}(y^{k+1}; x) \ge \phi_k(y^{k+1}; x)$, and condition (8) hold for every k.*

Remark. In a traditional bundle method we would refer to $\overline{g}_k = F'(x)^\star \overline{G}_k$ as the aggregate subgradient. Here we use the term aggregate for both $\overline{g}_k$ and $\overline{G}_k$ because

there is no risk of ambiguity. But are both elements really needed? The authors of [25] point out that storing $\bar{g}_k$ is cheaper than storing $\overline{G}_k$, so naturally they work in g-space and not in G-space.

Now observe an important difference of our present case with the convex case in [25], where $F'(x)^* = A^*$ is independent of x. Since our $F'(x)^*$ depends on x, as soon as a serious step $x \to x^+$ is taken, $\bar{g}_k = F'(x)^*\overline{G}_k$ is no longer useful at x^+, because it is no longer a subgradient of f at x^+. However, $\overline{G}_k$ is still useful. It suffices to replace $\bar{g}_k$ by $\bar{g}_k^+ := F'(x^+)^*\overline{G}_k$, which is a subgradient for f at x^+. So if we want to "recycle" old aggregates in the next serious loop, we have to store $\overline{G}_k$ and not $\bar{g}_k$. On the other hand, if we are happy to keep aggregates only within one single inner loop, then we do not need $\overline{G}_k$ and our case is similar to [25].

Remark. Conditions (G_1)–(G_3) leave a lot of freedom for the choice of the bases Q_k. In [24] Helmberg and Oustry investigate the convex case and discuss ways to combine their two approaches [25] and [48] into a unified method. An alternative approach is Polak and Wardi [54]. For nonconvex eigenvalue functions we have proposed in [45, 46] an extension of Oustry's approach.

Spectral bundle algorithm for $\min_{x\in\mathbb{R}^n} f(x)$.

Parameters: $0 < \gamma < \widetilde{\gamma} < \Gamma < 1$.

0. **Initialize outer loop.** Find initial iterate x and compute $f(x)$.
1. **Outer loop.** Stop if $0 \in \partial f(x)$ at current outer iterate x. Otherwise goto inner loop.
2. **Initialize inner loop.** Let $\mathcal{G}_1 \subset \partial\lambda_1(F(x))$, $\overline{G}_1 = \frac{1}{m} I_m$, put inner loop counter $k = 1$, and choose $\tau_1 > 0$. If old value for τ from previous sweep is memorized, use it to initialize τ_1.
3. **Tangent program.** At counter k with given $\tau_k > 0$ and $\mathcal{G}_k$ solve
$$\min_{y\in\mathbb{R}^n} \phi_k(y;x) + \frac{\tau_k}{2}\|y - x\|^2.$$
Solution is y^{k+1}. Find $G_k^* \in \mathcal{G}_k$ where (3) at y^{k+1} is attained. Write $G_k^* = \alpha_k^*\overline{G}_k + Q_k Y_k Q_k^\top$ according to (7).
4. **Acceptance test.** Compute $f(y_{k+1})$ and check whether
$$\rho_k = \frac{f(x) - f(y^{k+1})}{f(x) - \phi_k(y^{k+1};x)} \geq \gamma.$$
If this is the case, put $x^+ = y^{k+1}$ (serious step). Compute new memory element τ^+ as
$$\tau^+ = \begin{cases} \dfrac{\tau_k}{2} & \text{if } \rho_k > \Gamma, \\ \tau_k & \text{else} \end{cases}$$
Then go back to step 1 to commence a new sweep of outer loop. On the other hand, if $\rho_k < \gamma$, then continue inner loop with step 5.
5. **Agreement test.** Compute $\phi(y^{k+1};x)$ and control parameter
$$\widetilde{\rho}_k = \frac{f(x) - \phi(y^{k+1};x)}{f(x) - \phi_k(y^{k+1};x)}.$$
Put
$$\tau_{k+1} = \begin{cases} \tau_k & \text{if } \rho_k < \gamma \text{ and } \widetilde{\rho}_k < \widetilde{\gamma}, \\ 2\tau_k & \text{if } \rho_k < \gamma \text{ and } \widetilde{\rho}_k \geq \widetilde{\gamma}. \end{cases}$$
6. **Aggregate subgradient.** Compute new set $\mathcal{G}_{k+1}$ according to (7), with (G_1)–(G_3) satisfied. New aggregate subgradient is $\overline{G}_{k+1} = G_k^*$.
7. **Inner loop.** Increase counter $k \to k+1$ and go back to step 3.

7. Convergence analysis of inner loop. We have to show that the inner loop is finite, that is, finds a trial point y^{k+1} accepted in step 4 after a finite number k of steps. We prove this by showing that if the inner loop turns forever, that is, $\rho_k < \gamma$ for all k, then $0 \in \partial f(x)$. (Since the inner loop is not entered when $0 \in \partial f(x)$, this is an argument by contradiction.) There are two subcases to be discussed, depending on the decision in step 5. These will be addressed in Lemmas 4 and 5.

Our first concern is when $\rho_k < \gamma$ but $\widetilde{\rho}_k \geq \widetilde{\gamma}$. This is indeed the situation where we are far from the convex case. Namely, $\widetilde{\rho}_k \geq \widetilde{\gamma}$ means that ϕ_k is in good agreement with ϕ, but unfortunately $\rho_k < \gamma$ says that ϕ is not a good model of f, which is usually due to the fact that f is nonconvex in a neighborhood of the current x. In consequence, ϕ_k cannot be expected to be a good model of f either. This is addressed in step 5 of the algorithm by increasing the proximity parameter τ_k, which as we know is equivalent to reducing the trust region radius. This is the only way to improve the agreement between ϕ and f.

LEMMA 4. *Suppose the algorithm generates an infinite sequence of trial steps y^{k+1} such that always $\rho_k < \gamma$. Then $\widetilde{\rho}_k < \widetilde{\gamma}$ for some k_0 and all $k \geq k_0$.*

Proof. (i) Assume on the contrary that $\rho_k < \gamma$ for all k, but at the same time $\widetilde{\rho}_k \geq \widetilde{\gamma}$ for infinitely many $k \in \mathbb{N}$. Then according to the update rule in step 5 of the algorithm, the sequence τ_k tends to $+\infty$. As a consequence of the necessary optimality condition we have $0 \in \partial\phi_k(y^{k+1}; x) + \tau_k(y^{k+1} - x)$. Now observe that due to the special form (7), the subgradients of all functions $\phi_k(\cdot; x)$ are uniformly bounded by $\|F'(x)^\star\|$. Given that $\tau_k \to \infty$, we then must have $y^{k+1} \to x$. Using $\tau_k(x - y^{k+1}) \in \partial\phi_k(y^{k+1}; x)$ calls for the subgradient inequality, which gives

$$\tau_k \left(x - y^{k+1}\right)^\top \left(x - y^{k+1}\right) \leq \phi_k(x; x) - \phi_k(y^{k+1}; x) = f(x) - \phi_k(y^{k+1}; x),$$

the latter by Lemma 3. In other words,

$$\frac{\tau_k \|x - y^{k+1}\|^2}{f(x) - \phi_k(y^{k+1}; x)} \leq 1. \tag{11}$$

(ii) Now we expand the test parameters as follows:

$$\begin{aligned}
\widetilde{\rho}_k &= \rho_k + \frac{f(y^{k+1}) - \phi(y^{k+1}; x)}{f(x) - \phi_k(y^{k+1}; x)} \\
&\leq \rho_k + \frac{L\|x - y^{k+1}\|^2}{f(x) - \phi_k(y^{k+1}; x)} \qquad \text{(using Proposition 1)} \\
&\leq \rho_k + \frac{L}{\tau_k} \qquad \text{(using (11))}.
\end{aligned}$$

Since $\tau_k \to \infty$, we deduce $\limsup_{k\to\infty} \widetilde{\rho}_k \leq \limsup_{k\to\infty} \rho_k \leq \gamma$, contradicting $\widetilde{\rho}_k \geq \widetilde{\gamma} > \gamma$ for infinitely many k. ∎

Remark. Notice that the proof of Lemma 4 uses Lemma 3, which in turn exploits axiom (G_1). Axioms (G_2) and (G_3) will be needed in the next lemma.

As a consequence of Lemma 4 we see that the algorithm, when faced with the bad case $\rho_k < \gamma$, will continue to increase τ_k, until eventually $\widetilde{\rho}_k < \widetilde{\gamma}$, too. From some index k_0 onwards, we will then be in the first case in step 5 of the algorithm, where the parameter τ_k is frozen, i.e., $\tau_k =: \tau$ for $k \geq k_0$. We then have no easy argument to deduce $y^{k+1} \to x$. Here, indeed, we will have to exploit properties (G_2) and (G_3) of the update rule $\mathcal{G}_k \to \mathcal{G}_{k+1}$. We follow the line of [25, Lemma 4.2].

LEMMA 5. *Let x be the current iterate and suppose the algorithm generates an infinite sequence of trial steps y^{k+1}, where $\rho_k < \gamma$ for all k while $\widetilde{\rho}_k < \widetilde{\gamma}$ for some k_0 and all $k \geq k_0$. Then $0 \in \partial f(x)$.*

Proof. (i) As we mentioned already, $\rho_k < \gamma$ and $\widetilde{\rho}_k < \gamma$ for all $k \geq k_0$ implies that in step 5 of the algorithm, $\tau_k =: \tau$ is frozen for $k \geq k_0$. A priori we therefore do not know whether $y^{k+1} \to x$, as we did in the proof of Lemma 4. This complicates the following analysis.

(ii) Let us introduce the function

$$\psi_k(y;x) = \phi_k(y;x) + \tfrac{\tau}{2}\|y-x\|^2;$$

then by its definition, y^{k+1} is the global minimum of $\psi_k(\cdot;x)$ for $k \geq k_0$. Let $G_k^* \in \mathcal{G}_k$ be the subgradient where the supremum $\phi_k(y^{k+1};x)$ is attained and which is retained in $\mathcal{G}_{k+1}$ in accordance with rule (G_3) and also with step 3 of the algorithm. That means

$$\phi_k(y^{k+1};x) = G_k^* \bullet \left[F(x) + F'(x)(y^{k+1}-x)\right], \tag{12}$$

and also

$$\psi_k(y^{k+1};x) = G_k^* \bullet \left[F(x) + F'(x)(y^{k+1}-x)\right] + \tfrac{\tau}{2}\|y^{k+1}-x\|^2. \tag{13}$$

We introduce the function

$$\psi_k^*(y;x) = G_k^* \bullet [F(x) + F'(x)(y-x)] + \tfrac{\tau}{2}\|y-x\|^2.$$

Then $\psi_k^*(y^{k+1};x) = \psi_k(y^{k+1};x)$ and

$$\psi_k^*(y;x) \leq \psi_{k+1}(y;x) \tag{14}$$

for $k \geq k_0$, because $G_k^* \in \mathcal{G}_{k+1}$. We claim that

$$\psi_k^*(y;x) = \psi_k^*(y^{k+1};x) + \tfrac{\tau}{2}\|y-y^{k+1}\|^2. \tag{15}$$

An easy way to see this is to observe that ψ_k^* is quadratic and expand it, using $\nabla\psi_k^*(y;x) = F'(x)^\star G_k^* + \tau(y-x)$ and $\nabla^2\psi_k^*(y;x) = \tau I$. Then clearly,

$$\psi_k^*(y;x) = \psi_k^*(y^{k+1};x) + \nabla\psi_k^*(y^{k+1};x)^\top(y-y^{k+1}) + \tfrac{\tau}{2}(y-y^{k+1})^\top(y-y^{k+1}).$$

Formula (15) will therefore be established as soon as we show that the first order term in this expansion vanishes. But this term is

$$\begin{aligned}
&\nabla\psi_k^*(y^{k+1};x)^\top(y-y^{k+1}) \\
&\quad = \left(F'(x)^\star G_k^*\right)^\top(y-y^{k+1}) + \tau(y^{k+1}-x)^\top(y-y^{k+1}) \\
&\quad = \tau(x-y^{k+1})^\top(y-y^{k+1}) + \tau(y^{k+1}-x)^\top(y-y^{k+1}) \quad \text{(using (8))} \\
&\quad = 0.
\end{aligned}$$

That proves formula (15).

(iii) From (ii) we have

$$\begin{aligned}
\psi_k(y^{k+1};x) &\le \psi_k^*(y^{k+1};x) + \tfrac{\tau}{2}\|y^{k+2}-y^{k+1}\|^2 \quad (\text{using } \psi_k^*(y^{k+1};x) = \psi_k(y^{k+1};x)) \\
&= \psi_k^*(y^{k+2};x) \qquad (\text{using (15)}) \\
&\le \psi_{k+1}(y^{k+2};x) \qquad (\text{using (14)}) \\
&\le \psi_{k+1}(x;x) \qquad (y^{k+2} \text{ is minimizer of } \psi_{k+1}) \qquad (16) \\
&= \phi_{k+1}(x;x) \le \phi(x;x).
\end{aligned}$$

We deduce that the sequence $\psi_k(y^{k+1};x)$ is monotonically increasing and bounded above by $\phi(x;x)$. It therefore converges to some $\psi^* \le \phi(x;x)$.

Going back to (16) with this information, we see that the term $\frac{\tau}{2}\|y^{k+2}-y^{k+1}\|^2$ is now squeezed in between two convergent terms with the same limit ψ^*, and must therefore tend to zero. Consequently, $\|y^{k+1}-x\|^2-\|y^{k+2}-x\|^2$ also tends to 0, because the sequence y^k is bounded. (Boundedness of the y^{k+1} was already used in the proof of the previous lemma and follows from the particular form (7) of the subgradients and the fact that the sequence τ_k is nondecreasing and therefore bounded away from 0.)

Recalling $\phi_k(y;x) = \psi_k(y;x) - \frac{\tau}{2}\|y-x\|^2$, we deduce, using both convergence results, that

$$\begin{aligned}
&\phi_{k+1}(y^{k+2};x) - \phi_k(y^{k+1};x) \\
&\qquad = \psi_{k+1}(y^{k+2};x) - \psi_k(y^{k+1};x) - \tfrac{\tau}{2}\|y^{k+2}-x\|^2 + \tfrac{\tau}{2}\|y^{k+1}-x\|^2 \to 0. \qquad (17)
\end{aligned}$$

(iv) Let e_{k+1} be the normalized eigenvectors of $F(x)+F'(x)(y^{k+1}-x)$ associated with λ_1, which we pick in step 5 of the algorithm and according to rule (G_2). Then $g_k = F'(x)^\star e_{k+1}e_{k+1}^\top$ is a subgradient of $\phi_{k+1}(\cdot;x)$ at y^{k+1}. Hence by the subgradient inequality

$$\phi_{k+1}(y^{k+1};x) + g_k^\top\left(y-y^{k+1}\right) \le \phi_{k+1}(y;x).$$

Since $\phi_{k+1}(y^{k+1};x) = \phi(y^{k+1};x)$ by Lemma 3, respectively, rule (G_2), we have the estimate

$$\phi(y^{k+1};x) + g_k^\top\left(y-y^{k+1}\right) \le \phi_{k+1}(y;x). \qquad (18)$$

Now observe that

$$\begin{aligned}
0 &\le \phi(y^{k+1};x) - \phi_k(y^{k+1};x) \\
&= \phi(y^{k+1};x) + g_k^\top\left(y^{k+2}-y^{k+1}\right) - \phi_k(y^{k+1};x) - g_k^\top\left(y^{k+2}-y^{k+1}\right) \\
&\le \phi_{k+1}(y^{k+2};x) - \phi_k(y^{k+1};x) + \|g_k\|\left\|y^{k+2}-y^{k+1}\right\| \qquad (\text{using (18)}),
\end{aligned}$$

and this term tends to 0 due to (17), because $y^{k+2}-y^{k+1} \to 0$, and because the sequence g_k is bounded. We deduce that $\phi(y^{k+1};x) - \phi_k(y^{k+1};x) \to 0$.

(v) We now show that $\phi_k(y^{k+1};x) \to f(x)$, and then of course also $\phi(y^{k+1};x) \to f(x)$. Assume, contrary to what is claimed, that $\limsup_{k\to\infty} f(x) - \phi_k(y^{k+1};x) =: \eta > 0$. Choose $\delta > 0$ such that $\delta < (1-\widetilde{\gamma})\eta$. It follows from part (iv) that there exists $k_1 \geq k_0$ such that

$$\phi(y^{k+1};x) - \delta \leq \phi_k(y^{k+1};x) \tag{19}$$

for all $k \geq k_1$. Using $\widetilde{\rho}_k \leq \widetilde{\gamma}$ for $k \geq k_1$ then gives

$$\begin{aligned} \widetilde{\gamma}\left(\phi_k(y^{k+1};x) - f(x)\right) &\leq \phi(y^{k+1};x) - f(x) \\ &\leq \phi_k(y^{k+1};x) + \delta - f(x), \end{aligned} \tag{20}$$

which implies $\widetilde{\gamma}\eta \geq \eta - \delta$. This contradicts the choice of δ and therefore shows $\eta = 0$.

(vi) Having shown $\phi(y^{k+1};x) \to f(x)$ and $\phi_k(y^{k+1};x) \to f(x)$, we argue that $y^{k+1} \to x$. This follows from the definition of y^{k+1}, because

$$\psi_k(y^{k+1};x) = \phi_k(y^{k+1};x) + \tfrac{\tau}{2}\|y^{k+1}-x\|^2 \leq \psi_k(x;x) = \phi_k(x;x) = f(x).$$

Since $\phi_k(y^{k+1};x) \to f(x)$, we have indeed $\|y^{k+1}-x\| \to 0$ by a sandwich argument.

To finish the proof, let us show that $0 \in \partial f(x)$. Notice first that the necessary optimality condition gives $0 \in \partial\psi_k(y^{k+1};x) = \partial\phi_k(y^{k+1};x) + \tau(y^{k+1}-x)$, which implies

$$\tau(x - y^{k+1}) \in \partial\phi_k(y^{k+1};x).$$

The subgradient inequality gives

$$\begin{aligned} \tau(x-y^{k+1})^\top(y-y^{k+1}) &\leq \phi_k(y;x) - \phi_k(y^{k+1};x) \\ &\leq \phi(y;x) - \phi_k(y^{k+1};x) \quad (\text{using } \phi_k \leq \phi) \end{aligned}$$

for every y. Passing to the limit, observing $\tau(y^{k+1}-x) \to 0$ and $\phi_k(y^{k+1};x) \to \phi(x;x)$, we obtain the estimate

$$0 \leq \phi(y;x) - \phi(x;x)$$

for every y, which by convexity of $\phi(\cdot;x)$ implies $0 \in \partial\phi(x;x)$. Since $\partial\phi(x;x) = \partial f(x)$, we have shown $0 \in \partial f(x)$, as claimed. □

Remark. Various modifications of our algorithm may be considered. For instance, whenever a null step y^{k+1} is made, that is, $\rho_k < \gamma$, we should first check whether y^{k+1} gives descent in f:

$$f(x) - f(y^{k+1}) \geq \delta_1 > 0. \tag{21}$$

If this is not the case, the trust region radius is certainly too large, so we should increase τ_k right away. As presented, this will also happen, but after several null steps, bringing ϕ_k closer to ϕ, until the criterion in step 4 is met.

In the same vein, even when y^{k+1} gives descent in f, but slightly, so that $\rho_k > \gamma$ fails, we may check whether

$$\sigma_k := \frac{f(x) - f(y^{k+1})}{f(x) - \phi(y^{k+1};x)} \geq \delta_2 \tag{22}$$

for some $\frac{1}{2} < \delta_2 < 1$. If $\sigma_k < \delta_2$, then f and ϕ are not in good agreement. In this case, our algorithm will keep τ_k fixed and will start driving ϕ_k closer to ϕ. Eventually this will lead to a moment where $\widetilde{\rho}_k \geq \widetilde{\gamma}$, and then τ_k will be increased to $2\tau_k$. As above one may argue that in this case we have lost time, because we have brought ϕ_k closer to ϕ even though ϕ is too far away from f, only to notice in the end that we could not avoid increasing τ_k. This loss of time and energy could be avoided by adding the test (22). If $\sigma_k < \delta_2$, then we increase τ_k right away but keep incrementing ϕ_k. The convergence analysis of this is covered by the two central lemmas of this section.

8. Convergence analysis of the outer loop. All we have to do now is piece things together and show subsequence convergence of the sequence of serious steps x^j retained in the outer loop. We have the following.

THEOREM 6. *Let $f = \lambda_1 \circ F$ be a maximum eigenvalue function and let $x^1 \in \mathbb{R}^n$ be such that the set $\{x \in \mathbb{R}^n : f(x) \leq f(x^1)\}$ is compact. Then every accumulation point of the sequence x^j of serious steps generated by the algorithm is a critical point of f.*

Proof. (i) From the previous section we know that the inner loop always ends after a finite number of steps k with a new x^+ satisfying the acceptance test, unless we have finite termination due to $0 \in \partial f(x)$. Excluding this case, let us assume that x^j is the sequence of serious steps, satisfying the acceptance test in step 4 of the algorithm. Since y^{k+1} accepted in step 4 becomes the new x^{j+1}, that means

$$f(x^j) - f(x^{j+1}) \geq \gamma \left(f(x^j) - \phi_{k_j}(x^{j+1}; x^j) \right), \tag{23}$$

where j is the counter of the outer loop, k the counter of the inner loop, and where at the outer step j the inner loop was stopped at $k = k_j$. Now recall from the construction that $\tau_{k_j}\left(x^j - x^{j+1}\right) \in \partial\phi_{k_j}(x^{j+1}; x^j)$. The subgradient inequality for $\phi_{k_j}(\cdot; x^j)$ at x^{j+1} therefore gives

$$\tau_{k_j}\left(x^j - x^{j+1}\right)^\top (x^j - x^{j+1}) \leq \phi_{k_j}(x^j; x^j) - \phi_{k_j}(x^{j+1}; x^j) = f(x^j) - \phi_{k_j}(x^{j+1}; x^j),$$

using $\phi_{k_j}(x^j; x^j) = f(x^j)$. That means

$$\tau_{k_j}\|x^{j+1} - x^j\|^2 \leq f(x^j) - \phi_{k_j}(x^{j+1}; x^j) \leq \gamma^{-1}\left(f(x^j) - f(x^{j+1})\right)$$

using (23). Summing up from $j = 1$ to $j = J - 1$ gives

$$\sum_{j=1}^{J-1} \tau_{k_j}\|x^{j+1} - x^j\|^2 \leq \gamma^{-1} \sum_{j=1}^{J-1} f(x^j) - f(x^{j+1}) = \gamma^{-1}\left(f(x^1) - f(x^J)\right),$$

which is bounded above due to the hypothesis that $\{x \in \mathbb{R}^n : f(x) \leq f(x^1)\}$ is bounded. We deduce convergence of the series

$$\sum_{j=1}^{\infty} \tau_{k_j}\|x^{j+1} - x^j\|^2 < \infty.$$

In particular, $\tau_{k_j}\|x^{j+1} - x^j\|^2 \to 0$.

(ii) Let us prove that this implies $g_j := \tau_{k_j}\left(x^j - x^{j+1}\right) \to 0$, $(j \to \infty)$. Assume on the contrary that there exists an infinite subset $\mathcal{N}$ of $\mathbb{N}$ and some $\mu > 0$ such that $\|g_j\| = \tau_{k_j}\|x^j - x^{j+1}\| \geq \mu > 0$ for every $j \in \mathcal{N}$. In tandem with the summability of $\tau_{k_j}\|x^j - x^{j+1}\|^2$ shown in part (i) this could only mean $x^j - x^{j+1} \to 0$, and at the

same time, $\tau_{k_j} \to \infty$, $j \in \mathcal{N}$. We now argue that there exists yet another infinite subsequence $\mathcal{N}'$ of $\mathbb{N}$ with $\tau_j \to \infty$, $(j \in \mathcal{N}')$, such that for each $j \in \mathcal{N}'$, the doubling rule in step 5 of the algorithm was applied at least once before the step $x^{j+1} = y^{k_j+1}$ was accepted. Indeed, to construct $\mathcal{N}'$ we let, for every $j \in \mathcal{N}$, $j' \leq j$ be that outer-loop instant where the τ-parameter was increased for the last time before j, and we let $\mathcal{N}'$ consist of all these j', $j \in \mathcal{N}$. It is possible that $j' = j$, but in general we may have $j' < j$, and we know only that

$$2\tau_{j'-1} \leq \tau_{j'} \quad \text{and} \quad \tau_{j'} \geq \tau_{j'+1} \geq \cdots \geq \tau_j.$$

However, since $\tau_j \to \infty$, $j \in \mathcal{N}$, we know that we must have $\tau_{j'} \to \infty$, $j' \in \mathcal{N}'$. Since the doubling rule was applied at least once at the outer-loop counter j', $\mathcal{N}'$ is as claimed.

Let us say that for $j \in \mathcal{N}'$ the doubling rule was applied for the last time at stage $\tau_{k_j-\nu_j}$ for some $\nu_j \geq 1$. That means, $\tau_{k_j-\nu_j+1} = 2\tau_{k_j-\nu_j}$, while the τ-parameter remained unchanged during the following inner steps before acceptance:

$$\tau_{k_j} = \tau_{k_j-1} = \cdots = \tau_{k_j-\nu_j+1} = 2\tau_{k_j-\nu_j}. \tag{24}$$

Now recall that in step 5 of the algorithm we have $\rho_k < \gamma$ and $\widetilde{\rho}_k \geq \widetilde{\gamma}$ for those k where the trial step was not accepted and the doubling rule was applied. Since this is the case at stage $k_j - \nu_j$ we have

$$\rho_{k_j-\nu_j} = \frac{f(x^j) - f(y^{k_j-\nu_j+1})}{f(x^j) - \phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}; x^j)} < \gamma$$

and

$$\widetilde{\rho}_{k_j-\nu_j} = \frac{f(x^j) - \phi(y^{k_j-\nu_j+1}; x^j)}{f(x^j) - \phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}; x^j)} \geq \widetilde{\gamma}.$$

By (24) we now have

$$\frac{1}{2}\tau_{k_j}\left(x^j - y^{k_j-\nu_j+1}\right) \in \partial\phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}; x^j).$$

Using the subgradient inequality for $\phi_{k_j-\nu_j}(\cdot; x^j)$ at $y^{k_j-\nu_j+1}$ and $\phi_{k_j-\nu_j}(x^j; x^j) = f(x^j)$, we obtain

$$\begin{aligned}\frac{1}{2}\tau_{k_j}\left(x^j - y^{k_j-\nu_j+1}\right)^\top\left(x^j - y^{k_j-\nu_j+1}\right) &\leq \phi_{k_j-\nu_j}(x^j; x^j) - \phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}; x^j)\\ &= f(x^j) - \phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}; x^j),\end{aligned}$$

which could also be written as

$$\frac{\tau_{k_j}\|x^j - y^{k_j-\nu_j+1}\|^2}{f(x^j) - \phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}; x^j)} \leq 2. \tag{25}$$

Substituting (25) into the expression for $\widetilde{\rho}_{k_j-\nu_j}$ and expanding gives

$$\begin{aligned}\widetilde{\rho}_{k_j-\nu_j} &= \rho_{k_j-\nu_j} + \frac{f(y^{k_j-\nu_j+1}) - \phi(y^{k_j-\nu_j+1}; x^j)}{f(x^j) - \phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}; x^j)}\\ &\leq \rho_{k_j-\nu_j} + \frac{L\|x^j - y^{k_j-\nu_j+1}\|^2}{f(x^j) - \phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}; x^j)} \qquad \text{(using Proposition 1)}\\ &\leq \rho_{k_j-\nu_j} + \frac{2L}{\tau_{k_j}} \qquad \text{(using (25))}.\end{aligned}$$

Here Proposition 1 is applied to the set B of all x^j and $y^{k_j-\nu_j+1}$, $j \in \mathcal{N}'$, which is bounded because $\|y^{k_j-\nu_j+1}\| \le \|F'(x^j)^\star\|$ due to (7), and because the serious steps x^j belong to the level set $\{x \in \mathbb{R}^n : f(x) \le f(x^1)\}$, which is bounded by hypothesis. Since $\rho_{k_j-\nu_j} < \gamma$ and $L/2\tau_{k_j} \to 0$, we have $\limsup_{j\to\infty} \widetilde{\rho}_{k_j-\nu_j} \le \gamma$ in the estimate above, contradicting $\widetilde{\rho}_{k_j-\nu_j} \ge \widetilde{\gamma} > \gamma$ for all $j \in \mathcal{N}'$.

(iii) Having shown that $g_j := \tau_{k_j}(x^j - x^{j+1}) \to 0$, $(j \to \infty)$, let us argue that every accumulation point $\bar{x}$ of the sequence x^j of serious steps must be a critical point. Notice again that since $\{x \in \mathbb{R}^n : f(x) \le f(x^1)\}$ is compact by hypothesis, and since our algorithm is of descent type in the serious steps, the sequence x^j is bounded. Select a convergent subsequence $x^j \to \bar{x}$, $j \in \mathcal{N}$. The same argument applies to the sequence x^{j+1}, $j \in \mathcal{N}$. We may therefore assume that this sequence also has a limit, $\tilde{x}$. Notice that in general we might have $\tilde{x} \ne \bar{x}$. Only in those cases where the τ_{k_j}, $j \in \mathcal{N}$, are bounded away from 0 can we conclude that $x^{j+1} - x^j \to 0$, $j \in \mathcal{N}$. In general, however, according to step 4 of the algorithm, the τ-parameter may very well shrink to 0, and here $x^j - x^{j+1} \to 0$ cannot be assured.

Since g_j is a subgradient of $\phi_{k_j}(\cdot; x^j)$ at $x^{j+1} = y^{k_j+1}$, we have

$$\begin{aligned} g_j^\top h &\le \phi_{k_j}(x^{j+1}+h; x^j) - \phi_{k_j}(x^{j+1}; x^j) \\ &\le \phi(x^{j+1}+h; x^j) - \phi_{k_j}(x^{j+1}; x^j) \qquad (\text{using } \phi_{k_j} \le \phi) \end{aligned}$$

for every test vector h. Now we use the fact that $y^{k_j+1} = x^{j+1}$ was accepted in step 4 of the algorithm, which means

$$\gamma^{-1}\left(f(x^j) - f(x^{j+1})\right) \ge f(x^j) - \phi_{k_j}(x^{j+1}; x^j).$$

Combining these two estimates gives

$$\begin{aligned} g_j^\top h &\le \phi(x^{j+1}+h; x^j) - f(x^j) + f(x^j) - \phi_{k_j}(x^{j+1}; x^j) \\ &\le \phi(x^{j+1}+h; x^j) - f(x^j) + \gamma^{-1}\left(f(x^j) - f(x^{j+1})\right). \end{aligned}$$

Passing to the limit $j \in \mathcal{N}$ and using, in the order named, $g_j \to 0$, $x^{j+1} \to \tilde{x}$, $x^j \to \bar{x}$, and $f(\bar{x}) = \phi(\bar{x}; \bar{x})$, implies

$$0 \le \phi(\tilde{x}+h; \bar{x}) - \phi(\bar{x}; \bar{x})$$

for every test vector h, where the last term $f(x^j) - f(x^{j+1}) \to 0$ by monotonicity. Choosing $h = \bar{x} - \tilde{x} + h'$ therefore implies

$$0 \le \phi(\bar{x}+h'; \bar{x}) - \phi(\bar{x}; \bar{x})$$

for every test vector $h' \in \mathbb{R}^n$, which proves $0 \in \partial\phi(\bar{x}; \bar{x})$. Hence also $0 \in \partial f(\bar{x})$. ∎

In practical tests we observe convergence, and the theoretical possibility of a sequence of iterates with several accumulation points never occurs. This is explained to some extent by the following.

COROLLARY 7. *Suppose $f = \lambda_1 \circ F$ is convex in a closed and bounded neighborhood Ω of x^* and that the iterates (serious steps) x^j remain in Ω. Then the sequence x^j converges to some local minimum $x^\sharp \in \Omega$ with $f(x^*) = f(x^\sharp)$.*

Proof. As the sequence of serious steps x^j satisfies the acceptance condition in step 4 of the algorithm, we are in a situation similar to the one in the convex algorithm

discussed in [25]. The argument presented there can be used and shows convergence to a local minimum $x^\sharp \in \Omega$. □

Remark. The present trust region method should be compared to the approach of Fuduli, Gaudioso, and Giallombardo [20, 21] for general nonsmooth and nonconvex locally Lipschitz functions, where the authors design a trust region with the help of the first order affine approximations $a(y) = g^\top(y - y^{k+1}) + f(y^{k+1})$, $g \in \partial f(y^{k+1})$ of the objective f at the trial points y^{k+1}. As these affine models are not support functions to the objective, the authors classify them according to whether $a(x) > f(x)$ or $a(x) \le f(x)$, using this information to devise a trust region around the current x. Their approach is certainly appealing, because it uses genuine information from the objective f. In contrast, our method uses information from the model $\phi(\cdot; x)$ at the trial points y^{k+1}.

9. Minimizing the H_∞-norm. In this section we extend our algorithm to a larger class of functions which are suprema of an infinite family of maximum eigenvalue functions. The application we have primarily in mind is the H_∞-norm, but the results are applicable to a much larger class.

To introduce our case, we consider a parametrized family of stable linear time-invariant dynamical systems

$$P(\theta) : \begin{cases} \dot{x} = A(\theta)x + B(\theta)w, \\ z = C(\theta)x + D(\theta)w \end{cases} \tag{26}$$

with data $A(\theta) \in \mathbb{R}^{n_x \times n_x}$, $B(\theta) \in \mathbb{R}^{n_w \times n_x}$, $C(\theta) \in \mathbb{R}^{n_x \times n_z}$, $D(\theta) \in \mathbb{R}^{n_w \times n_z}$ depending smoothly on a decision parameter $\theta \in \mathbb{R}^n$. The transfer function of $P(\theta)$ is $G(\theta, s) = C(\theta)\left(sI - A(\theta)\right)^{-1} B(\theta) + D(\theta)$. Here n_x is the order of the system, $x(t)$ its state, n_w the number of inputs, $w(t)$ the input vector, n_z the number of outputs, and $z(t)$ the output vector. As a typical example, in feedback control synthesis, $P(\theta)$ may represent a closed-loop system, depending on the unknown (to be designed) feedback controller θ. The closed-loop transfer function then depends on the decision vector θ, which regroups the controller gains and possibly other decision parameters, e.g., from the open-loop system [44], or scalings/multipliers in robust synthesis [7].

Typically, the performance of the unknown feedback controller might be assessed in the H_∞-norm. Recall that the H_∞-norm $\|G(\theta, \cdot)\|_\infty$ of a stable system is the $L^2(j\mathbb{R}) \to L^2(j\mathbb{R})$ operator norm of the channel $w \to z$, where $z(s) = G(\theta, s)w(s)$. An explicit expression is

$$\|G(\theta, \cdot)\|_\infty = \sup_{\omega \in \mathbb{R} \cup \{\infty\}} \overline{\sigma}\left(G(\theta, j\omega)\right) = \sup_{\omega \in \mathbb{R} \cup \{\infty\}} \lambda_1\left(G(\theta, j\omega)^H G(\theta, j\omega)\right)^{1/2},$$

where X^H is the conjugate transpose of a matrix X. We are interested in that choice of θ which minimizes the H_∞-norm,

$$\min_{\theta \in \mathbb{R}^n} \|G(\theta, \cdot)\|_\infty. \tag{27}$$

We introduce the function

$$f(\theta) = \|G(\theta, \cdot)\|_\infty^2,$$

which is then an infinite maximum of maximum eigenvalue functions

$$f(\theta) = \max_{\omega \in \mathbb{R} \cup \{\infty\}} f(\theta, \omega), \qquad f(\theta, \omega) = \lambda_1\left(F(\theta, \omega)\right),$$

where

$$F(\theta,\omega) = G(\theta,j\omega)^H G(\theta,j\omega) \in \mathbb{S}^m.$$

Program (27) is semi-infinite with two sources of nonsmoothness: the infinite maximum operator and the nonsmoothness of each maximum eigenvalue function $f(\cdot,\omega)$.

Yet another difficulty arises in (27). Namely, given the fact that the H_∞-norm is defined only for stable transfer functions, the objective function $f(\theta)$ is defined only on the set $\mathcal{S}$ of those parameters θ where $G(\theta,\cdot)$ is stable. In other words, program (27) has the hidden constraint $\theta \in \mathcal{S}$. But $\mathcal{S}$ is an open set, because $G(\theta,\cdot)$ depends continuously on θ, so $\theta \in \mathcal{S}$ is not a constraint in the usual sense of mathematical programming. The following known fact is therefore useful.

LEMMA 8. *Suppose $(A(\theta),B(\theta),C(\theta),D(\theta))$ is observable and controllable for every $\theta \in \mathcal{S}$. Then $\|G(\theta,\cdot)\|_\infty \to +\infty$ for $\theta \in \mathcal{S}$ and $\theta \to \bar{\theta} \in \partial\mathcal{S}$. In other words, $f(\theta) = \|G(\theta,\cdot)\|_\infty^2$ behaves like a barrier function as θ approaches the boundary $\partial\mathcal{S}$ of the hidden constraint $\mathcal{S}$.*

The following result is yet another key property for the analysis of f; see, e.g., [11], [10, Lemma 1] for a proof.

LEMMA 9. *Suppose $G(\theta)$ is stable, i.e., $\theta \in \mathcal{S}$. Then the set $\Omega(\theta) = \{\omega \in \mathbb{R}\cup\{\infty\} : f(\theta) = f(\theta,\omega)\}$ of active frequencies is either finite or $\Omega(\theta) = \mathbb{R}\cup\{\infty\}$, i.e., $f(\theta) = f(\theta,\omega)$ for every $\omega \in \mathbb{R}\cup\{\infty\}$.*

We refer to $\Omega(\theta)$ as the set of active frequencies at θ. A system where $\Omega(\theta) = \mathbb{R}\cup\{\infty\}$ is called *all-pass*. In practical cases, iterates θ where $G(\theta,\cdot)$ is all-pass are rarely encountered.

For the following we switch back to the more standard notation in optimization, where the decision variable θ is denoted by $x \in \mathbb{R}^n$. Let x be our current iterate and consider the case where $\Omega(x) = \{\omega_1,\ldots,\omega_p\}$ is finite. Any Ω with $\Omega(x) \subset \Omega \subset \mathbb{R}\cup\{\infty\}$ is called an extension of $\Omega(x)$. For a given extension Ω we consider the function $f_\Omega(y) = \max_{\omega\in\Omega} f(y,\omega)$. If Ω is finite, then f_Ω is a maximum eigenvalue function, namely, $f_\Omega(y) = \lambda_1(F_\Omega(y))$, where $F_\Omega(y)$ is block diagonal with diagonal blocks $F(y,\omega)$, $\omega \in \Omega$ arranged in any convenient order. We have $f_\Omega \le f$ and $f_{\Omega(x)}(x) = f_\Omega(x) = f(x)$ for every extension Ω of $\Omega(x)$. The subdifferential of f at x is determined by $\Omega(x)$ in as much as

$$\partial f(x) = \partial f_\Omega(x) = \partial f_{\Omega(x)}(x).$$

Our goal is to extend the eigenvalue optimization algorithm to the case of the H_∞-norm. We use the following simple idea:

i. For a finite extension Ω of $\Omega(x)$ we know how to generate descent steps for f_Ω at x, because f_Ω is a maximum eigenvalue function.
ii. Suppose y^{k+1} is a serious step for f_Ω satisfying the acceptance test in step 4 of the algorithm. If Ω is large enough, f_Ω is close to f, so that we may hope that the acceptance test will also be satisfied for f.

This leads to a convergent algorithm for the H_∞-norm. What is needed is an increasing sequence $\Omega^1 \subset \Omega^2 \subset \cdots$ of finite sets whose union is dense in $\mathbb{R}_\infty$. Then we use the scheme of our algorithm to generate descent steps for f_{Ω_ℓ}, where $\Omega(x)\cup\Omega^\ell \subset \Omega_\ell$. If the approximation of f by f_{Ω_ℓ} is not good enough, we replace Ω_ℓ by the larger $\Omega_{\ell+1}$, where $\Omega(x)\cup\Omega^{\ell+1} \subset \Omega_{\ell+1}$, etc. This approach is inspired by the theory of consistent approximations of [53].

Spectral bundle algorithm for program (27).

Parameters $0 < \gamma^\sharp < \gamma < \frac{1}{2}$.	
0.	**Initialize outer loop.** Choose initial x such that $f(x) < \infty$.
1.	**Outer loop.** If $0 \in \partial f(x)$ at current x stop, else goto inner loop.
2.	**Initialize inner loop.** Let $x_1 = x$ and choose finite Ω_1 containing $\Omega(x_1)$. Put inner loop counter $\ell = 1$.
3.	**Subprogram.** At inner loop counter ℓ and current Ω_ℓ, f_{Ω_ℓ}, $\phi_{\Omega_\ell}(\cdot;x)$, and $\phi^k_{\Omega_\ell}(\cdot;x)$ use inner loop of the first algorithm (with counter k) to generate trial step x_ℓ satisfying the test $\dfrac{f(x)-f_{\Omega_\ell}(x_\ell)}{f(x)-\phi^k_{\Omega_\ell}(x_\ell;x)} \geq \gamma.$
4.	**Reality check.** Test whether $\dfrac{f(x)-f(x_\ell)}{f(x)-\phi^k_{\Omega_\ell}(x_\ell;x)} \geq \gamma^\sharp.$
5.	**Decision.** If this is the case, let $x^+ = x_\ell$ and go back to step 1. Otherwise add new frequencies to the set Ω_ℓ to obtain $\Omega_{\ell+1}$ and go back to step 3.

Here $\phi_{\Omega_\ell}(\cdot;x)$ relates to f_{Ω_ℓ} as $\phi(\cdot;x)$ relates to f in the algorithm of section 6, and, similarly, $\phi^k_{\Omega_\ell}(\cdot;x)$ used here plays the role of $\phi_k(\cdot;x)$ there.

Remarks. (i) Notice that our algorithm now has three iterative levels: the outer loop generating the serious iterates $x, x^+, x^{++}, \ldots$; the inner loop with counter ℓ, which corresponds in fact to the outer loop in the first algorithm, now applied to the function f_{Ω_ℓ}; and the innermost loop, which corresponds to the inner loop in the first algorithm, and which has its own counter k.

(ii) Notice that Ω_ℓ could in principle be *any* increasing sequence of finite sets of frequencies whose union is dense, but it is preferable to adapt this sequence to the local situation at the current iterate x. Ideas of how $\Omega_\ell(x)$ could be chosen at each step are discussed in [4].

THEOREM 10. *Suppose $(A(x), B(x), C(x), D(x))$ is observable and controllable for every $x \in \mathcal{S}$. Let x^1 be a starting point such that $f(x^1) < \infty$ and such that $\{x \in \mathbb{R}^n : f(x) \leq f(x^1)\}$ is bounded. Suppose the approximating sequence Ω_k is such that $f_{\Omega_k} \to f$ uniformly on bounded sets as $k \to \infty$. Then every accumulation point of the sequence of iterates x^j generated by the above algorithm with starting point x^1 is a critical point of f.*

Proof. (i) Observe first that due to the barrier property of the objective, the boundedness of the initial level set, and the fact that our method is of descent type in the outer iterates, every accumulation point $\bar{x}$ of the sequence of serious iterates x^j is necessarily inside the stability region $\mathcal{S}$. This means criticality of $\bar{x}$ is still described by $0 \in \partial f(\bar{x})$. In other words, the hidden constraint $x \in \mathcal{S}$ can be disregarded in what follows.

(ii) Let x be the current iterate of the outer loop and consider the inner loop with function f_{Ω_ℓ} and its models ϕ_{Ω_ℓ} and $\phi^k_{\Omega_\ell}$ for a fixed set Ω_ℓ. Applying the lemmas of section 7 to the maximum eigenvalue function f_{Ω_ℓ} shows finite termination of step 3 of the semi-infinite algorithm at a suitable x_ℓ (that is, after a finite number of steps y^{k+1}, where k is the counter of the innermost loop). Notice here that Lemmas 4 and 5 do not use compactness of the level sets of the objective, which is good news, because the objective is f_{Ω_ℓ}, and we know nothing about compactness of the level sets of f_{Ω_ℓ}. Only compactness of the level set of f is assumed in the statement. Instead, what made that argument in section 7 work was the special structure (7) of the subgradients

of the maximum eigenvalue function, and this applies to each f_{Ω_ℓ}.

(iii) The trial iterate x_ℓ found in step 3 corresponds in fact to the latest y^{k+1} of the innermost loop in the terminology of section 7, and the test in step 3 is precisely the acceptance test in the first algorithm. But being built on f_{Ω_ℓ}, x_ℓ does not necessarily pass the reality check in step 4, so a restart with a larger $\Omega_{\ell+1}$ may be required. What we have to prove, then, is that after a finite number of such updates $\Omega_\ell \to \Omega_{\ell+1}$, the $x_\ell = y^{k+1}$ will pass the test in step 4 and become the new outer iterate x^+. This is where we have to use the fact that f_{Ω_ℓ} gets closer to f as Ω_ℓ increases. More precisely, exploiting again the special structure of the subgradients of the different λ_1 involved, and using that $F'(z,\omega)$ is uniformly bounded for z in a bounded set and $\omega \in \mathbb{R}\cup\{\infty\}$, we see that the sequence of trial steps x_ℓ is bounded. Since $f_{\Omega_\ell} \to f$ uniformly on bounded sets, we conclude using $\gamma^\sharp < \gamma$ that ultimately the test in step 3 is sharper than the test in step 4. That proves finiteness of the loop in ℓ.

(iv) Finally, relabeling the outer iterates $x, x^+, \ldots$ as x^j, we are back in the situation analyzed in section 8. Using compactness of $\{x \in \mathbb{R}^n : f(x) \le f(x^1)\}$, we can use the same argument, now involving the parameter $\gamma^\sharp$ from step 4 of the algorithm. This completes the argument. □

Remark. The theory of consistent approximations [53] allows us in principle to apply this method in a fairly general context. However, a difficulty arises in step 4 of the algorithm, where the reality check requires computing values $f(x_\ell)$. This is what makes the case of the H_∞-norm special, because here we have an efficient way to compute function values [11]. The same idea can be used to solve problems with integral quadratic constraints (IQCs); see [7, 3].

10. Numerical tests. We present numerical tests with BMIs arising in feedback controller synthesis. Consider a closed-loop system of the form (26), where $A(K)$, $B(K), \ldots$ depend on the feedback controller K to be designed. The bounded real lemma [12] asserts that the closed-loop transfer channel $w \to z$ has H_∞-norm bounded by γ_∞ if and only if there exists a Lyapunov matrix $X \succ 0$ such that

$$\widetilde{\mathcal{B}}(K,X,\gamma_\infty) = \begin{bmatrix} A(K)^\top X + XA(K) & XB(K) & C(K)^\top \\ B(K)^\top X & -\gamma_\infty I & D(K)^\top \\ C(K) & D(K) & -\gamma_\infty I \end{bmatrix} \prec 0.$$

Fixing a small threshold $\epsilon > 0$, we consider the following nonlinear semidefinite program:

$$\begin{aligned} \text{minimize} \quad & \gamma_\infty \\ \text{subject to} \quad & \widetilde{\mathcal{B}}(K,X,\gamma_\infty) \preceq -\epsilon I, \\ & \epsilon I - X \preceq 0, \end{aligned}$$

which may be solved as an eigenvalue optimization program with decision variable $x = (\mathbf{vec}(K), \mathbf{svec}(X), \gamma_\infty)$ if exact penalization is used. To this end, put $\mathcal{B}(K,X,\gamma_\infty) := \mathrm{diag}[\epsilon I - X; \widetilde{B}(K,X,\gamma_\infty)]$ and fix a penalty parameter $\alpha > 0$ to solve the eigenvalue program $\min_x \lambda_1\left(\gamma_\infty I + \alpha\mathcal{B}(K,X,\gamma_\infty)\right)$. An alternative approach is to fix the performance level γ_∞ and to solve the eigenvalue program

$$\min_{K,X} \lambda_1\left(\mathcal{B}(K,X,\gamma_\infty)\right) \tag{28}$$

until a value < 0 is found. The gain γ_∞ could then be updated a few times to improve performance. This approach has been adopted in our numerical tests, while the exact

penalty approach was used in [8]. Further testing of this approach for control problems with IQCs is presented in [7].

10.1. Numerical implementation. We have performed six numerical experiments using models known in the control literature (VTOL helicopter, chemical reactor, transport airplane, piezoelectric actuator, coupled springs model, and binary distillation tower). This allows comparison with previous studies. We present both static and reduced order controller designs. The state space matrices of these models can be found in [35, 8], with the results of H_2 and H_∞ synthesis problems.

To solve the nonconvex eigenvalue optimization problem (28), we use our MATLAB implementation of the spectral bundle algorithm. The tangent subproblem to compute the trial step y^+ requires minimizing a quadratic cost function subject to an SDP constraint (an LMI (linear matrix inequality)). In order to solve the tangent subproblem efficiently, our specSDP routine [8] was used.

10.1.1. Initialization of the algorithm. The parameter values of the spectral bundle algorithm in section 6 have been set to $\gamma = 0.01$, $\tilde{\gamma} = 0.4$, and $\Gamma = 0.6$. We use $\text{tol} = 10^{-5}$ as a tolerance parameter to stop the algorithm as soon as progress in function values is minor, that is, $f(x) - f(x^+) < 10^{-5}(|f(x)| + 1)$.

Initialization of the variables X and K in program (28) is a difficult task. Indeed, the cost function (28) is nonconvex and the behavior of the algorithm could dramatically change for a bad choice of X and K. For instance, simple initializations such as $K = 0$ and $X = I$ are bound to fail. We have decided to start with a closed-loop stabilizing K, which is easily obtained via minimization of the spectral abscissa $\alpha(K) = \max \text{Re}\Lambda(A(K))$, where $\Lambda(A(K))$ is the spectrum of $A(K)$; see [9]. Once the initial K^0 is fixed in (28), minimizing the cost function with respect to X alone is a convex program, which can be solved using standard LMI techniques. A possible way to initialize X is therefore to choose X^0 optimal with respect to K^0. Unfortunately, this often leads to numerical problems since X^0 can be ill-conditioned and have an exceedingly large norm. We have observed during our numerical testing that the algorithm may crash because of the difference of magnitude between X and K. To avoid these effects, we have sometimes used a scaled version of X^0 to obtain decision variables K^0, X^0 of the same order of magnitude.

Initializing γ_∞ is easier, because the standard full order H_∞ controller gives a lower bound.

A delicate point is the initialization and choice of the number $r_k \in \mathbb{N}$ defining the dimension of the set $\mathcal{G}_k$ used to define the local model ϕ_k at each sweep k. As there does not seem to be any theoretical way to set the value of r_k, we have adjusted it heuristically during the computations. Figure 1 compares, for the helicopter model, static choices $r_k = \text{const}$ and displays the behavior of the algorithm for $r_k \in \{1, 2, 3, 4\}$. The ratio $f_i(x_k)/f_4(x_k)$ is plotted for $i = 1, \ldots, 3$, and $k = 1, \ldots, 100$, where $f_i(x_k)$ is the value of the cost function for $r_k = i$, and for the kth step x_k of the algorithm. As we can see in this plot, after some iterations, the algorithm behaves best for $r_k = 4$, indicating that larger r_k should give better results. The results in [48] seem to indicate that r_k should be chosen in such a way that the gap between λ_{r_k} and λ_{r_k+1} is as large as possible, but our testing in [2, 4, 8] has not confirmed this. The situation is far from clear, and dynamic choices of r_k ought to be tested. The advantage of our present approach motivated by [25] over the line motivated by [48] is that convergence of the method no longer hinges on the choice of r_k, respectively, r_ϵ.

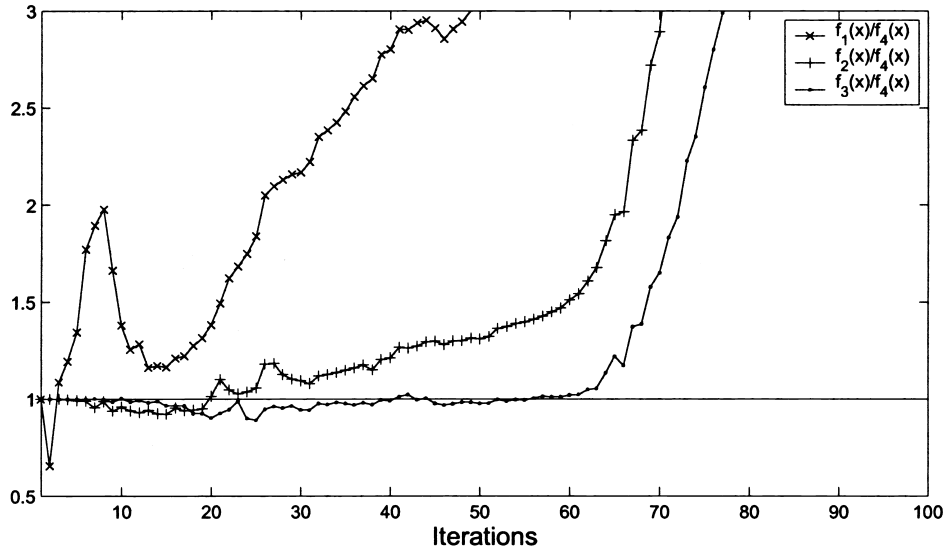


FIG. 1. *Behavior of the cost function during iterations for different values of r_k. This plot shows the ratio $f_i(x_k)/f_4(x_k)$, where $f_i(x_k)$ is the value of the cost function at step x_k for $r_k = i$.*

10.1.2. Nonsmooth optimality tests. To check whether the algorithm has reached a critical point, respectively, a minimum $x^* = (X^*, K^*)$, we have implemented two nonsmooth tests of optimality. The first one uses the ϵ-*enlarged subdifferential* of [48] for the maximum eigenvalue function, $\delta_\epsilon \lambda_1(X)$, to compute the criticality measure

$$\sigma_\epsilon = \operatorname{dist}\left(0, \delta_\epsilon f(x^*)\right),$$

where $f(x) = \lambda_1\left(\mathcal{B}(K, X, \gamma_\infty)\right) = \lambda_1(F(x))$. This parameter is the minimum value of the small size semidefinite program (computed using specSDP [8]):

$$\text{minimize } \left\{||F'(x^*)^\star G||_2 : \quad G \in \delta_\epsilon \lambda_1\left(\mathcal{B}(X^*, K^*, \gamma_\infty)\right)\right\},$$

where $\delta_\epsilon \lambda_1(X) = \left\{Q_\epsilon Y Q_\epsilon^\top \ : \ Y \succeq 0, \operatorname{Tr}(Y) = 1\right\}$. Here Q_ϵ is an $m \times r_\epsilon$ matrix whose r_ϵ columns form an orthonormal basis of eigenvectors associated with those eigenvalues $\lambda_i(X)$ satisfying

$$i \in I_\epsilon := \left\{i \ : \ \lambda_i(X) > \lambda_1(X) - \epsilon\right\}.$$

The number $r_\epsilon := \max\{i \ : \ i \in I_\epsilon\}$ is called the ϵ-*multiplicity* of $\lambda_1(X)$. In [48] it is shown that $\partial\lambda_1(X) \subset \delta_\epsilon \lambda_1(X) \subset \partial_\epsilon \lambda_1(X)$, so that ϵ could be roughly interpreted in the following way: If $0 \in \delta_\epsilon f(x^*) = F'(x^*)^\star \delta_\epsilon \lambda_1\left(\mathcal{B}(X^*, K^*, \gamma_\infty)\right)$, then it is not possible to further decrease f locally around x^* by more than ϵ. See [45, Lemma 2] for more details on this optimality test. Notice that this test may indeed be used as a stopping test in step 1 of the algorithm.

Our second optimality test is heuristic and is designed for a posteriori testing of criticality. It uses random perturbations x of $x^* = (X^*, K^*)$ to see whether the cost function value can be further decreased locally. Denoting by n_v the number of real optimization variables, we generate $100n_v$ random perturbations around $x^* = (X^*, K^*)$. The cost function values $f(X_i, K_i)$ for each perturbation $i = 1, \ldots, 100n_v$

are used to define

$$m_f := \min_i f(X_i, K_i),\ M_f := \max_i f(X_i, K_i),$$

and

$$p_f = \frac{\#\{i : f(X_i, K_i) < f(X^*, K^*)\}}{100 n_v}.$$

Parameter p_f (in percent) gives the proportion of perturbations for which the cost function value is improved.

10.2. Numerical results for H_∞-synthesis. Tables 1 and 2 present results obtained with the spectral bundle algorithm applied to the H_∞-synthesis problem. Table 1 is for static controllers, and Table 2 is for dynamic controllers. Comparison of performance γ_∞ with results in the literature, for static controllers $n_k = 0$, is given in Table 3. Table 4 gives some measured CPU times for three models on the static synthesis case. All numerical experiments have been performed on a Linux computer with a 2Ghz processor.

10.2.1. VTOL helicopter. State space data for the VTOL helicopter model are from [35, 8]; the model is described in [30]. The H_∞ gain was fixed at $\gamma_\infty = 0.1542$, the optimization variables were initialized as $K = [1, 1]^\top$, and $X = I$. The algorithm successfully solved the problem and obtained the H_∞ controller

$$K_\infty = \begin{bmatrix} 14.06432 \\ 239.5975 \end{bmatrix}.$$

We decided to look for a dynamic controller of order $n_k = 2$ with prescribed closed-loop performance $\gamma_\infty = 0.133$. The algorithm was initialized with a closed-loop stabilizing K^0 and $X^0 = I$. The optimal Lyapunov matrix X with respect to the given K^0 was used neither in the static nor in the dynamic case, because it has a very high norm and is likely to introduce numerical problems. The algorithm successfully computed

$$A_K = \begin{bmatrix} 1.672546 & 1.851477 \\ 1.849434 & 1.670218 \end{bmatrix},\ B_K = \begin{bmatrix} 73.76900 \\ 73.68110 \end{bmatrix},$$

$$C_K = \begin{bmatrix} 1.309171 & 1.308932 \\ 3.245753 & 3.241668 \end{bmatrix} \text{e-2},\ D_K = \begin{bmatrix} 0.4300008 \\ 0.7486698 \end{bmatrix}.$$

10.2.2. Chemical reactor. The chemical reactor model and numerical data can be found in [29]. We fixed the performance level $\gamma_\infty = 1.1830$ and initialized our algorithm with a closed-loop stabilizing K^0 together with the associated optimal X^0 (scaled for numerical convenience). This kind of initialization was used for both static and dynamic cases. The obtained static controller is

$$K_\infty = \begin{bmatrix} -3.791707 & -9.704666 \\ -7.166853 & -35.27994 \end{bmatrix}.$$

We also computed a dynamic controller of order 2. For $\gamma_\infty = 1.1420$, we obtained

$$A_K = \begin{bmatrix} -2.197969 & -0.341903 \\ -0.334860 & 2.205355 \end{bmatrix},\ B_K = \begin{bmatrix} 0.4983757 & 1.096069 \\ -6.452330 & -13.82350 \end{bmatrix},$$

$$C_K = \begin{bmatrix} -0.1090918 & 1.556743 \\ -0.2769249 & 3.964105 \end{bmatrix},\ D_K = \begin{bmatrix} -3.637238 & -6.382226 \\ -3.506708 & -19.39329 \end{bmatrix}.$$

The criticality measure was quite low in the static and the dynamic case, with $\sigma_\epsilon = 2.0041\text{e-}4$, respectively, $\sigma_\epsilon = 9.7570\text{e-}4$. In the static case, for the purpose of testing, we run the algorithm with a more severe stopping criterion to see if criticality decreased further. The stopping criteria were

$$f(x) - f(x^+) < 10^{-5}(|f(x)| + 1) \text{ and } ||x - x^+|| < 10^{-5}(||x|| + 1). \tag{29}$$

With this rule the algorithm stops after 26448 iterations. The final point verifies $\lambda_1(\widetilde{\mathcal{B}}) = -0.0079$ and criticality measure $\sigma_\epsilon = 4.2865\text{e-}06$, with $r_\epsilon = 4$ and $\epsilon = 1.3813\text{e-}5$. At this numerical precision, we can consider that the algorithm has reached a critical point.

10.2.3. Transport airplane. Model and state space data for the transport airplane are from [22]. We used a closed-loop stabilizing K^0 and the associated optimal X^0 for initialization. For $\gamma_\infty = 3.1770$, our algorithm computed the static controller

$$K_\infty = \begin{bmatrix} 0.6340988 & -0.5964908 & -0.7923650 & 5.166775e-2 & 1.055142 \end{bmatrix}.$$

In the static case we have also made a test of the algorithm with the stopping criterion (29). We have observed that the criticality of the final point of the algorithm has decreased: $\sigma_0 = 2.4819\text{e-}5$ after 241 iterations. Again, we can consider that the algorithm has reached a critical point with regard to the chosen numerical precision.

We failed to find a dynamic controller of order 2 for the airplane model. We computed a dynamic controller of order 1 with performance $\gamma_\infty = 2.860$. The H_∞ controller is

$$A_K = \begin{bmatrix} -0.4589498 \end{bmatrix},$$

$$B_K = \begin{bmatrix} -1.133331 & 1.441023 & 1.107071 & -0.116483 & -1.873279 \end{bmatrix}\text{e-}2,$$

$$C_K = \begin{bmatrix} -7.108071\text{e-}3 \end{bmatrix},$$

$$D_K = \begin{bmatrix} 1.740566 & -0.8878559 & -0.9477933 & 0.1000800 & 2.542508 \end{bmatrix}.$$

10.2.4. Piezoelectric actuator. The model of the piezoelectric actuator can be found in [13]. This study turned out to be one of the most difficult. As can be seen in Tables 1 and 2, the spectral bundle algorithm at first failed to solve the control problem both in the static and in the dynamic case. The algorithm converged to a couple (X^*, K^*) with slightly positive objective value around 1.45e-5 for the dynamical case, with the criticality parameter σ_ϵ quite small in both cases, indicating optimality. While it is perfectly possible that our algorithm, which is a local optimization method, may converge to a local minimum with positive values, failing to solve the underlying control problem, the present case turned out to be special. Namely, upon testing

the obtained controller K^*, we realized that it *is* closed-loop stabilizing and therefore solves the control problem. What happened is that the X^* computed by our algorithm was not suitable, as it fails to solve the Lyapunov inequality. Nonetheless, within the prescribed numerical precision both stopping tests indicated a local minimum.

The static controller for $\gamma_\infty = 0.0578$ reads

$$K_\infty = \left[\begin{array}{ccc} -0.3880673 & -1.837152 & -10.00377 \end{array} \right] \text{e+7},$$

and the dynamical controller of order 2, for $\gamma_\infty = 0.030$, is

$$A_K = \left[\begin{array}{cc} -6.770630 & -7.974432\text{e-}1 \\ -7.973471\text{e-}1 & -5.118056 \end{array} \right] \text{e+6},$$

$$B_K = \left[\begin{array}{ccc} 8.842279\text{e-}2 & 4.722429\text{e-}1 & 2.812368 \\ 3.574044\text{e-}2 & 1.910070\text{e-}1 & 1.136064 \end{array} \right] \text{e+7},$$

$$C_K = \left[\begin{array}{cc} -1.167814 & -4.720797\text{e-}1 \end{array} \right] \text{e+6},$$

$$D_K = \left[\begin{array}{ccc} -4.930334\text{e-}1 & -1.977292 & -1.480471\text{e+}1 \end{array} \right] \text{e+7}.$$

The described phenomenon indicates the numerical difficulty of synthesis problems with joint variable $x = (X, K)$, where important disparities between the numerical ranges of the two variables K and X may occur. In particular, the Lyapunov matrix X may be very ill-conditioned. The idea to keep K^* and compute a new Lyapunov variable X associated with K^* using a convex technique is systematically used in D-K iterations, where K and X variables are optimized alternatingly. The advantage of this approach is that both subproblems are then convex and can be solved by standard SDP solvers. However, intensive testing [27, 18] has shown that D-K techniques tend to get stalled and should in general be avoided. We believe that joint minimization in $x = (X, K)$ is the method of choice, despite the indicated difficulties. This does not exclude occasional restarts.

10.3. Coupled springs model. This is model CSE1 from [35] and consists of a string of coupled springs with dash-pots and masses. Input forces act on both the left and on the right ends of the spring system. The feedback controller has to stabilize the positions of the masses. We focus on the synthesis of a dynamic controller of order 4. To begin with, an initial closed-loop stabilizing controller was computed. Then X^0 was set to identity. An optimal controller was then synthesized for the performance level $\gamma_\infty = 0.0235$. The algorithm stopped at (K^*, X^*) with criticality measure sufficiently low in comparison with the numerical precision: $\sigma_\epsilon = 4.36e-6$, with $r_\epsilon = 5$ and $\epsilon = 3.64e-5$.

10.4. Distillation tower. Finally, to test the efficiency of the algorithm on a larger model, we used the BDT2 model from COMPL$_e$IB library [35], a binary distillation tower with 82 states, 4 outputs, and 4 controller inputs. The complete model is described in [56, section 12.4]. As can be seen in Table 1, the number of optimization variables $n_v = 3419$ is large in comparison with the previous examples. It should be highlighted that approximatively 99.5% (3403) of these variables are needed for the Lyapunov matrix X^*, but only 0.5% (16) are needed for controller K^*.

TABLE 1
Results of static H_∞-synthesis.

	γ_∞	α	$\lambda_n(X)$	$\lambda_1(\widetilde{\mathcal{B}})$	$\lambda_{n+n_y+n_u}(\widetilde{\mathcal{B}})$
Helicopter	0.1542	-0.1288	8.3236	-1.8146e-4	-1.8866e+5
Chemical reactor	1.1830	-1.8721	0.4327	-4.4606e-3	-2.4616e+3
Airplane	3.1770	-0.2486	1.1127e-3	-5.8368e-5	-44.4910
Piezo actuator	0.0578	-1.4202	7.0198e-10	8.2255e-3	-399.3432
BDT2	1.0722	-8.7144e-2	0.3223	-4.3202e-3	-6.2023e+2

#It.	$[n, n_y, n_u]$	n_v	r_k	σ_ϵ	r_ϵ	ϵ	p_f	m_f	M_f
600	[4,2,1]	12	4	1.15e-4	3	3.77e-7	0%	6.43e-4	1.95e-2
1800	[4,2,2]	14	4	2.01e-4	4	1.23e-5	0%	6.63e-3	9.28e-2
77	[9,1,5]	50	12	0.0114	4	3.24e-5	0%	4.00e-7	4.38e-5
1191	[5,1,3]	18	10	7.05e-3	5	7.94e-4	0%	570	2.29e+4
2934	[82,4,4]	3419	15	1.27e-3	6	6.20e-5	0.3%	-1.02e-5	8.87e-4

TABLE 2
Results of dynamic H_∞-synthesis.

	γ_∞	α	$\lambda_n(X)$	$\lambda_1(\widetilde{\mathcal{B}})$	$\lambda_{n+n_y+n_u}(\widetilde{\mathcal{B}})$
Helicopter	0.1334	-0.1583	2.7722e-3	-5.7903e-5	-39.5053
Chemical reactor	1.1420	-0.8245	0.2944	-3.2326e-3	-735.4831
Airplane	2.8600	-0.3161	6.2022e-4	-1.7815e-4	-150.1174
Piezo actuator	0.0300	-0.4796	-9.7169e-6	1.4503e-5	-16.1523
CSE1	0.0235	-2.1309e-1	8.2865e-1	-1.5574e-4	-6.6618e+2

#It.	$[n, n_k, n_y, n_u]$	n_v	r_k	σ_ϵ	r_ϵ	ϵ	p_f	m_f	M_f
9334	[6,2,4,3]	33	10	5.14e-4	5	1.11e-5	0%	4.01e-2	5.20
1379	[6,2,4,4]	37	6	9.75e-4	3	2.94e-4	1.81%	-1.81e-9	2.80e-8
1781	[10,1,2,6]	67	12	4.26e-2	3	4.50e-5	0%	2.11e-3	2.25e-2
8962	[6,2,3,5]	43	10	3.27e-2	5	1.45e-5	0%	1.15e-3	7.44e-2
2e4	[20,4,10,2]	384	6	4.34e-6	5	3.64e-5	0%	0.0025	6.51

TABLE 3
Comparison of γ_∞ with results in the literature for static controllers $n_k = 0$. The nonconvex spectral bundle method (NSBM) is shown on left.

	NSBM	[34]	[8]
Helicopter	0.1542	0.3455	0.157
Chemical reactor	1.1420	-	1.202
Airplane	3.1770	3.1774	2,220
Piezo actuator	0.0578	6.6256	3.055e-3

TABLE 4
Comparison of mean CPU times, in seconds, on three models of different sizes. Mean CPU times are given for computation of $f(x)$, computation of $F'(x)$, resolution of tangent program TP, serious step, and null step. %Serious gives the percentage of serious steps with respect to the total number of iterations.

Mean time	$f(x)$	$F'(x)$	TP	Serious	Null	%Serious
Helicopter	3.23e-04	8.80e-04	3.75e-02	3.81e-02	3.89e-02	35.2
Airplane	5.11e-04	1.46e-03	8.64e-02	8.80e-02	9.09e-02	66.8
BDT2	1.08e-02	1.12	4.45	5.51	4.66	44.5

Synthesis of a static controller K_∞ was obtained for the performance level $\gamma_\infty = 1.0722$. An initial stabilizing controller K^0 was computed with its corresponding optimal Lyapunov matrix X^0. The synthesized controller is

$$K_\infty := \begin{bmatrix} 5.748949e-1 & 1.751953 & 9.954549e-1 & 3.725248e-1 \\ -6.313297e-1 & 1.133587 & 9.815346e-1 & 2.909215 \\ 1.986992 & 1.789245 & 3.988785e-1 & 2.468048 \\ -1.061248e-1 & 5.597463e-1 & 3.635867 & 4.772583 \end{bmatrix}.$$

Criticality was fairly low compared to the numerical precision: $\sigma_\epsilon = 1.27e-3$ with $r_\epsilon = 6$ and $\epsilon = 6.20e-5$. However, the same numerical phenomenon as in the piezoelectric actuator example was observed: (X^*, K^*) was not a stationary point, and the cost could be further reduced by using a convex optimization technique to compute a new Lyapunov matrix X with K^* fixed.

Acknowledgment. The authors would like to thank the anonymous referees for their valuable comments and suggestions.

REFERENCES

[1] P. Apkarian and D. Noll, *Nonsmooth optimization for multidisk H_∞ synthesis*, European J. Control, 12 (2006), pp. 229–244.

[2] P. Apkarian and D. Noll, *Controller design via nonsmooth multidirectional search*, SIAM J. Control Optim., 44 (2006), pp. 1923–1949.

[3] P. Apkarian and D. Noll, *IQC analysis and synthesis via nonsmooth optimization*, Systems Control Lett., 55 (2006), pp. 971–981.

[4] P. Apkarian and D. Noll, *Nonsmooth H_∞ synthesis*, IEEE Trans. Automat. Control, 51 (2006), pp. 71–86.

[5] P. Apkarian and D. Noll, *Nonsmooth optimization for multiband frequency domain control design*, Automatica, 43 (2007), pp. 724–731.

[6] P. Apkarian and D. Noll, *Nonsmooth structured control design with applications to PID loopshaping of a process*, Internat. J. Robust Nonlinear Control, 17 (2007), pp. 1320–1342.

[7] P. Apkarian, D. Noll, and O. Prot, *Nonsmooth methods for analysis and synthesis with integral quadratic constraints*, submitted.

[8] P. Apkarian, D. Noll, J.-B. Thevenet, and H. D. Tuan, *A spectral quadratic-SDP method with applications to fixed-order H_2 and H_∞ synthesis*, Eur. J. Control, 10 (2004), pp. 527–538.

[9] V. Bompart, P. Apkarian, and D. Noll, *Non-smooth techniques for stabilizing linear systems*, in Proceedings of the 2007 American Control Conference (New York, NY), IEEE Press, Piscataway, NJ, 2007, pp. 1245–1250.

[10] V. Bompart, D. Noll, and P. Apkarian, *Second-order nonsmooth optimization of the H_∞ norm*, Numer. Math., 107 (2007), pp. 433–454.

[11] S. Boyd and V. Balakrishnan, *A regularity result for the singular values of a transfer matrix and a quadratically convergent algorithm for computing its L_∞-norm*, Systems Control Lett., 15 (1990), pp. 1–7.

[12] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM Stud. Appl. Math. 15, SIAM, Philadelphia, 1994.

[13] B. M. Chen, *H_∞ Control and its Applications*, Lecture Notes in Control and Inform. Sci. 235, Springer-Verlag, New York, Heidelberg, Berlin, 1998.

[14] F. Clarke, *Optimization and Nonsmooth Analysis*, Canadian Math. Society Series, John Wiley & Sons, New York, 1983.

[15] S. Cox and R. Lipton, *Extremal eigenvalue problems for two-phase conductors*, Arch. Ration. Mech. Anal., 136 (1996), pp. 101–117.

[16] J. Cullum, W. Donath, and P. Wolfe, *The minimization of certain nondifferentiable sums of eigenvalues of symmetric matrices*, Math. Programming Stud., 3 (1975), pp. 35–55.

[17] A. R. Diaaz and N. Kikudu, *Solution to shape and topology eigenvalue optimization problems using a homogenization method*, J. Numer. Methods Engrg., 35 (2005), pp. 1487–1502.

[18] B. Fares, D. Noll, and P. Apkarian, *Robust control via sequential semidefinite programming*, SIAM J. Control Optim., 40 (2002), pp. 1791–1820.

[19] R. Fletcher, *Semi-definite matrix constraints in optimization*, SIAM J. Control Optim., 23 (1985), pp. 493–513.

[20] A. Fuduli, M. Gaudioso, and G. Giallombardo, *A DC piecewise affine model and a bundling technique in nonconvex nonsmooth optimization*, Optim. Methods Softw., 19 (2004), pp. 89–102.

[21] A. Fuduli, M. Gaudioso, and G. Giallombardo, *Minimizing nonconvex nonsmooth functions via cutting planes and proximity control*, SIAM J. Optim., 14 (2004), pp. 743–756.

[22] D. Gangsaas, K. Bruce, J. Blight, and U.-L. Ly, *Applications of modern synthesis to aircraft control: Three case studies*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 995–1014.

[23] C. Helmberg and K. C. Kiwiel, *A spectral bundle method with bounds*, Math. Programming, 93 (2002), pp. 173–194.

[24] C. Helmberg and F. Oustry, *Bundle methods to minimize the maximum eigenvalue function*, in Handbook of Semidefinite Programming. Theory, Algorithms and Applications, Internat. Ser. Oper. Res. Management Sci. 27, L. Vandenberghe, R. Saigal, and H. Wolkowitz, eds., Kluwer, Boston, 2000, pp. 307–337.

[25] C. Helmberg and F. Rendl, *Spectral bundle method for semidefinite programming*, SIAM J. Optim., 10 (2000), pp. 673–696.

[26] C. Helmberg, F. Rendl, and R. Weismantel, *A semidefinite programming approach to the quadratic knapsack problem*, J. Comb. Optim., 4 (2000), pp. 197–215.

[27] J. W. Helton and O. Merino, *Coordinate optimization for bi-convex matrix inequalities*, in Proceedings of the 45th IEEE Conference on Decision and Control (San Diego, CA), IEEE Press, Piscataway, NJ, 1997, pp. 3609–3613.

[28] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms*, Vols. I, II, Grundlehren Math. Wiss. 305, 306, Springer-Verlag, New York, Heidelberg, Berlin, 1993.

[29] Y. S. Hung and A. G. J. MacFarlane, *Multivariable Feedback: A Quasi-classical Approach*, Lecture Notes in Control and Inform. Sci. 40, Springer-Verlag, New York, Heidelberg, Berlin, 1982.

[30] L. H. Keel, S. P. Bhattacharyya, and J. W. Howe, *Robust control with structured perturbations*, IEEE Trans Automat. Control, 36 (1988), pp. 68–77.

[31] K. C. Kiwiel, *Methods of Descent for Nondifferentiable Optimization*, Lecture Notes in Math. 1133, Springer-Verlag, Berlin, 1985.

[32] K. C. Kiwiel, *A linearization algorithm for computing control systems subject to singular value inequalities*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 595–602.

[33] K. C. Kiwiel, *Proximity control in bundle methods for convex nondifferentiable optimization*, Math. Programming, 46 (1990), pp. 105–122.

[34] F. Leibfritz, *Computational Design of Stabilizing Static Output Feedback Controllers*, Tech. Report 99-01, Universität Trier, Trier, Germany, 1999.

[35] F. Leibfritz, *$COMPL_eIB$, COnstrained Matrix-optimization Problem LIbrary—A Collection of Test Examples for Nonlinear Semidefinite Programs, Control System Design and Related Problems*, Tech. Report, Universität Trier, Trier, Germany, 2003.

[36] C. Lemaréchal, *An extension of Davidson's method to nondifferentiable problems*, in Nondifferentiable Optimization, Math. Programming Stud., M. L. Balinski and P. Wolfe, eds., North–Holland, Amsterdam, 1975, pp. 95–109.

[37] C. Lemaréchal, *Bundle methods in nonsmooth optimization*, in Nonsmooth Optimization, Proceedings of the IIASA Workshop (1977), C. Lemaréchal and R. Mifflin, eds., Pergamon, Oxford, Elmsford, NY, 1978, pp. 79–102.

[38] C. Lemaréchal, *Nondifferentiable optimization*, in Optimization, Handbooks Oper. Res. Management Sci. 1, North–Holland, Amsterdam, 1989, pp. 529-572.

[39] C. Lemaréchal, *Lagrangian relaxation*, in Computational Combinatorial Optimization, M. Junger and D. Naddef, eds., Springer, New York, 2001, pp. 112–156.

[40] C. Lemaréchal, A. Nemirovskii, and Y. Nesterov, *New variants of bundle methods*, Math. Programming, 69 (1995), pp. 111–147.

[41] C. Lemaréchal and F. Oustry, *Nonsmooth algorithms to solve semidefinite programs*, in Advances in Linear Matrix Inequality Methods in Control, L. El Ghaoui and S.-I. Niculescu, eds., SIAM, Philadalphia, 1999, pp. 57–77.

[42] D. Mayne and E. Polak, *Algorithms for the design of control systems subject to singular value inequalities*, Math. Programming Stud., 18 (1982), pp. 112–134.

[43] D. Mayne, E. Polak, and A. Sangiovanni, *Computer aided design via optimization*, Automatica, 18 (1982), pp. 147–154.

[44] K. Mombaur, *Stability Optimization of Open-Loop Controlled Walking Robots*, Ph.D. Thesis, Universität Heidelberg, Heidelberg, Germany, 2001.
[45] D. Noll and P. Apkarian, *Spectral bundle method for nonconvex maximum eigenvalue functions: First-order methods*, Math. Programming Ser. B, 104 (2005), pp. 701–727.
[46] D. Noll and P. Apkarian, *Spectral bundle method for nonconvex maximum eigenvalue functions: Second-order methods*, Math. Programming Ser. B, 104 (2005), pp. 729–747.
[47] D. Noll, M. Torki, and P. Apkarian, *Partially augmented Lagrangian method for matrix inequality constraints*, SIAM J. Optim., 15 (2004), pp. 161–184.
[48] F. Oustry, *A second-order bundle method to minimize the maximum eigenvalue function*, Math. Programming Ser. A, 89 (2000), pp. 1–33.
[49] M. L. Overton, *On minimizing the maximum eigenvalue of a symmetric matrix*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 256–268.
[50] M. L. Overton, *Large-scale optimization of eigenvalues*, SIAM J. Optim., 2 (1992), pp. 88–120.
[51] M. L. Overton and R. S. Womersley, *On the sum of the largest eigenvalues of a symmetric matrix*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 41–45.
[52] E. Polak, *On the mathematical foundations of nondifferentiable optimization in engineering design*, SIAM Rev., 29 (1987), pp. 21–89.
[53] E. Polak, *Optimization: Algorithms and Consistent Approximations*, Appl. Math. Sci. 124, Springer, New York, 1997.
[54] E. Polak and Y. Wardi, *A nondifferential optimization algorithm for the design of control systems subject to singular value inequalities over the frequency range*, Automatica, 18 (1982), pp. 267–283.
[55] F. Pukelsheim, *Optimal Design of Experiments*, John Wiley & Sons, New York, 1993.
[56] S. Skogestad and I. Postlethwaite, *Multivariate Feedback Control*, John Wiley & Sons, New York, 1996.
[57] J.-B. Thevenet, D. Noll, and P. Apkarian, *Nonlinear spectral SDP method for BMI-constrained problems: Applications to control design*, in Informatics in Control, Automation and Robotics I, J. Braz, H. Arajo, A. Viera, and B. Encarnaco, eds., Springer-Verlag, Berlin, 2006, pp. 61–72.
[58] L. Vandenberghe and S. Boyd, *Semidefinite programming*, SIAM Rev., 38 (1996), pp. 49–95.
[59] P. Wolfe, *A method of conjugate subgradients for minimizing nondifferentiable functions*, in Nondifferentiable Optimization, Math. Programming Stud. 3, M. L. Balinski and P. Wolfe, eds., North–Holland, Amsterdam, 1975, pp. 145–173.
[60] Z. Zhao, B. Braams, M. Fukuda, M. Overton, and J. Percus, *The reduced density matrix method for electronic structure calculations and the role of three-index representability conditions*, J. Chem. Phys., 120 (2004), pp. 2095–2104.

SIAM J. OPTIM.
Vol. 19, No. 1, pp. 307–325

A FIRST-ORDER CONVERGENCE ANALYSIS OF TRUST-REGION METHODS WITH INEXACT JACOBIANS*

ANDREA WALTHER†

Abstract. A class of trust-region sequential quadratic programming algorithms for the solution of minimization problems with nonlinear equality constraints is analyzed. The considered class of optimization methods does not require the exact evaluation of the constraint Jacobian in each optimization step but uses only an approximation of this first-order derivative information. Hence, the presented approach is especially well suited for equality constrained optimization problems where the Jacobian of the constraints is dense. The accuracy requirements for the presented first-order global convergence result are based on the feasibility and the optimality of the iterates. The corresponding criteria can be verified easily during the optimization process to adjust the approximation quality of the constraint Jacobian.

Key words. trust-region algorithms, inexact Jacobians, global convergence

AMS subject classifications. 90C30, 90C55, 49M05, 49M37

DOI. 10.1137/050634530

1. Introduction.

Trust-region successive quadratic programming (SQP) algorithms have been applied efficiently to solve a wide range of nonlinear optimization problems given by

$$\min_{x \in \mathbb{R}^N} f(x) \qquad \text{subject to } \; c(x) = 0, \tag{1}$$

where the objective $f : \mathbb{R}^N \to \mathbb{R}$ and the vector of the constraints $c : \mathbb{R}^N \to \mathbb{R}^M$ with $N \geq M$ are given smooth functions. For the majority of the trust-region SQP type algorithms, the computation of the next iterate requires the solution of a linear system of the form

$$A(x_k)A(x_k)^T v = b,$$

where

$$A(x) = (\nabla c_1(x), \ldots, \nabla c_M(x))^T \in \mathbb{R}^{M \times N}$$

is the exact matrix of the constraint gradients at x. Furthermore, a representation $Z(x)$ of the null space of $A(x)$ is needed frequently for the computation of the next step. For these reasons, the explicit forming and factoring of the constraint Jacobian $A(x)$ provides an efficient step calculation if $A(x)$ is sparse and well structured; see, e.g., [1]. As an alternative, one may use iterative system solves up to a certain accuracy, for example, Krylov subspace or multigrid methods, for the step calculation in each iteration; see, e.g., [17, 20, 29]. However, both approaches may result in very time-consuming computations, especially if the Jacobian of the constraints is dense or unstructured. Therefore, we present and analyze in this paper a class of trust-region SQP algorithms that does not require the exact evaluation of the constraint Jacobian or an iterative solution of a linear system with a system matrix that involves the con-

*Received by the editors June 27, 2005; accepted for publication (in revised form) August 14, 2007; published electronically March 28, 2008.
http://www.siam.org/journals/siopt/19-1/63453.html
†Institute of Scientific Computing, Technische Universität Dresden, 01062 Dresden, Germany (andrea.walther@tu-dresden.de).

straint Jacobian. Instead the proposed algorithm works only with an approximation of this first-order information. Hence, the algorithm presented here is well suited for optimization problems of moderate size but with a special structure of the constraint Jacobian. The corresponding applications cover the wide range of periodic adsorption processes including for example the purification of hydrogen. In these cases, the Jacobian of the equality constraints is dense due to the periodicity of the underlying chemical process. As a consequence, the run-time needed for the optimization process may be dominated significantly by the computation of the dense Jacobian and its factorization; see, e.g., [19]. For these optimization tasks and problems with a similar structure, the algorithm proposed in this paper may allow a considerable reduction of the computing time required to calculate a solution.

For numerous optimization problems, the considered system is described by ordinary or partial differential equations the discretization of which yields the equality constraints. Exploiting the direct sensitivity equation or the adjoint differential equation, one can evaluate products of the Jacobian $A(x)$ and a given vector v, i.e., $A(x)v$ and $A(x)^T v$. Related derivative information can be computed also by applying automatic differentiation [14]. Hence, it is reasonable to assume that one can evaluate exact products of the Jacobian multiplied from the right or from the left by a given vector. However, the computation of the complete Jacobian matrix $A(x)$ may be very time consuming, especially if $A(x)$ is dense or unstructured, since many Jacobian-vector products are required to build the full matrix $A(x)$ in these cases. Therefore, we present an algorithm that uses only Jacobian-vector and vector-Jacobian products but avoids the calculation and factorization of $A(x)$ in each optimization step or the iterative solve of a linear system involving $A(x)$ as part of the system matrix.

To solve the optimization problem (1), we follow the approach proposed by Byrd [2] and Omojokun [23]. For composite-step trust-region methods that employ exact information, a comprehensive treatment of the convergence properties can be found in [7]. Implementations of the Byrd–Omojokun trust-region method are used successfully to solve equality constrained nonlinear problems (NLPs) [1, 20]. Related implementations using augmented Lagrangian merit functions are proposed and analyzed in [9]. Extensions of this approach to a more general class of trust-region methods can be found in [8]. Box trust-region methods are analyzed in [13]. More recently, trust-region methods without penalty functions have been developed by Fletcher and others [10, 11, 12] as well as Ulbrich and Ulbrich [26].

The effects of inexact problem information on the global convergence of inexact SQP methods can be found, for example, in [18, 21, 27]. In a line search setting, the effects of inexact information on the global convergence are studied in [3]. For an inexact composite step trust-region SQP method a first proof of global convergence is given in [17], where the analysis is focused on inexactness arising from iterative system solves. Our analysis and assumptions on inexactness differ from [17] in the following way: we do not consider a splitting of the variables into state and control variables. Hence, we allow general unstructured approximations of the Jacobian $A(x)$ and the corresponding null space representation as well as inexactness due to iterative solves. The proofs of first-order convergence given in this paper are based on ideas presented in [4]. Since we concentrate our analysis on the effects of inexact Jacobian information, the present paper does not examine the performance of the algorithm in the presence of dependent constraint gradients. Therefore, we assume in contrast to [4] throughout that the constraint Jacobian has full rank. Furthermore, we do not incorporate inequality constraints as in [4], since the efficient handling of inequalities in the case of inexact constraint Jacobians is subject of further research.

This paper has the following structure. In section 2 we introduce the notation and the main assumptions that are used for the proof of global first-order convergence. Subsequently, we discuss our inexact trust-region SQP algorithm in section 3. The well-posedness of this algorithm will be shown in section 4. Section 5 contains the proof of global convergence to first-order critical points. Finally, some conclusions and possible extensions are presented in section 6.

2. Notation and assumptions. The Lagrangian of (1) is defined by

$$\mathcal{L}(x,\lambda) = f(x) + \lambda^T c(x).$$

Assuming that a suitable constraint qualification is fulfilled, the first-order optimality conditions yield for an optimal solution x_* of (1) that

$$\begin{aligned}\nabla_x \mathcal{L}(x_*,\lambda_*) &= \nabla f(x_*) + A(x_*)^T\lambda_* = 0,\\ \nabla_\lambda \mathcal{L}(x_*,\lambda_*) &= c(x_*) = 0\end{aligned}$$

holds for a certain Lagrange multiplier $\lambda_* \in \mathbb{R}^M$. To apply an SQP trust-region algorithm, we approximately solve in the kth iteration the quadratic program

$$\begin{aligned}&\min_{d\in\mathbb{R}^N} \nabla f(x_k)^T d + \frac{1}{2} d^T B_k d\\ &\text{subject to } A(x_k)d + c(x_k) = 0,\\ &\|d\| \le \Delta_k,\end{aligned} \tag{2}$$

to compute a new step d_k for a given iterate x_k, a given trust-region radius Δ_k, and Lagrange multipliers λ_k. Here and throughout, B_k may stand for the exact second-order information $\nabla^2_{xx}\mathcal{L}(x_k,\lambda_k)$. Then, the functions $f(\cdot)$ and $c(\cdot)$ have to be twice continuously differentiable. Alternatively, one may use a symmetric matrix approximating the Hessian $\nabla^2_{xx}\mathcal{L}(x_k,\lambda_k)$. Furthermore, $\|\cdot\|$ denotes the Euclidean norm $\|\cdot\|_2$.

Since problem (2) may have no feasible solution, relaxation strategies were studied; see, e.g., [5, 24, 25]. As an alternative to overcome this difficulty, one can use a composite-step method. Following the approach of Byrd and Omojokun, we define the merit function

$$\phi(x;\mu) = f(x) + \mu\|c(x)\|$$

with the penalty parameter $\mu > 0$ to judge the progress toward the solution. This merit function is exact but nondifferentiable due to the Euclidean norm in the second term. A model of $\phi(\cdot;\mu_k)$ around an iterate x_k is given by the function

$$m_k(d) = f(x_k) + \nabla f(x_k)^T d + \frac{1}{2} d^T B_k d + \mu_k\|c(x_k) + A(x_k)d\|.$$

For measuring the progress of our algorithm, we define for a given iterate x_k and a step d the actual reduction in the merit function as

$$\operatorname{ared}_k(d) = \phi(x_k;\mu_k) - \phi(x_k + d;\mu_k). \tag{3}$$

The predicted reduction in the merit function is defined as the change of the model m_k caused by a step d, i.e.,

$$\begin{aligned}\operatorname{pred}_k(d) &= m_k(0) - m_k(d)\\ &= -\nabla f(x_k)^T d - \frac{1}{2} d^T B_k d + \mu_k(\|c(x_k)\| - \|c(x_k) + A(x_k)d\|).\end{aligned} \tag{4}$$

We suppose that for each iteration k one can provide an approximation A_k of the exact Jacobian $A(x_k)$ and an approximation Z_k of an exact null space basis $Z(x_k)$ with $A(x_k)Z(x_k) = 0$ and $A_k Z_k = 0$. Hence, we refer to the exact matrix information as $A(x_k)$ and $Z(x_k)$ and to the corresponding approximation as A_k and Z_k, respectively. The approximation of the derivative matrices using quasi-Newton update formulas fits into this setting. For this purpose, one may employ the well-known symmetric rank one (SR1) update formula to approximate the Hessian $\nabla^2_{xx}\mathcal{L}(x_k, \lambda_k)$. This approach is examined for unconstrained optimization in [6]. The two-sided rank one (TR1) update formula as proposed in [15] can be used to approximate the constraint Jacobian. Another possibility would be to compute the required Hessian-vector products exactly employing, for example, automatic differentiation. For the first-order information, the exact information $A(x_k)$ and $Z(x_k)$ could be computed for the iterate k and used for the following iterates as long as the restrictions on the inexactness are fulfilled. This is a promising approach since the iterates converge frequently in a tangential way toward the optimal solution. This observation holds when the Hessian is approximated for example by a quasi-Newton formula and the exact Jacobian of the constraints is used. We observe the same behavior in our first numerical experiments for numerous test problems using the TR1 update to approximate the Jacobian. Therefore, the changes in the null space will be hopefully rather small at the end of the optimization procedure.

To prove the convergence results presented in this paper, we define $D \equiv N - M$ and make the following assumptions:

(AS1) $A(x_k)$ has full row rank for all iterates x_k with $\sigma_D(A(x_k)) \geq \hat{\sigma} > 0$, where $\sigma_D(A(x_k))$ denotes the smallest singular value of $A(x_k)$.

(AS2) A_k has full row rank for all iterations with $\sigma_D(A_k) \geq \tilde{\sigma} > 0$.

(AS3) $Z_k \in \mathbb{R}^{N\times D}$ has full column rank D for all iterates x_k with $\sigma_D(Z_k) \geq \check{\sigma} > 0$ and remains bounded.

(AS4) The sequence $\{f(x_k)\}$ is bounded below. The sequences $\{\nabla f(x_k)\}$, $\{c(x_k)\}$, $\{A(x_k)\}$, and $\{B_k\}$ are bounded.

(AS5) The functions $\nabla f(\cdot)$, $c(\cdot)$, and $A(\cdot)$ are Lipschitz continuous on an open convex set $\mathcal{X}$ containing all iterates.

(AS6) The gradients $\nabla f(x)$, $\nabla_x \mathcal{L}(x, \lambda)$, the gradient-vector product $\nabla f(x)^T d$ and the products $A(x)v$, $w^T A(x)$ can be evaluated exactly.

(AS7) For fixed x_k, the approximation Z_k can be improved in a finite number of steps such that an exact null space representation $Z(x_k)$ is obtained.

Assumption AS1 is needed to prove the feasibility of all limit points and to derive upper bounds for the normal steps in section 5. A similar assumption is made in [17, sec. 3.3] to prove first-order global convergence. In the paper [4], the upper bound for the normal step is derived using an assumption similar to AS1. Furthermore, the analysis in [4] explicitly studies the rank deficiency of the constraint Jacobian $A(x_k)$ and its influence on the overall algorithm. That is, the iterates could converge either to a feasible point or to a limit point failing the linear independence constraint qualification. Therefore, an assumption similar to AS1 is not made for this part of [4]. However, the present paper focuses mainly on the convergence of a trust-region algorithm with inexact Jacobian information. For that reason, we decided not to explore the possibility that $A(x_k)$ is rank deficient since this would complicate the analysis considerably. The convergence to a limit point not satisfying the linear independence constraint qualification may be the subject of future research.

In AS7, we assume that we can improve the approximation Z_k such that it represents an exact null space $Z(x_k)$ of $A(x_k)$ after a finite number of improvement steps.

This is possible, for example, for the TR1 approach by performing M rank one updates without changing the current iterate x_k since the TR1 update procedure yields the exact Jacobian $A(x_k)$ for fixed x_k after at most M updates. This can be verified in the following way. Starting with an approximation $\tilde{A}_0 = A_k$ one performs M TR1 updates of the form

$$\tilde{A}_{i+1} \equiv \tilde{A}_i + \frac{(y_i - \tilde{A}_i v_i)(\tau_i^T - w_i^T \tilde{A}_i)}{(\tau_i^T - w_i^T \tilde{A}_i) v_i}$$

with $y_i \equiv A(x_k)v_i$ and $\tau_i \equiv w_i^T A(x_k)$ for arbitrary linearly independent vectors v_i and w_i chosen such that $(\tau_i^T - w_i^T \tilde{A}_i)v_i \neq 0$ holds. Due to the heredity of the rank one update, one obtains after M updates that

$$w_i^T \tilde{A}_M = w_i^T A(x_k) \qquad \text{for all} \qquad i = 0, \ldots, M-1.$$

The proof of this identity is similar to the proof of a related result for the SR1 update and can be found in [28]. Since the w_i, $0 \leq i < M$, are M linearly independent vectors, it follows that $\tilde{A}_M = A(x_k)$. Using an equivalent update procedure for a factorized null space representation, one obtains an exact null space representation $Z_k = Z(x_k)$ after at most M updates [16]. If one freezes the Jacobian and null space information as proposed above, one can evaluate new exact Jacobian information if the restrictions on the inexactness are no longer valid. This approach ensures that assumption AS7 holds. Hence, one can use the approximation $Z_k = Z_{k-1}$ and improve the approximation of the null space if required.

All other assumptions are either standard assumptions required also for the global convergence analysis in other papers, i.e., AS4 and AS5, or motivated by the applications that we had in mind when designing the algorithm, i.e., AS2, AS3, and AS6.

3. A Jacobian-free trust-region method. To apply a composite step trust-region method as proposed by Byrd and Omojokun, we first compute a normal step n that lies well inside the trust-region radius and that attempts to satisfy the linear constraints in (2). Subsequently, we take a tangential step t toward optimality. Putting both steps together, we obtain the total step $d = n + t$.

3.1. The normal subproblem. For the current iterate x_k, we compute a normal step n_k that best satisfies the linearized constraints by solving the *normal subproblem*

$$\begin{aligned} &\min_{n \in \mathbb{R}^N} \ \|c(x_k) + A(x_k)n\|^2 \\ &\text{subject to } \|n\| \leq \tilde{\Delta}_k \end{aligned} \tag{5}$$

with $\tilde{\Delta}_k = \kappa \Delta_k$ and $\kappa \in (0, 1)$. This optimization problem may have infinitely many solutions. The exact Cauchy step for (5) is given by

$$n_k^C = -\alpha_k^C A(x_k)^T c(x_k), \tag{6}$$

where α_k^C is the optimal solution of the problem

$$\begin{aligned} &\min_{\alpha \geq 0} \ \|c(x_k) - \alpha A(x_k)A(x_k)^T c(x_k)\| \\ &\text{subject to } \|\alpha A(x_k)^T c(x_k)\| \leq \tilde{\Delta}_k. \end{aligned} \tag{7}$$

Hence, due to our assumption AS6 that we can evaluate $A(x_k)v$ and $A(x_k)^T w$ for given v and w exactly, we are able to compute the exact Cauchy step. Nevertheless, employing only the exact Cauchy step may yield very slow convergence [22]. To accelerate the optimization process, one could use in addition also the exact Newton step. This global minimizer of the unconstrained version of (5) is given by

$$n^N(x_k) = -A(x_k)^+ c(x_k) = -A(x_k)^T (A(x_k)A(x_k)^T)^{-1} c(x_k).$$

However, we do not want to compute the vector $(A(x_k)A(x_k)^T)^{-1}c(x_k)$ exactly. Alternatively, if one assumes that an approximation $(A_k A_k^T)^{-1}c(x_k)$ can be evaluated at low computational cost, for example, by maintaining a factorized approximation of $A(x_k)$ as described in [16], then one could use the approximation

$$n_k^N = -A(x_k)^T (A_k A_k^T)^{-1} c(x_k)$$

of the exact Newton step. In combination with the exact Cauchy step, then one may compute the inexact dogleg step of Powell by setting

$$n_k^D = \eta n_k^N + (1-\eta) n_k^C$$

with $\eta = 1$ if $\|n_k^N\| \leq \tilde{\Delta}_k$. Otherwise $\eta \in [0,1]$ would be adjusted such that the length of n_k^D is equal to $\tilde{\Delta}_k$.

For obtaining convergence, one has to analyze the reduction in the linearized constraints caused by the normal step. For that purpose, we define the *normal predicted reduction* for a vector n as

$$\text{npred}_k(n) = \|c(x_k)\| - \|c(x_k) + A(x_k)n\| \tag{8}$$

and require that the normal step n_k computed in the kth iteration satisfies the following condition.

Normal Cauchy decrease condition. *An approximate solution n_k of the normal subproblem* (5) *must satisfy*

$$\text{npred}_k(n_k) \geq \gamma_n \text{npred}_k(n_k^C) \tag{9}$$

for some constant $\gamma_n > 0$.

To guarantee that a sufficient normal Cauchy decrease is achieved, one may use the exact Cauchy step itself as a normal step. Then (9) is obviously fulfilled with $\gamma_n = 1$. If one uses the inexact dogleg step, one can ensure that (9) holds by maximizing $\text{npred}_k(.)$ over the dogleg path. For our convergence analysis, the normal steps n_k have to fulfill the *range space condition*

$$\exists\, v_k \in \mathbb{R}^M \quad \text{such that} \quad n_k = A^T(x_k)v_k, \quad \text{i.e.,} \quad n_k \perp \ker\left(A(x_k)\right), \tag{10}$$

holds for all iterations $k \in \mathbb{N}$. Note that the normal steps n_k^D, n_k^C, and a linear combination of n_k^D and n_k^C are of the form $A^T(x_k)v_k$ such that they fulfill (10).

Since $\alpha = 0$ is feasible for the optimization problem (7), it follows from (9) that

$$\text{npred}_k(n_k) \geq 0 \tag{11}$$

holds. One can improve the bound on the normal predicted reduction as shown, for example, in [7, Lemma 15.4.17] and [4, Lemma 2]. The main ingredients of the proofs are the normal Cauchy decrease condition and the property $\|A(x_k)u_k^C\| > 0$

for $u_k^C \equiv -A(x_k)^T c(x_k) \neq 0$ due to the full rank of $A(x_k)$, i.e., assumption AS1, and $u_k^C \perp \ker(A(x_k))$. Since the inexactness of the Jacobian does not influence the derivation of the result, we skip the proof of the following lemma. It can be proved exactly along the lines of Lemma 2 in [4].

LEMMA 3.1. *Suppose that assumption* AS1 *holds. Let n_k be an approximate solution of the normal subproblem* (5) *such that* (9) *holds. Then*

$$\|c(x_k)\|\mathrm{npred}_k(n_k) \geq \frac{\gamma_n}{2}\|A(x_k)^T c(x_k)\| \min\left\{\tilde{\Delta}_k, \frac{\|A(x_k)^T c(x_k)\|}{\|A(x_k)\|^2}\right\}. \tag{12}$$

3.2. The tangential subproblem. Given a current iterate x_k, we compute the tangential step towards optimality. Usually, one tries to maintain linearized feasibility, i.e., the exact tangential step $t(x_k) = Z(x_k)p_k$ should be in the exact null space of the constraints. Since we have only an approximation Z_k of the exact null space $Z(x_k)$ available, we will have to safeguard the computation of the tangential step $t_k = Z_k p_k$ by limiting the amount of inexactness, as will be explained later.

However, first we concentrate on computing an approximate solution of the inexact *tangential subproblem*

$$\begin{aligned} \min_{p\in\mathbb{R}^{N-M}} \ & (\nabla f(x_k)+B_k n_k)^T Z_k p + \frac{1}{2} p^T Z_k^T B_k Z_k p, \\ & \|Z_k p\| \leq \hat{\Delta}_k, \end{aligned} \tag{13}$$

with $\hat{\Delta}_k = (1-\kappa)\Delta_k$.

The steepest descent direction in the null space basis variables for this optimization problem at $p = 0$ is given by

$$p_k^C = -Z_k^T(\nabla f(x_k)+B_k n_k); \tag{14}$$

see, e.g., [4, 17]. For judging the improvement provided by the tangential step, we define the *tangential predicted reduction* produced by a tangential step $t = Z_k p$ as change in the objective function of the tangential subproblem. Hence, we have

$$\mathrm{tpred}_k(t) = -(\nabla f(x_k)+B_k n_k)^T t - \frac{1}{2} t^T B_k t.$$

To ensure global convergence of our trust-region algorithm, we will impose the following condition on the tangential step.

Tangential Cauchy decrease condition. *An approximate solution t_k of the tangential subproblem* (13) *must satisfy*

$$\mathrm{tpred}_k(t_k) \geq \gamma_t \,\mathrm{tpred}_k(\theta_k^C Z_k p_k^C) \tag{15}$$

for some constant $\gamma_t > 0$, where θ_k^C solves the problem

$$\begin{aligned} & \min_{\theta\geq 0}\left[-\mathrm{tpred}_k(\theta Z_k p_k^C)\right] \\ & \text{subject to } \ \|\theta Z_k p_k^C\| \leq \hat{\Delta}_k. \end{aligned} \tag{16}$$

Since $\theta = 0$ is feasible for the optimization problem (16), it follows that

$$\mathrm{tpred}_k(t_k) \geq 0. \tag{17}$$

For deriving a sharper bound on the tangential predicted reduction that is needed for the convergence analysis, we cite the following result [4, Lemma 1].

LEMMA 3.2. *Consider the one-dimensional problem*

$$\min_{z\geq 0} \psi(z) \equiv \frac{1}{2}az^2 - bz$$
$$\text{subject to } z \leq y,$$

where $b \geq 0$ *and* $y > 0$. *Then the optimal value* ψ_* *satisfies*

$$\psi_* \leq -\frac{b}{2}\min\left\{y, \frac{b}{|a|}\right\} \quad \textit{if } a \neq 0 \qquad \textit{and} \qquad \psi_* \leq -by \quad \textit{if } a = 0.$$

The derivation of a tighter lower bound for the tangential predicted reduction is based also on the representation of the null space of the constraint Jacobian. In the corresponding proofs of [7, Lemma 15.4.2] and [4, Lemma 3], the steepest descent direction is computed with an exact null space representation. The same holds true for the corresponding estimate in [17, section 3.1.2]. We do not require that an exact null space representation is available but use only the inexact tangential subproblem (13). Therefore, we state the full proof of the following result, where we use ideas applied to prove Lemma 3 in [4].

LEMMA 3.3. *Suppose that assumptions* AS3 *and* AS4 *hold. Let* t_k *be an approximate solution of the tangential subproblem* (13) *that satisfies* (15). *Then*

$$\text{tpred}_k(t_k) \geq \hat{\gamma}\|p_k^C\| \min\left\{\hat{\Delta}_k, \|p_k^C\|\right\} \tag{18}$$

for a constant $\hat{\gamma} > 0$.

Proof. Inequality (18) clearly holds if $p_k^C = 0$. Hence, we now assume that $p_k^C \neq 0$. Then, problem (16) is equivalent to

$$\begin{aligned} &\min_{\theta\geq 0} -\frac{1}{2}(p_k^C)^T Z_k^T B_k Z_k p_k^C \theta^2 - \|p_k^C\|^2\theta \\ &\quad \text{subject to } \theta \leq \frac{\hat{\Delta}_k}{\|Z_k p_k^C\|}. \end{aligned} \tag{19}$$

First assume that $(p_k^C)^T Z_k^T B_k Z_k p_k^C \neq 0$. Lemma 3.2 applied to problem (19) yields

$$-\text{tpred}_k(\theta_k^C Z_k p_k^C) \leq -\frac{1}{2}\|p_k^C\|^2 \min\left\{\frac{\hat{\Delta}_k}{\|Z_k p_k^C\|}, \frac{\|p_k^C\|^2}{|(p_k^C)^T Z_k^T B_k Z_k p_k^C|}\right\}.$$

Combining this inequality with (15) and using norm inequalities, we obtain

$$\text{tpred}_k(t_k) \geq \frac{\gamma_t}{2}\|p_k^C\| \min\left\{\frac{\hat{\Delta}_k}{\|Z_k^T Z_k\|^{1/2}}, \frac{\|p_k^C\|}{\|Z_k^T B_k Z_k\|}\right\}.$$

Since we assume that the approximations $\{Z_k\}$ remain bounded, we have that $\{Z_k^T Z_k\}$ are bounded. In addition, $\{B_k\}$ is bounded, which yields that $Z_k^T B_k Z_k$ is bounded. Hence, we can deduce from the last inequality that there exists a positive constant $\hat{\gamma}$ such that (18) holds.

We do not assume that B_k has full rank. Therefore, it may happen even if $p_k^C \neq 0$ that $(p_k^C)^T Z_k^T B_k Z_k p_k^C = 0$. Then the solution of (19) is given by $\theta_k^C = \frac{\hat{\Delta}_k}{\|Z_k p_k^C\|}$. It follows that

$$-\mathrm{tpred}_k(\theta_k^C Z_k p_k^C) \leq -\|p_k^C\|^2 \frac{\hat{\Delta}_k}{\|Z_k p_k^C\|} \leq -\|p_k^C\| \frac{\hat{\Delta}_k}{\|Z_k^T Z_k\|^{1/2}} \ .$$

Since $\{Z_k\}$ remains bounded, this inequality proves the assertion. □

To accelerate the convergence, one may use not the steepest descent direction given by (14) but an approximation of the Newton step. For this purpose, we may apply the Steihaug CG algorithm (see, e.g., [4, 22]) as long as (15) is fulfilled for the tangential step t_k.

The matrix Z_k only approximates the null space $Z(x_k)$ of the exact Jacobian $A(x_k)$. Hence, one has for the combined step $d_k = n_k + t_k$ that the identity $A(x_k)d_k = A(x_k)n_k$ is not necessarily valid. Therefore, we obtain for the predicted reduction (4) of the function m_k the equation

$$\begin{aligned}\mathrm{pred}_k(d_k) = & -\nabla f(x_k)^T(n_k + t_k) - \frac{1}{2}(n_k + t_k)^T B_k(n_k + t_k) \\ & + \mu_k(\|c(x_k)\| - \|c(x_k) + A(x_k)(n_k + t_k)\|) \\ = & \ \mathrm{tpred}(t_k) + \mu_k \mathrm{npred}(n_k) + \chi_k + \mathrm{err}_k(d_k),\end{aligned}$$

where

$$\chi_k = -\nabla f(x_k)^T n_k - \frac{1}{2} n_k^T B_k n_k, \tag{20}$$

$$\mathrm{err}_k(d_k) = \mu_k(\|c(x_k) + A(x_k)n_k\| - \|c(x_k) + A(x_k)d_k\|).$$

As can be seen, $\mathrm{err}_k(d_k)$ is a measure for the error in Z_k, i.e., in the approximation of $Z(x_k)$. Since the usual identity for the predicted reduction is not valid, we define an inexact predicted reduction

$$\mathrm{ipred}_k(d_k) = \mathrm{tpred}(t_k) + \mu_k \mathrm{npred}(n_k) + \chi_k \tag{21}$$

by omitting the error term. We will use this inexact measure for our trust-region algorithm. However, to ensure well-posedness and convergence for the considered class of trust-region methods, we need a bound on the error term $\mathrm{err}_k(d_k)$. Obviously, one can derive that

$$\begin{aligned}|\mathrm{err}_k(d_k)| &= \mu_k \big| \|c(x_k) + A(x_k)n_k\| - \|c(x_k) + A(x_k)d_k\| \big| \\ &\leq \mu_k \|A(x_k)t_k\| \leq \mu_k \, \nu \Delta_k^2.\end{aligned}$$

Hence, one may use a criterion like

$$\|A(x_k)t_k\| \leq \nu \Delta_k^2 \tag{22}$$

for a constant $\nu > 0$ to bound the inexactness that is due to the tangential step. This inequality can be easily verified by evaluating one Jacobian-vector product. Similar requirements on the inexactness can be found in [17, section 4.1.4] in the context of the convergence analysis of inexact trust-region methods for PDE-constrained optimization problems and in [7, section 10.4] for trust-region methods in the unconstrained

case. However, using (22) it may happen that $\mathrm{pred}_k(d_k)$ may become negative if $\mathrm{err}_k(d_k)$ is large relative to $\mathrm{ipred}_k(d_k)$. For this reason, we will use the direct criterion

$$-\mathrm{err}_k(d_k) < (1-\eta-(1-\eta)/2)\,\mathrm{ipred}_k(d_k) \tag{23}$$

for a constant $\eta \in (0,1)$. This inequality can be used to control the error in the inexact predicted reduction and therefore allows us to ensure well-posedness of the algorithm. Note that one only has to bound a negative $\mathrm{err}_k(d_k)$ since a positive error leads to an even larger $\mathrm{pred}_k(d_k)$. If (23) holds, one has

$$\begin{aligned} \mathrm{pred}_k(d_k) &= \mathrm{ipred}_k(d_k) + \mathrm{err}_k(d_k) \\ &> \mathrm{ipred}_k(d_k) - (1-\eta-(1-\eta)/2)\,\mathrm{ipred}_k(d_k) \\ &> (\eta + (1-\eta)/2)\,\mathrm{ipred}_k(d_k) \geq 0 \end{aligned} \tag{24}$$

if $\mathrm{ipred}_k(d_k) \geq 0$. Once more, (23) can be easily verified by evaluating two Jacobian-vector products.

3.3. The trust-region algorithm. After specifying the computation of the normal and tangential step, we can now state a detailed description of our algorithm for solving the NLP (1).

Algorithm I.

Start: Set initial values x_0, λ_0, $\mu_{-1} > 0$, A_0, Z_0, Δ_0, $\rho \in (0,1)$, $\eta \in (0,1)$, $\omega \in (0, \frac{1}{2})$, and $\nu > 0$

for $k = 0, 1, \ldots$

1. Compute a normal step n_k such that (9) and (10) hold.
2. Compute a tangential step t_k such that (15) holds.
   Compute the total step $d_k = n_k + t_k$.
3. Compute the smallest value $\tilde{\mu}_k$ such that

$$\mathrm{ipred}_k(d_k) = \mathrm{tpred}(t_k) + \tilde{\mu}_k \mathrm{npred}(n_k) + \chi_k \geq \rho\,\tilde{\mu}_k\,\mathrm{npred}_k(n_k). \tag{25}$$

   If $\tilde{\mu}_k \leq \mu_{k-1}$, set $\mu_k = \mu_{k-1}$, otherwise set $\mu_k = \max\{\tilde{\mu}_k, 1.5\mu_{k-1}\}$.
4. If (23) does not hold, update A_k and Z_k and go to step 1.
5. If
$$\mathrm{ared}_k(d_k) < \eta\,\mathrm{ipred}_k(d_k)$$
   decrease Δ_k by a constant factor and go to 1.
6. Set $x_{k+1} = x_k + d_k$ and choose a Δ_{k+1} such that $\Delta_{k+1} \geq \Delta_k$
7. Compute new A_{k+1}, Z_{k+1}, and Lagrange multipliers λ_{k+1} using

$$\lambda_{k+1} = -(A_{k+1}A_{k+1}^T)^{-1}A(x_{k+1})\nabla f(x_{k+1}) \tag{26}$$

   such that $\|Z_{k+1}^T A(x_{k+1})^T \lambda_{k+1}\| \leq \omega \|Z_{k+1}^T \nabla f(x_{k+1})\|$.
8. If $Z_{k+1}^T \nabla f(x_{k+1}) = 0$ and $c(x_{k+1}) = 0$ go to 7 to improve Z_{k+1},
   else increase k by 1 and go to 1.

Algorithm I represents a Byrd–Omojokun trust-region algorithm that takes the inexactness of the Jacobian and its null space representation into account. To clarify this point we will discuss now each step of Algorithm I and compare it to a standard Byrd–Omojokun approach. The computation of a normal direction in step 1 is identical to a standard approach where the normal Cauchy decrease condition and the range

space condition have to be fulfilled. Note that the inexactness of the Jacobian may enter into the normal direction due to the choice of the normal step. The tangential direction computed in step 2 has to fulfill the tangential Cauchy decrease condition, i.e., a standard requirement for a Byrd–Omojokun algorithm.

In step 3, χ_k can be of any sign. Furthermore, we have that $\mathrm{npred}_k(n_k)$ and $\mathrm{tpred}_k(t_k)$ are nonnegative due to (11) and (17). Hence, if $\mathrm{npred}_k(n_k) > 0$ holds it follows that $\mathrm{ipred}_k(d_k) \geq \rho\,\mu_k\,\mathrm{npred}_k(n_k)$ is valid when

$$\mu_k \geq \frac{-\chi_k}{(1-\rho)\mathrm{npred}_k(n_k)}.$$

This lower bound is a sufficient condition but is not necessary, as condition (25) may hold also for smaller values of μ_k. If $\mathrm{npred}_k(n_k) = 0$ one can conclude from Lemma 3.1 that $c(x_k) = 0$ due to assumption AS1. Therefore, $n_k = 0$ solves the normal subproblem (5). The solution of (5) must be unique because of the range space condition (10). It follows for $\mathrm{npred}_k(n_k) = 0$ that $n_k = 0$, $\chi_k = 0$, and that (25) is satisfied for any value of μ_k.

The additional test on (23) in step 4 ensures that the inexactness of the Jacobian and its null space representation does not harm the tangential direction too much. Due to assumption AS7, we need only a finite number of improvement steps for fixed x_k to obtain an exact $Z_k = Z(x_k)$ such that (23) is fulfilled.

Steps 5 and 6 are standard update procedures of a trust-region algorithm. One only has to remember that $\mathrm{ipred}(d_k)$ is not equal to the predicted reduction $\mathrm{pred}(d_k)$ due to the inexactness allowed here. We will see later that the algorithm converges despite this inexactness.

In step 7, we compute an approximation Z_{k+1} of the exact null space such that the inexactness is limited to a certain amount in the direction λ_{k+1}. Such an approximation can be found due to assumption AS7. Subsequently, we test whether the approximation Z_{k+1} is good enough. A stationary point of the NLP (1) would satisfy the equations

$$Z(x_{k+1})^T \nabla f(x_{k+1}) = 0, \qquad c(x_{k+1}) = 0$$

due to the first-order optimality condition. However, we do not have an exact null space representation $Z(x_{k+1})$. Therefore, in step 8 we check whether x_{k+1} is a stationary point of the inexact problem, i.e., whether the equations

$$Z_{k+1}^T \nabla f(x_{k+1}) = 0, \qquad c(x_{k+1}) = 0$$

hold. If this is the case but x_{k+1} is not a Karush–Kuhn–Tucker (KKT) point of the NLP (1), we have that $Z(x_{k+1})^T \nabla f(x_{k+1}) \neq 0$. Hence, our approximation Z_{k+1} of the null space $Z(x_{k+1})$ must be improved to obtain well-posedness. Due to assumption AS7, we need only a finite number of improvement steps for fixed x_k to obtain $Z_{k+1}^T \nabla f(x_{k+1}) \neq 0$. Hence, it follows that there can be only an infinite cycling between steps 7 and 8 if x_{k+1} is an KKT point of the NLP (1).

4. Well-posedness of Algorithm I. An important property of a trust-region algorithm is the well-posedness. Here, one has to show that the trust-region radius cannot shrink to zero if an iterate x_k is not a stationary point of the NLP (1). For this purpose, we analyze the relation of the actual and predicted reduction. We will employ ideas used in the proof of Lemma 4 in [4]. In addition, we must take into account the inexactness of the Jacobian and its null space representation. That is, we

have to ensure that the error term $\mathrm{err}_k(d_k)$ does not dominate the model. In step 4 of Algorithm I, we require that (23) holds. Employing this inequality, we can prove the following result that is related to Lemma 4 in [4].

LEMMA 4.1. *Assume that the assumptions* AS4, AS5, *and* AS7 *hold on the open convex set $\mathcal{X}$ containing all iterates. Then there exists a positive constant ζ such that for any iterate x_k and any step $d_k = n_k + t_k$ generated by Algorithm* I *with $[x_k, x_k + d_k] \subset \mathcal{X}$ and $\mathrm{ared}_k(d_k) \leq \eta\, \mathrm{ipred}_k(d_k)$, it follows that*

$$0 \leq \eta \mathrm{ipred}_k(d_k) - \mathrm{ared}_k(d_k) \leq \zeta(1 + \mu_k)\Delta_k^2 \tag{27}$$

Proof. Since $A(\cdot)$ is Lipschitz continuous, there exists a constant $\zeta' > 0$ such that

$$\begin{aligned} \big|\, \|c(x_k + d_k)\| - \|c(x_k) + A(x_k)d_k\,\| \big| &\leq \|c(x_k + d_k) - c(x_k) - A(x_k)d_k\,\| \\ &\leq \sup_{\tilde{x}\,\in[x_k, x_k+d_k]} \|A(\tilde{x}) - A(x_k)\|\, \|d_k\| \\ &\leq \zeta' \Delta_k^2\,. \end{aligned}$$

As in Lemma 4 of [4], the last inequality, the definitions (3) and (4), the Lipschitz continuity of ∇f, and the boundedness of B yield

$$\begin{aligned} |\,\mathrm{pred}_k(d_k) - \mathrm{ared}_k(d_k)\,| &\leq \Big| f(x_k + d_k) - f(x_k) - \nabla f(x_k)^T d_k - \frac{1}{2} d_k^T B_k d_k \\ &\quad + \ \mu_k(\|c(x_k + d_k)\| - \|c(x_k) + A(x_k)d_k\|) \Big| \\ &\leq \zeta(1 + \mu_k)\Delta_k^2 \end{aligned}$$

for some positive constant ζ. Combining the last two inequalities with the bound (23) on the error and therefore (24), we obtain

$$\begin{aligned} 0 < \eta \mathrm{ipred}_k(d_k) - \mathrm{ared}_k(d_k) &\leq \left(\eta + \frac{1 - \eta}{2}\right) \mathrm{ipred}_k(d_k) - \mathrm{ared}_k(d_k) \\ &\leq \mathrm{pred}_k(d_k) - \mathrm{ared}_k(d_k) \leq \zeta(1 + \mu_k)\Delta_k^2\,. \end{aligned}$$

∎

Next, we have to prove that Algorithm I cannot generate an infinite cycling between steps 1 and 5. To show that an acceptable step is determined with a finite number of reductions of Δ_k even if the Jacobian and its null space representation are inexact, we employ two properties. First, it follows for $c(x_k) = 0$ from (8), (11), and assumption AS1 that $\mathrm{npred}_k(n_{k,i}) = 0$, $n_{k,i} = 0$, and therefore $p_k^C = -Z_k^T \nabla f(x_k) \neq 0$ due to steps 7 and 8 of Algorithm I. Second, it follows for $c(x_k) \neq 0$ from assumption AS1 that $A(x_k)^T c(x_k) \neq 0$. Using these properties of our inexact setting, the proof of the following result is similar to the one of Proposition 1 in [4] taking the modified estimate (27) into account. Therefore, we only will state the parts of the proof that differ from the proof of [4, Proposition 1].

PROPOSITION 4.2. *Let assumption* AS1 *hold. Suppose that x_k is not a stationary point of the NLP* (1). *Then there exists a Δ_k^0 such that*

$$\mathrm{ared}_k(d_k) \geq \eta\, \mathrm{ipred}_k(d_k)$$

for any $\Delta \in (0, \Delta_k^0)$.

Proof. Let the iterate x_k be fixed. To prove the assertion, we assume that there is a subsequence indexed by i of trust radii $\Delta_{k,i}$ such that $\Delta_{k,i}$ converges to zero and

that $\text{ared}_k(d_{k,i}) < \eta\,\text{ipred}_k(d_{k,i})$ for the corresponding steps $d_{k,i} = n_{k,i} + t_{k,i}$ and the penalty parameter $\mu_{k,i}$ for all i.

For $\eta \in (0,1)$, the inequality $\text{ared}_k(d_{k,i}) < \eta\,\text{ipred}_k(d_{k,i})$ yields

$$\left(\eta + \frac{1-\eta}{2}\right)\text{ipred}_k(d_{k,i}) - \text{ared}_k(d_{k,i}) > \frac{1-\eta}{2}\text{ipred}_k(d_{k,i}) \geq 0.$$

Then, it follows from Lemma 4.1 in combination with $\Delta_{k,i} \to 0$ that

$$\text{ipred}_k(d_{k,i}) = (1+\mu_{k,i})o(\|d_{k,i}\|). \tag{28}$$

This equation can be used exactly along the lines of the proof of Proposition 1 in [4] to produce a contradiction proving the assertion of the proposition. Therefore, we skip the rest of the proof here. □

Hence, to obtain well-posedness of Algorithm I even in the presence of inexact first-order information one has to ensure that the approximation Z_k of the exact null space representation is not too bad. In our approach the effects of the inexactness are bounded for the tangential step by the additional condition (23). This suffices to show the bound (27). Additionally, the test on the quality of Z_k in steps 7 and 8 of Algorithm I ensures that there cannot be an infinite cycling between steps 1 and 5, i.e., an acceptable step can be computed with a finite number of iterations. Note that only the inexactness of the null space approximation Z_k but not the inexactness of the constraint Jacobian approximation A_k has to be controlled to achieve well-posedness.

5. Convergence analysis. Comparing the following theorem with its counterpart in [4], one finds that the result presented here is less general. This is due to the fact that we concentrate the analysis in this paper mainly on the influence of inexact Jacobian information. That is, we do not want to study the performance of Algorithm I in the presence of dependent constraint gradients as in [4] but focus on the effects caused by inexact constraint Jacobian information. Therefore, we assume in contrast to [4] that the exact constraint Jacobian $A(x_k)$ has full row rank, i.e., assumption AS1 holds. Otherwise, the iterates generated by Algorithm I may converge to a limit point failing the linear independence constraint qualification. For the derivation of the next result, it is not required to handle the inexactness of A_k and Z_k directly. The inexact first-order information are taken into account by Lemma 4.1 which is used in the proof of the following assertion. Due to the estimate in Lemma 4.1 that differs from [4, Proposition 1], we state the parts of the proof that differ from [4, Lemma 7] but skip the rather long remaining parts of the proof.

THEOREM 5.1 (feasibility of all limit points). *Assume that* AS1–AS7 *hold. Then we have*

$$\lim_{k\to\infty} c(x_k) = 0.$$

Proof. We define the function

$$\Psi(x) = \|A(x)^T c(x)\|\,.$$

Using the assumptions AS4 and AS5, we obtain that there are constants $\epsilon_1, \epsilon_2, \epsilon_3 > 0$ such that

$$\begin{aligned} |\Psi(x) - \Psi(x_l)| &= \|A(x)^T c(x) - A(x)^T c(x_l) + A(x)^T c(x_l) - A(x_l)^T c(x_l)\| \\ &\leq \epsilon_1 \|x - x_l\| \ + \epsilon_2 \|x - x_l\| \leq \epsilon_3 \|x - x_l\| \end{aligned} \tag{29}$$

holds for any two points x and x_l in $\mathcal{X}$. Now, consider an arbitrary iterate x_l with $\Psi_l \equiv \Psi(x_l) \neq 0$. First, we will show that Algorithm I accepts all sufficiently small steps that are in a neighborhood of the iterate x_l. If the current step d_k is acceptable, nothing has to be shown; otherwise one has $\mathrm{ared}_k(d_k) < \eta\,\mathrm{ipred}_k(d_k)$ and Lemma 4.1 can be applied. We define the ball

$$\mathcal{B}_l = \{x \,:\, \|x - x_l\| < \Psi_l/(2\epsilon_3)\}\,.$$

Applying (29) yields for any $x \in \mathcal{B}_l$ that $\Psi(x) \geq \Psi_l/2 > 0$. It follows that there exists a constant $\bar{c}$ with $\|c(x)\| \geq \bar{c} > 0$. Using Lemma 3.1 and assumption AS4 yields the existence of a constant $\epsilon_4 > 0$ such that for any iterate $x_k \in \mathcal{B}_l$ the inequality

$$\mathrm{ipred}_k(d_k) \geq \rho\,\mu_k \mathrm{npred}_k(n_k) \geq \mu_k \epsilon_4 \Psi_l \min\{\tilde{\Delta}_k, \Psi_l\} \tag{30}$$

holds. For sufficient small Δ_k it follows that

$$\mathrm{ipred}_k(d_k) \geq \mu_k \epsilon_4 \Psi_l \tilde{\Delta}_k. \tag{31}$$

Employing this inequality together with the estimate that was derived in the proof of Lemma 4.1, we have

$$0 \leq \frac{\left(\eta + \frac{1-\eta}{2}\right)\mathrm{ipred}_k(d_k) - \mathrm{ared}_k(d_k)}{\mathrm{ipred}_k(d_k)} \leq \frac{\zeta(1+\mu_k)\Delta_k^2}{\mu_k \epsilon_4 \Psi_l \tilde{\Delta}_k}$$

and therefore

$$\mathrm{ared} \geq \eta \mathrm{ipred}_k(d_k) + \left(\frac{1-\eta}{2} - \frac{\zeta(1+\mu_k)\Delta_k}{\mu_k \epsilon_4 \Psi_l}\right)\mathrm{ipred}_k(d_k)\,.$$

For Δ_k sufficiently small, the second term on the right-hand side is nonnegative. Hence, for all $x_k \in \mathcal{B}_l$ and all such Δ_k, we have

$$\mathrm{ared}_k(d_k) \geq \eta\,\mathrm{ipred}_k(d_k), \tag{32}$$

which results in acceptance of d_k due to step 4 of Algorithm I. The remainder of this proof follows exactly along the lines of [4, Lemma 7]. ∎

To prove the first-order optimality of all limit points, we need that the normal step can be bounded by the normal predicted reduction and that the penalty factor μ_k eventually becomes constant. For that purpose, we present the next two lemmas. For the proofs of the following two results, it is not necessary to handle the inexactness of A_k and Z_k directly. Nevertheless, we state the two proofs since the derivation differs slightly from the proofs contained in [4] due to the different setting.

LEMMA 5.2 (upper bound on normal step). *Let assumptions* AS1 *and* AS4 *be fulfilled. Then there exists a positive constant γ such that*

$$\|n_k\| \leq \gamma\,\mathrm{npred}_k(n_k). \tag{33}$$

Proof. Using Lemma 3.1, we have for the normal step

$$\|c(x_k)\|\mathrm{npred}_k(n_k) \geq \frac{\gamma_n}{2}\|A(x_k)^T c(x_k)\| \min\left\{\tilde{\Delta}_k, \frac{\|A(x_k)^T c(x_k)\|}{\|A(x_k)\|^2}\right\}.$$

If $c(x_k) = 0$, then inequality (33) is trivially satisfied. Therefore assume $c(x_k) \neq 0$. Since $A(x_k)$ is supposed to remain bounded there exists a constant $\bar{\sigma} = \sup_k \|A(x_k)\|$. Together with assumption AS1, this gives

$$\text{npred}_k(n_k) \geq \frac{\gamma_n}{2}\hat{\sigma} \min\left\{\tilde{\Delta}_k, \frac{\hat{\sigma}\|c(x_k)\|}{\bar{\sigma}^2}\right\}. \tag{34}$$

Now, we have to consider two cases. First, let $\|c(x_k)\| \geq \hat{\sigma}\tilde{\Delta}_k/2$. Using $\bar{\sigma} \geq \hat{\sigma}$ and the trust-region constraint, we obtain

$$\text{npred}_k(n_k) \geq \frac{\gamma_n}{2}\hat{\sigma} \min\left\{1, \frac{\hat{\sigma}^2}{2\bar{\sigma}^2}\right\}\tilde{\Delta}_k \geq \frac{\gamma_n\hat{\sigma}^3}{4\bar{\sigma}^2}\|n_k\|.$$

This yields (33). Second, assume that $\|c(x_k)\| < \hat{\sigma}\tilde{\Delta}_k/2$. To derive the upper bound (33) in this case, we employ (10) and Lemma 3.1. Hence, there exists a vector $v_k \in \mathbb{R}^M$ such that

$$\begin{aligned}\|c(x_k)\|^2 &\geq \|c(x_k) + A(x_k)n_k\|^2 \\ &= \|c(x_k)\|^2 + 2c(x_k)^T A(x_k)n_k + \|A(x_k)A(x_k)^T v_k\|^2.\end{aligned}$$

One obtains

$$\|A(x_k)A(x_k)^T v_k\|^2 \leq -2c(x_k)^T A(x_k)n_k.$$

Using the Cauchy–Schwarz inequality, it follows that

$$\|A(x_k)A(x_k)^T v_k\| \leq 2\|c(x_k)\|.$$

Due to assumption AS1, this inequality implies that

$$\|n_k\| = \|A(x_k)^T v_k\| \leq \frac{2}{\hat{\sigma}}\|c(x_k)\|.$$

Employing the last inequality and (34), we have

$$\begin{aligned}\text{npred}_k(n_k) &\geq \frac{\gamma_n}{2}\hat{\sigma} \min\left\{\tilde{\Delta}_k, \frac{\hat{\sigma}\|c(x_k)\|}{\bar{\sigma}^2}\right\} \geq \frac{\gamma_n}{2}\hat{\sigma} \min\left\{\frac{2}{\hat{\sigma}}, \frac{\hat{\sigma}}{\bar{\sigma}^2}\right\}\|c(x_k)\| \\ &\geq \gamma_n \min\left\{\frac{2}{\hat{\sigma}}, \frac{\hat{\sigma}}{\bar{\sigma}^2}\right\}\|n_k\|,\end{aligned}$$

which concludes the proof. □

LEMMA 5.3 (bound on hpred and constant μ_k for $k \geq k_1$). *Suppose that assumptions* AS1 *and* AS4 *are satisfied. Then the sequence of penalty parameters* $\{\mu_k\}$ *is bounded. Furthermore, there exist an index* k_1 *and positive constants* $\bar{\mu}$ *and* ξ *such that* $\mu_k = \bar{\mu}$ *holds for all* $k \geq k_1$ *and*

$$\text{ipred}_k(d_k) \geq \xi \text{tpred}_k(t_k). \tag{35}$$

Proof. The sequences $\{\nabla f(x_k)\}$ and $\{B_k\}$ are bounded due to assumption AS4. It follows from (8) that $\text{npred}_k(n_k) \leq \|c(x_k)\|$. Furthermore, $\|c(x_k)\|$ is bounded due to assumption AS4. Hence, $\text{npred}_k(n_k)$ is bounded. Using (33), we obtain that there exists a constant ξ_1 such that

$$-\nabla f(x_k)^T n_k - \frac{1}{2}n_k^T B_k n_k \geq -\xi_1 \text{npred}_k(n_k).$$

Then, we can deduce from the definition (21) of $\mathrm{ipred}_k(d_k)$ that

$$\mathrm{ipred}_k(d_k) \geq \mathrm{tpred}(t_k) + \mu_k \mathrm{npred}(n_k) - \xi_1 \mathrm{npred}_k(n_k). \tag{36}$$

Employing that $\mathrm{npred}_k(n_k) \geq 0$ and $\mathrm{tpred}_k(t_k) \geq 0$, we can derive from this inequality that (25) in step 3 of Algorithm I holds for $\mu_k \geq \xi_1/(1-\rho)$. Hence, if μ_k becomes larger than $\xi_1/(1-\rho)$, it will never be increased. Taking into account that Algorithm I increases μ_k by a constant factor this yields that after some iterate, e.g., k_1, μ_k will remain unchanged at some value $\bar{\mu}$.

Then, it follows from (21) and (25) that

$$\mathrm{ipred}_k(d_k) \geq \mathrm{tpred}(t_k) - \xi_1 \mathrm{npred}_k(n_k) \geq \mathrm{tpred}(t_k) - \frac{\xi_1}{\rho\,\mu_k} \mathrm{ipred}_k(d_k).$$

Hence, (35) is satisfied with $1/\xi = 1 + \xi_1/(\rho\,\bar{\mu})$. ∎

Now, the field is prepared to prove the main result of this paper, namely, the convergence to a first-order critical point from an arbitrary starting point. That is, we prove global convergence for our trust-region method given by Algorithm I. For this purpose, we have to take the inexactness of Z_k explicitly into account: the bound on the error in the null space representation provided by step 7 of Algorithm I is directly required to prove the following result. Therefore, we state the full proof, where we also employ ideas from [4, Lemma 12].

THEOREM 5.4 (all limit points are first-order optimal). *Suppose that* AS1–AS7 *hold. Then, it follows that*

$$\lim_{k\to\infty} \nabla_x \mathcal{L}(x_k, \lambda_k) = \lim_{k\to\infty} (\nabla f(x_k) + A(x_k)^T \lambda_k) = 0,$$

where the multipliers λ_k are defined as in (26).

Proof. Step 7 of Algorithm I ensures that

$$\|Z_k^T A(x_k)^T \lambda_k\| \leq \omega \|Z_k^T \nabla f(x_k)\|$$

for $\omega \in (0, \frac{1}{2})$ and $k > 0$. This yields for $q_k = \nabla_x \mathcal{L}(x_k, \lambda_k)$

$$\begin{aligned} \|Z_k^T q_k\| &= \|Z_k^T (\nabla f(x_k) + A(x_k)^T \lambda_k)\| \geq \|Z_k^T \nabla f(x_k)\| - \|Z_k^T A(x_k)^T \lambda_k\| \\ &\geq (1-\omega) \|Z_k^T \nabla f(x_k)\| \geq \frac{1-\omega}{\omega} \|Z_k^T A(x_k)^T \lambda_k\|. \end{aligned}$$

Setting $\varrho = \omega/(1-\omega) \in (0,1)$, we obtain

$$\varrho \|Z_k^T q_k\| \geq \|Z_k^T A(x_k)^T \lambda_k\|.$$

It follows that there exists a constant γ_1' such that AS3 and AS4 yield

$$\begin{aligned} \|p_k^C\| &= \| - Z_k^T (\nabla f(x_k) + B_k n_k)\| \\ &= \| - Z_k^T \nabla f(x_k) - Z_k^T A(x_k)^T \lambda_k + Z_k^T A(x_k)^T \lambda_k - Z_k^T B_k n_k\| \\ &= \| - Z_k^T q_k + Z_k^T A(x_k)^T \lambda_k - Z_k^T B_k n_k\| \\ &\geq \check{\sigma} \|q_k\| - \varrho \check{\sigma} \|q_k\| - \gamma_1' \|n_k\| = (1-\varrho) \check{\sigma} \|q_k\| - \gamma_1' \|n_k\| \end{aligned}$$

is valid.

To obtain a contradiction, suppose that $\lim_{k\to\infty} q_k = 0$ does not hold. Then, there exists a constant $\vartheta > 0$ such that $0 < \vartheta \leq \frac{1}{4} \limsup_{k\to\infty} \|q_k\|$. Lemma 5.1

ensures that $c(x_k) \to 0$. Together with Lemma 5.2 this yields $\|n_k\| \to 0$. Hence, there is an arbitrarily large l such that for the iterate x_l and all $k \geq l$, we have $\|q_l\| > 3\vartheta$ and $\gamma_1'\|n_k\| < (1-\varrho)\check{\sigma}\vartheta$. Let γ_L be the Lipschitz constant for q_k. We define the ball $\mathcal{B}_l = \{x : \|x - x_l\| \leq \vartheta/\gamma_L\}$. Now, assume that the iterates x_k with $k > l$ do not leave $\mathcal{B}_l$. Then, it follows for all k that

$$\begin{aligned}\|p_k^C\| &\geq (1-\varrho)\check{\sigma}(\|q_l\| - \|q_l - q_k\|) - \gamma_1'\|n_k\| \\ &\geq (1-\varrho)\check{\sigma}(3\vartheta - \vartheta - \vartheta) = (1-\varrho)\check{\sigma}\vartheta > 0.\end{aligned}$$

Employing Lemmas 3.3 and 5.3 gives with $\gamma_2' = \xi\,\hat{\gamma}(1-\varrho)\check{\sigma}$ that

$$\operatorname{ipred}_k(d_k) \geq \xi \operatorname{tpred}_k(t_k) \geq \gamma_2'\,\vartheta \min\{\hat{\Delta}_k, (1-\varrho)\check{\sigma}\,\vartheta\}. \tag{37}$$

Furthermore, the boundedness of $f(x_k)$ due to AS4 and the boundedness of the μ_k due to Lemma 5.3 gives

$$\phi(x_k;\mu_k) = f(x_k) + \mu_k\|c(x_k)\| \geq K \tag{38}$$

for a constant $K \in \mathbb{R}$. This yields that $\operatorname{ipred}_k(d_k) \to 0$. Together with (37) this implies $\hat{\Delta}_k \to 0$. Taking l sufficiently large yields for any $k \geq l$ with $x_k \in \mathcal{B}_l$ that $\hat{\Delta}_k \leq \min\{1, (1-\varrho)\check{\sigma}\vartheta\}$ and therefore

$$\operatorname{ipred}_k(d_k) \geq \gamma_2'\,\vartheta\hat{\Delta}_k. \tag{39}$$

If $x_k \in \mathcal{B}_l$, we employ the same argument as in the proof of Theorem 5.1 to show that an acceptable step is generated for sufficiently small Δ_k. Hence, if $x_k \in \mathcal{B}_l$ for all $k > l$, then Δ_k would eventually stop decreasing. However, this contradicts the fact that $\hat{\Delta}_k \to 0$. Thus the sequence $\{x_k\}$ must leave $\mathcal{B}_l$ for some $k > l$.

In that case, suppose that x_{k+1} is the first iterate after x_l that is not contained in $\mathcal{B}_l$. We deduce from (39) and $\hat{\Delta}_k = (1-\kappa)\Delta_k$ that

$$\begin{aligned}\phi(x_{k+1};\mu_{k+1}) &\leq \phi(x_l;\mu_l) - \eta\sum_{j=l}^{k}\operatorname{ipred}(x_j,\mu_j) \\ &\leq \phi(x_l;\mu_l) - \gamma_2'\vartheta(1-\kappa)\sum_{j=l}^{k}\Delta_j \\ &\leq \phi(x_l;\mu_l) - \gamma_2'(1-\kappa)\vartheta^2/\gamma_L.\end{aligned} \tag{40}$$

One can derive the last inequality from the fact x_{k+1} has left the ball $\mathcal{B}_l$ with radius ϑ/γ_L. The sequence $\{\phi(x_k;\mu_k)\}$ is decreasing and bounded below due to (38). Hence, it converges. This is a contradiction to the fact that l can be chosen arbitrarily large in (40) and the fact that $\vartheta > 0$. Therefore, $q_k \to 0$. ∎

Once more, one only has to limit the error due to the inexact null space representation Z_k for the proof of global convergence. Therefore, an implementation of Algorithm I will have to handle the approximation of the null space representation carefully. One possibility is to employ the TR1 update of the Jacobian that also provides an approximation of the null space representation [16]. We will present corresponding numerical results in a forthcoming paper [28].

6. Conclusion. In this paper, we have proposed and analyzed for the first time a class of trust-region methods based only on inexact information on the constraint Jacobian and the null space representation without any assumption on the method to approximate these matrices. Using two conditions measuring the inexactness of the null space representation, we prove global first-order convergence for the presented algorithm under quite mild conditions. The two required conditions on the inexactness can be easily verified during the optimization process.

Due to the nondifferentiable merit function and the weak assumptions on the inexactness, one may need to accelerate the convergence rate using additional safe-guard strategies for the inexactness possibly in combination with a second-order correction or a watch-dog technique.

In addition to this subject, future work will also comprise the handling of inequality constraints. The introduction of slack variables in combination with interior point techniques would be one possibility. Alternatively, one may analyze projection methods to incorporate, for example, bound constraints.

Acknowledgments. The author is very thankful to Lorenz T. Biegler for numerous discussions on trust-region methods. Furthermore, the author is grateful to Jorge Nocedal and Hubert Schwetlick for their motivating comments on the subject of this paper. The valuable comments and hints of the anonymous referees are also gratefully acknowledged.

REFERENCES

[1] A. Arora and L. Biegler, *A trust region SQP algorithm for equality constrained parameter estimation with simple parameter bounds*, Comput. Optim. Appl., 28 (2004), pp. 51–86.
[2] R. Byrd, *Robust trust region methods for constrained optimization*, in Proceedings of the Third SIAM Conference on Optimization, Houston, 1987.
[3] R. Byrd, F. Curtis, and J. Nocedal, *Inexact SQP Methods for Equality Constrained Optimization*, Tech. report, Northwestern University, Evanston, IL, 2006.
[4] R. Byrd, J. Gilbert, and J. Nocedal, *A trust region method based on interior point techniques for nonlinear programming*, Math. Program., 89A (2000), pp. 149–185.
[5] M. Celis, J. Dennis, and R. Tapia, *A trust region strategy for nonlinear equality constrained optimization*, in Numerical Optimization, 1984, P. T. Boggs, R. H. Byrd, and R. B. Schnabel, eds., SIAM, Philadelphia, 1985.
[6] A. Conn, N. Gould, and P. Toint, *Convergence of quasi-Newton matrices generated by the symmetric rank one update*, Math. Program., 50A (1991), pp. 177–196.
[7] A. Conn, N. Gould, and P. Toint, *Trust-region Methods*, SIAM, Philadelphia, 2000.
[8] J. Dennis, M. El-Alem, and M. Maciel, *A global convergence theory for general trust-region-based algorithms for equality constrained optimization*, SIAM J. Optim., 7 (1997), pp. 177–207.
[9] M. El-Alem, *A global convergence theory for the Celis–Dennis–Tapia trust-region algorithm for constrained optimization*, SIAM J. Numer. Anal., 28 (1991), pp. 266–290.
[10] R. Fletcher, N. Gould, S. Leyffer, P. Toint, and A. Wächter, *Global convergence of a trust-region SQP-filter algorithm for general nonlinear programming*, SIAM J. Optim., 13 (2003), pp. 635–659.
[11] R. Fletcher and S. Leyffer, *Nonlinear programming without a penalty function*, Math. Program., 91A (2002), pp. 239–269.
[12] R. Fletcher, S. Leyffer, and P. Toint, *On the global convergence of a filter-SQP algorithm*, SIAM J. Optim., 13 (2002), pp. 44–59.
[13] F. Gomes, M. Maciel, and J. Martinez, *Nonlinear programming algorithms using trust regions and augmented Lagrangians with nonmonotone penalty parameters*, Math. Program., 84A (1999), pp. 161–200.
[14] A. Griewank, *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*, Frontiers in Appl. Math. 19, SIAM, Philadelphia, 2000.
[15] A. Griewank and A. Walther, *On constrained optimization by adjoint-based quasi-Newton methods*, Optim. Methods Softw., 17 (2002), pp. 869–889.

[16] A. Griewank, A. Walther, and M. Korzec, *Maintaining factorized KKT systems subject to rank-one updates of Hessians and Jacobians*, Optim. Methods Softw., 22 (2007), pp. 279–295.

[17] M. Heinkenschloss and L. Vicente, *Analysis of inexact trust-region SQP algorithms*, SIAM J. Optim., 12 (2001), pp. 283–302.

[18] H. Jäger and E. Sachs, *Global convergence of inexact reduced SQP methods*, Optim. Methods Softw., 7 (1997), pp. 83–110.

[19] L. Jiang, L. Biegler, and G. Fox, *Optimization of pressure swing adsorption systems for air separation*, AIChE J., 49 (2003), pp. 1140–1157.

[20] M. Lalee, J. Nocedal, and T. Plantenga, *On the implementation of an algorithm for large-scale equality constrained optimization*, SIAM J. Optim., 8 (1998), pp. 682–706.

[21] F. Leibfritz and E. Sachs, *Inexact SQP interior point methods and large scale optimal control problems*, SIAM J. Control Optim., 38 (1999), pp. 272–293.

[22] J. Nocedal and S. Wright, *Numerical Optimization*, Springer, New York, 1999.

[23] E. Omojokun, *Trust Region Algorithms for Optimization with Nonlinear Equality and Inequality Constraints*, Ph.D. thesis, Department of Computer Science, University of Colorado, 1989.

[24] M. Powell and Y. Yuan, *A trust region algorithm for equality constrained optimization*, Math. Program., 49A (1990), pp. 189–211.

[25] G. Shultz, R. Schnabel, and R. Byrd, *A family of trust region based algorithms for unconstrained minimization with strong global convergence properties*, SIAM J. Numer. Anal., 22 (1985), pp. 47–67.

[26] M. Ulbrich and S. Ulbrich, *Non-monotone trust region methods for nonlinear equality constrained optimization without a penalty function*, Math. Program., 95B (2003), pp. 103–135.

[27] S. Volkwein and M. Weiser, *Affine invariant convergence analysis for inexact augmented Lagrangian-SQP methods*, SIAM J. Control Optim., 41 (2002), pp. 875–899.

[28] A. Walther and L. Biegler, *A Trust-Region Algorithm for Nonlinear Programming Problems with Dense Constraint Jacobians*, Tech. report MATH-WR-01-2007, TU Dresden, submitted.

[29] R. Waltz and J. Nocedal, *KNITRO User's Manual*, Tech. report OTC 05/2003, Optimization Technology Center, Northwestern University, Evanston, IL, 2003.

SIAM J. OPTIM.
Vol. 19, No. 1, pp. 326–350

GLOBAL CONVERGENCE OF A NONSMOOTH NEWTON METHOD FOR CONTROL-STATE CONSTRAINED OPTIMAL CONTROL PROBLEMS*

MATTHIAS GERDTS†

Abstract. We investigate a nonsmooth Newton method for the numerical solution of optimal control problems subject to mixed control-state constraints. The necessary conditions are stated in terms of a local minimum principle. By use of the Fischer–Burmeister function the local minimum principle is transformed into an equivalent nonlinear and nonsmooth equation in appropriate Banach spaces. This nonlinear and nonsmooth equation is solved by a nonsmooth Newton's method. We prove the global convergence and the locally superlinear convergence under certain regularity conditions. The globalized method is based on the minimization of the squared residual norm. Numerical examples for the Rayleigh problem conclude the article.

Key words. optimal control, nonsmooth Newton method, control-state constraints, global convergence

AMS subject classifications. 49J15, 49J52, 49M15

DOI. 10.1137/060657546

1. Introduction. We consider the following optimal control problem (OCP) subject to mixed control-state constraints:

$$\text{(OCP)}\qquad \begin{array}{ll} \text{Minimize} & \displaystyle\int_0^1 f_0(x(t),u(t))dt \\ \text{w.r.t.} & x \in W^{1,\infty}([0,1],\mathbb{R}^{n_x}), u \in L^\infty([0,1],\mathbb{R}^{n_u}), \\ \text{subject to (s.t.)} & x'(t) = f(x(t),u(t)) \text{ a.e. in } [0,1], \\ & \psi(x(0),x(1)) = 0, \\ & c(x(t),u(t)) \le 0 \text{ a.e. in } [0,1]. \end{array}$$

Without loss of generality the discussion is restricted to autonomous problems on the fixed time interval $[0,1]$. The functions $f_0 : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \to \mathbb{R}$, $f : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_x}$, $\psi : \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_\psi}$, $c : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_c}$ are supposed to be at least twice continuously differentiable w.r.t. to all arguments. As usual, the Banach space $L^\infty([0,1],\mathbb{R}^n)$ consists of all measurable functions $h : [0,1] \to \mathbb{R}^n$ with

$$\|h\|_\infty := \operatorname*{ess\,sup}_{0\le t\le 1} \|h(t)\| < \infty,$$

where $\|\cdot\|$ denotes the Euclidian norm on $\mathbb{R}^n$. The Banach space $W^{1,\infty}([0,1],\mathbb{R}^n)$ consists of all absolutely continuous functions $h : [0,1] \to \mathbb{R}^n$ with

$$\|h\|_{1,\infty} := \max\{\|h\|_\infty, \|h'\|_\infty\} < \infty.$$

Several approaches toward the numerical solution of OCP have been investigated in the literature. The so-called direct discretization method is based on a discretization of the infinite dimensional OCP and leads to a finite dimensional nonlinear program;

*Received by the editors April 19, 2006; accepted for publication (in revised form) August 28, 2007; published electronically March 28, 2008.
http://www.siam.org/journals/siopt/19-1/65754.html
†School of Mathematics, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK (gerdtsm@maths.bham.ac.uk).

see, e.g., Gerdts [10]. The latter can be solved numerically by suitable programming methods such as, e.g., sequential quadratic programming. The direct discretization method turns out to be very robust in practice. Nevertheless, the computational effort grows at a nonlinear rate with the number of grid points used for discretization. Convergence results for discretized OCPs can be found in Dontchev and others [7, 6], Hager [15], and Malanowski, Büskens, and Maurer [23].

The so-called indirect method for OCPs attempts to satisfy the necessary conditions that are provided by the well-known minimum principle numerically; see Hartl, Sethi, and Vickson [16] for an overview on minimum principles. The exploitation of the minimum principle leads to a nonlinear multipoint boundary value problem that has to be solved numerically; see Oberle and Grimm [28] for an implementation of a multiple shooting algorithm. Although the indirect method usually leads to the most accurate solutions, it suffers from the drawback that it requires a good initial guess in order to converge. One crucial task is to estimate the sequence of active and inactive intervals of the control-state constraint.

We refer to Büskens [3], Gerdts [12], chapter one of Grötschel, Krumke, and Rambau [13], Ioffe and Tihomirov [17], and the literature cited therein for an overview on direct discretization methods and indirect methods.

Our intention is to analyze the local and global convergence properties of an alternative method—the nonsmooth Newton method. The method is based on a nonsmooth reformulation of the necessary optimality conditions and it was introduced for the problem class OCP in Gerdts [11]. A brief outline of the essential ideas of the algorithm is as follows. The reformulation of the necessary conditions leads to the nonsmooth equation

$$F(z) = 0, \qquad F : Z \to Y,$$

where Z and Y are appropriate Banach spaces. Application of the globalized nonsmooth Newton's method generates sequences $\{z^k\}$, $\{d^k\}$, and $\{\alpha_k\}$ related by the iteration

$$z^{k+1} = z^k + \alpha_k d^k, \qquad k = 0, 1, 2, \ldots .$$

Herein, the search direction d^k is the solution of the linear operator equation $V_k(d^k) = -F(z^k)$ and the step length $\alpha_k > 0$ is determined by a line-search procedure of Armijo's type for a suitably defined merit function. The linear operator V_k is chosen from an appropriately defined generalized Jacobian $\partial_* F(z^k)$.

The nonsmooth Newton method was investigated in finite dimensions by, among others, Qi [29] and Qi and Sun [30]. Extensions to infinite spaces can be found in Kummer [19, 20], Chen, Nashed, and Qi [4], and Ulbrich [31, 32]. Our approach follows the general framework of Ulbrich [31, 32], which was used to solve certain OCPs subject to partial differential equations. The novelty of this paper is the application to the problem class OCP. The application of the nonsmooth Newton method to this problem class has not been investigated in detail before. The structure of the problem is exploited and leads to a new global convergence result in section 4. Moreover, sufficient conditions for the nonsingularity of the operator V_k are derived in section 3.

The paper is organized as follows. Section 2 introduces the nonsmooth Newton method and establishes the locally superlinear convergence under comparatively mild assumptions. In section 3, details of the computation of the search direction are shown. It turns out that the search direction solves a linear boundary value problem

with a differential-algebraic equation (DAE). If a certain operator is invertible, the so-called index of the DAE is 1 and the DAE can be transformed easily into an ordinary differential equation. A sufficient condition for the existence of the inverse operator is provided. Section 4 analyzes the global convergence properties of the nonsmooth Newton's method. Finally, numerical illustrations are presented in section 5.

2. Local convergence of the nonsmooth Newton method. The (augmented) Hamilton function $H : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_c} \to \mathbb{R}$ is defined by

$$H(x, u, \lambda, \eta) := f_0(x, u) + \lambda^\top f(x, u) + \eta^\top c(x, u).$$

We summarize the well-known minimum principle for OCP. Throughout the rest of the paper we will use the abbreviation $f[t]$ for $f(x(t), u(t))$ and likewise for other functions with time-dependent arguments. Moreover, for an index set I and a vector c with components c_i we define $c_I := (c_i)_{i\in I}$.

Let (x_*, u_*) be a (weak) local minimum of OCP and, in addition to the smoothness assumptions made above, let the following assumptions be satisfied at (x_*, u_*):

(i) Linear independence: there exist $\alpha > 0$ and $\beta > 0$ such that

$$\|c'_{I_\alpha(t),u}[t]^\top \zeta\| \geq \beta\|\zeta\|$$

for all ζ of appropriate dimension. Herein, the index set I_α is defined by $I_\alpha(t) := \{i \in \{1, \dots, n_c\} \mid c_i[t] \geq -\alpha\}$.

(ii) Controllability: for every $q \in \mathbb{R}^{n_\psi}$ there exists a solution of the linear system

$$\begin{aligned} x'(t) - f'_x[t]x(t) - f'_u[t]u(t) &= 0,\\ \psi'_{x_0}x(0) + \psi'_{x_1}x(1) &= q,\\ c'_x[t]x(t) + c'_u[t]u(t) + S_\alpha(t)\sigma(t) &= 0, \end{aligned}$$

where $S_\alpha(t) := \operatorname{diag}(c_{i,\alpha}(t))$ and $c_{i,\alpha}(t) := \min\{c_i[t] + \alpha, 0\}$.

Under these assumptions, Malanowski [22, p. 86] shows in Theorem 4.3 the regularity of the Lagrange multipliers associated with OCP. In particular, the multiplier l_0 associated with the objective function can be normalized to one and the linear operator defined by the linear system in (ii) is surjective under the assumptions (i) and (ii); see Lemma 4.1 in Malanowski [22]. Under assumptions (i) and (ii) there exist Lagrange multipliers $\lambda_* \in W^{1,\infty}([0, 1], \mathbb{R}^{n_x})$, $\eta_* \in L^\infty([0, 1], \mathbb{R}^{n_c})$, and $\sigma_* \in \mathbb{R}^{n_\psi}$ with

$$x'_*(t) - f(x_*(t), u_*(t)) = 0, \tag{2.1}$$

$$\lambda'_*(t) + H'_x(x_*(t), u_*(t), \lambda_*(t), \eta_*(t))^\top = 0, \tag{2.2}$$

$$\psi(x_*(0), x_*(1)) = 0, \tag{2.3}$$

$$\lambda_*(0) + \psi'_{x_0}(x_*(0), x_*(1))^\top\sigma_* = 0, \tag{2.4}$$

$$\lambda_*(1) - \psi'_{x_1}(x_*(0), x_*(1))^\top\sigma_* = 0, \tag{2.5}$$

$$H'_u(x_*(t), u_*(t), \lambda_*(t), \eta_*(t))^\top = 0. \tag{2.6}$$

Furthermore, the complementarity conditions hold a.e. in $[0, 1]$:

$$\eta_*(t) \geq 0, \qquad c(x_*(t), u_*(t)) \leq 0, \qquad \eta_*(t)^\top c(x_*(t), u_*(t)) = 0. \tag{2.7}$$

Remark 2.1. Similar necessary conditions can be found in Neustadt [27, Ch. VI.3] and Zeidan [34, Th. 3.1]. A regularity condition based on a controllability condition

can be found in Zeidan [34, Prop. 4.2]. The regularity assumptions (i) and (ii) with suitable extensions occur in the context of sufficient conditions [25], convergence of discretization methods [6], and sensitivity analysis [26].

Unfortunately, these necessary conditions are not directly solvable for the variable $(x_*, u_*, \lambda_*, \eta_*, \sigma_*)$ owing to the complementarity conditions. Therefore, the subsequent considerations aim at the reformulation of this set of equalities and inequalities as an equivalent system of equations, which will be solved by a generalized version of Newton's method. Notice that if the mixed-control state constraints were not present in the optimal control problem, then the generalized version of Newton's method will coincide with the (classical) Lagrange–Newton method. The Lagrange–Newton method for control constrained optimal control problems was analyzed in Alt and Malanowski [1]; later it was extended to problems involving pure state constraints in Alt and Malanowski [2]. Under suitable conditions, the authors obtain a locally quadratic convergence rate for problems with control constraints in the first paper (Theorem 4) and a superlinear convergence rate for problems with control and state constraints in the second (Theorem 5.2). This is more than we are able to show for our approach so far, as only a locally superlinear convergence rate will be established in this paper. The results of Alt and Malanowski [1, 2] suggest that it might be possible to improve the local convergence rate of our method.

The convex and locally Lipschitz continuous Fischer–Burmeister function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ is defined by

$$\varphi(a, b) := \sqrt{a^2 + b^2} - a - b, \tag{2.8}$$

(cf. Fischer [8]). The Fischer–Burmeister function has the nice property that $\varphi(a, b) = 0$ holds if and only if $a, b \geq 0$ and $ab = 0$. Hence, the complementarity conditions (2.7) are equivalent with the equality

$$\varphi(-c_i(x_*(t), u_*(t)), \eta_{i,*}(t)) = 0, \qquad i = 1, \ldots, n_c,$$

that has to hold almost everywhere in $[0, 1]$. Rather than working with the derivative of φ, which does not exist at the origin, we will work with Clarke's generalized Jacobian of φ:

$$\partial\varphi(a, b) = \begin{cases} \left\{ \left(\dfrac{a}{\sqrt{a^2 + b^2}} - 1, \dfrac{b}{\sqrt{a^2 + b^2}} - 1 \right) \right\} & \text{if } (a, b) \neq (0, 0), \\ \left\{ (s, r) \in \mathbb{R}^2 \;\middle|\; (s + 1)^2 + (r + 1)^2 \leq 1 \right\} & \text{if } (a, b) = (0, 0). \end{cases}$$

Notice that $\partial\varphi(a, b)$ is a nonempty, convex, and compact set. Let the Banach spaces

$$\begin{aligned} Z &= W^{1,\infty}([0, 1], \mathbb{R}^{n_x}) \times L^\infty([0, 1], \mathbb{R}^{n_u}) \times W^{1,\infty}([0, 1], \mathbb{R}^{n_x}) \times L^\infty([0, 1], \mathbb{R}^{n_c}) \times \mathbb{R}^{n_\psi}, \\ Y_1 &= L^\infty([0, 1], \mathbb{R}^{n_x}) \times L^\infty([0, 1], \mathbb{R}^{n_x}) \times \mathbb{R}^{n_\psi} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \times L^\infty([0, 1], \mathbb{R}^{n_u}), \\ Y_2 &= L^\infty([0, 1], \mathbb{R}^{n_c}) \end{aligned}$$

be equipped with the maximum norm for product spaces and $z_* = (x_*, u_*, \lambda_*, \eta_*, \sigma_*)$. Then, the necessary conditions (2.1)–(2.7) are equivalent with the nonlinear equation

$$F(z_*) = \begin{pmatrix} F_1(z_*) \\ F_2(z_*) \end{pmatrix} = 0, \tag{2.9}$$

where $F_1 : Z \to Y_1$ and $F_2 : Z \to Y_2$ denote the smooth and the nonsmooth part of $F : Z \to Y := Y_1 \times Y_2$, respectively:

$$F_1(z)(\cdot) := \begin{pmatrix} x'(\cdot) - f(x(\cdot), u(\cdot)) \\ \lambda'(\cdot) + H'_x(x(\cdot), u(\cdot), \lambda(\cdot), \eta(\cdot))^\top \\ \psi(x(0), x(1)) \\ \lambda(0) + \psi'_{x_0}(x(0), x(1))^\top \sigma \\ \lambda(1) - \psi'_{x_1}(x(0), x(1))^\top \sigma \\ H'_u(x(\cdot), u(\cdot), \lambda(\cdot), \eta(\cdot))^\top \end{pmatrix}, \quad F_2(z)(\cdot) := \omega(z(\cdot)), \tag{2.10}$$

where $\omega = (\omega_1, \ldots, \omega_{n_c})^\top : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_c} \times \mathbb{R}^{n_\psi} \to \mathbb{R}^{n_c}$ and

$$\omega_i(\bar{x}, \bar{u}, \bar{\lambda}, \bar{\eta}, \bar{\sigma}) := \varphi(-c_i(\bar{x}, \bar{u}), \bar{\eta}_i), \qquad i = 1, \ldots, n_c. \tag{2.11}$$

The standard approach to solve (2.9) numerically would be to apply the classical Newton method. Unfortunately, the derivative $F'(z^k)$ does not exist since the component F_2 is not differentiable. Hence, we have to find a substitute for the derivative F' in the classical Newton method. In finite dimensional spaces, such a substitute for locally Lipschitz continuous functions may be chosen from the generalized Jacobian of F defined by

$$\partial F(z) := \operatorname{co}\left\{ V \;\middle|\; V = \lim_{\substack{z_i \in D_F \\ z_i \to z}} F'(z_i) \right\},$$

where D_F denotes the set of points where F is differentiable [5]. However, in infinite dimensional spaces it is more difficult to define an appropriate generalized Jacobian since locally Lipschitz continuous functions in general are not differentiable almost everywhere. Motivated by the chain rule in finite dimensions we define the point to set mapping $\partial_* F : Z \rightrightarrows \mathcal{L}(Z, Y)$ according to

$$\partial_* F(z^k)(z) := \left\{ \begin{pmatrix} F'_1(z^k)(z) \\ -S\,(c'_x[\cdot]x + c'_u[\cdot]u) + R\eta \end{pmatrix} \;\middle|\; \begin{array}{l} S = \operatorname{diag}(s_1, \ldots, s_{n_c}), \\ R = \operatorname{diag}(r_1, \ldots, r_{n_c}), \\ (s_i, r_i) \in \partial\varphi[\cdot] \text{ a.e.}, \\ s_i(\cdot), r_i(\cdot) \text{ measurable} \end{array} \right\}$$

and use this set as a generalized Jacobian. The same idea was introduced earlier in Ulbrich [31, Def. 3.35]. Notice that the first component F_1 of F in (2.10) is continuously Fréchet-differentiable with

$$F'_1(z^k)(z) = \begin{pmatrix} x'(\cdot) - f'_x[\cdot]x(\cdot) - f'_u[\cdot]u(\cdot) \\ \lambda'(\cdot) + H''_{xx}[\cdot]x(\cdot) + H''_{xu}[\cdot]u(\cdot) + H''_{x\lambda}[\cdot]\lambda(\cdot) + H''_{x\eta}[\cdot]\eta(\cdot) \\ \psi'_{x_0}x(0) + \psi'_{x_1}x(1) \\ \lambda(0) + \psi''_{x_0x_0}(\sigma^k, x(0)) + \psi''_{x_0x_1}(\sigma^k, x(1)) + \left(\psi'_{x_0}\right)^\top \sigma \\ \lambda(1) - \psi''_{x_1x_0}(\sigma^k, x(0)) - \psi''_{x_1x_1}(\sigma^k, x(1)) - \left(\psi'_{x_1}\right)^\top \sigma \\ H''_{ux}[\cdot]x(\cdot) + H''_{uu}[\cdot]u(\cdot) + H''_{u\lambda}[\cdot]\lambda(\cdot) + H''_{u\eta}[\cdot]\eta(\cdot) \end{pmatrix},$$

provided that the functions f_0, f, c, ψ are twice continuously differentiable w.r.t. all arguments. All functions are evaluated at $z^k = (x^k, u^k, \lambda^k, \eta^k, \sigma^k) \in Z$.

Replacing the nonexisting Jacobian F' in the classical Newton method by the generalized Jacobian $\partial_* F(z^k)$ leads to the following algorithm.

ALGORITHM 2.2. LOCAL NONSMOOTH NEWTON'S METHOD.

(0) *Choose* z^0.

(1) *If some stopping criterion is satisfied, stop.*

(2) *Choose an arbitrary* $V_k \in \partial_* F(z^k)$ *and compute the search direction* d^k *from the linear equation*

$$V_k(d^k) = -F(z^k).$$

(3) *Set* $z^{k+1} = z^k + d^k$, $k = k+1$, *and goto* (1).

The assumptions needed to prove local convergence of the method are similar to those in [29], [30], [18], and [31]. $\partial_* F(z)$ is called nonsingular if for every $V \in \partial_* F(z)$ the inverse operator V^{-1} exists and if it is linear and bounded, i.e., $V^{-1} \in \mathcal{L}(Y, Z)$.

THEOREM 2.3. *Let* z_* *be a zero of* F. *Suppose that there exist constants* $\Delta > 0$ *and* $C > 0$ *such that for every* $z \in U_\Delta(z_*)$ *the generalized Jacobian* $\partial_* F(z)$ *is nonsingular and* $\|V^{-1}\|_{\mathcal{L}(Y,Z)} \le C$ *for every* $V \in \partial_* F(z)$.

(i) *Let*

$$\|F(z) - F(z_*) - V(z - z_*)\|_Y = o(\|z - z_*\|_Z) \qquad \forall V \in \partial_* F(z) \tag{2.12}$$

as $\|z - z_*\|_Z \to 0$. *Then, for* z^0 *sufficiently close to* z_* *the nonsmooth Newton method converges superlinearly to* z_*.

(ii) *Let*

$$\|F(z) - F(z_*) - V(z - z_*)\|_Y = \mathcal{O}(\|z - z_*\|_Z^{1+p}) \qquad \forall V \in \partial_* F(z) \tag{2.13}$$

as $\|z - z_*\|_Z \to 0$. *Then, for* z^0 *sufficiently close to* z_* *the nonsmooth Newton method converges at order* $1 + p$ *to* z_*.

Furthermore, if $F(z^k) \neq 0$ *for all* k, *then the residual values converge superlinearly:*

$$\lim_{k \to \infty} \frac{\|F(z^{k+1})\|_Y}{\|F(z^k)\|_Y} = 0.$$

Proof. Due to the first assumption, the algorithm is well defined in some neighborhood of z_*. It holds that

$$V_k(z^{k+1} - z_*) = V_k(z^k + d^k - z_*) = V_k(z^k - z_*) + V_k d^k = V_k(z^k - z_*) - F(z^k) + F(z_*).$$

The assertions in (i) and (ii) follow from

$$\begin{aligned} \|z^{k+1} - z_*\|_Z &= \|V_k^{-1}\left(V_k(z^k - z_*) - F(z^k) + F(z_*)\right)\|_Y \\ &\le \|V_k^{-1}\|_{\mathcal{L}(Y,Z)} \cdot \|F(z^k) - F(z_*) - V_k(z^k - z_*)\|_Y \\ &\le C \cdot \|F(z^k) - F(z_*) - V_k(z^k - z_*)\|_Y \\ &= \begin{cases} o(\|z^k - z_*\|_Z) & \text{in case (i)}, \\ \mathcal{O}(\|z^k - z_*\|_Z^{1+p}) & \text{in case (ii)}. \end{cases} \end{aligned} \tag{2.14}$$

Let $\varepsilon > 0$ be arbitrary. According to (2.14) there exists $\delta > 0$ with

$$\|z^{k+1} - z_*\|_Z \le \varepsilon \|z^k - z_*\|_Z \quad \text{whenever} \quad \|z^k - z_*\|_Z \le \delta.$$

Notice that for any $\delta > 0$ there exists some $k_0(\delta)$ such that $\|z^k - z_*\| \le \delta$ for every $k \ge k_0(\delta)$ since z^k converges to z_*. By the local Lipschitz continuity of F we get

$$\|F(z^{k+1})\|_Y = \|F(z^{k+1}) - F(z_*)\|_Y \le L\|z^{k+1} - z_*\|_Z \le L\varepsilon \|z^k - z_*\|_Z$$

locally around z_* and the Newton iteration implies

$$\|z^{k+1} - z^k\|_Z \le \|V_k^{-1}\|_{\mathcal{L}(Y,Z)} \cdot \|F(z^k)\|_Y \le C\|F(z^k)\|_Y.$$

Thus,

$$\begin{aligned}\|z^k - z_*\|_Z &\le \|z^{k+1} - z^k\|_Z + \|z^{k+1} - z_*\|_Z \\ &\le C\|F(z^k)\|_Y + \|z^{k+1} - z_*\|_Z \\ &\le C\|F(z^k)\|_Y + \varepsilon\|z^k - z_*\|_Z\end{aligned}$$

and

$$\|z^k - z_*\|_Z \le \frac{C}{1-\varepsilon}\|F(z^k)\|_Y.$$

Finally,

$$\|F(z^{k+1})\|_Y \le L\varepsilon\|z^k - z_*\|_Z \le \frac{L\varepsilon C}{1-\varepsilon}\|F(z^k)\|_Y.$$

Since $F(z^k) \neq 0$ and ε may be arbitrarily small, this shows the last assertion. ∎

Remark 2.4.

- The properties (2.12) and (2.13) can be written as

$$\sup_{V \in \partial_* F(z)} \|F(z) - F(z_*) - V(z - z_*)\|_Y = o(\|z - z_*\|_Z),$$
$$\sup_{V \in \partial_* F(z)} \|F(z) - F(z_*) - V(z - z_*)\|_Y = \mathcal{O}(\|z - z_*\|_Z^{1+p})$$

  as $\|z - z_*\|_Z \to 0$ and are referred to as semismoothness and p-order semismoothness of F at z_*; see Ulbrich [31, Def. 3.1].
- It suffices if the assumptions are satisfied for certain elements of $\partial_* F$ provided that only these elements are used in the algorithm. For the upcoming computations we used the element corresponding to the choices

$$s_i(t) = \begin{cases} -1 & \text{if } c_i[t] = 0,\ \eta_i(t) = 0, \\ \frac{-c_i[t]}{\sqrt{c_i[t]^2 + \eta_i(t)^2}} - 1 & \text{otherwise,} \end{cases}$$
$$r_i(t) = \begin{cases} 0 & \text{if } c_i[t] = 0,\ \eta_i(t) = 0, \\ \frac{\eta_i(t)}{\sqrt{c_i[t]^2 + \eta_i(t)^2}} - 1 & \text{otherwise.} \end{cases}$$

The first component F_1 is continuously Fréchet-differentiable if f_0, f, c, ψ are twice continuously differentiable. The Fréchet-differentiability immediately yields (2.12) for the component F_1. If the second derivatives of f_0, f, c, ψ are even locally Lipschitz continuous, then F_1' also satisfies a local Lipschitz condition of type

$$\|F_1'(z + d) - F_1'(z)\|_{\mathcal{L}(Z,Y_1)} \le L\|d\|_Z.$$

Using this property and the mean-value theorem we find

$$\begin{aligned}\|F_1(z + d) - F_1(z) - F_1'(z + d)(d)\|_{Y_1} &\le \int_0^1 \|(F_1'(z + td) - F_1'(z + d))(d)\|_{Y_1} dt \\ &\le \int_0^1 \|F_1'(z + td) - F_1'(z + d)\|_{\mathcal{L}(Z,Y_1)} dt \cdot \|d\|_Z \\ &\le \frac{L}{2}\|d\|_Z^2\end{aligned}$$

and thus (2.13) with $p = 1$ holds for F_1.

The second component $F_2(z)(t) = \omega(z(t))$ of F in (2.10) is a superposition operator as in [31, Sec. 3.3] with the difference that F_2 maps from some subset of L^∞ to L^∞. This allows us to consider the operator F_2 pointwise since $\|z - z_*\|_Z \to 0$ implies $\|z(t) - z_*(t)\| \to 0$ a.e. in $[0,1]$. Let us summarize some well-known results for finite dimensions. The Fischer–Burmeister function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ is 1-order semismooth (and particularly semismooth) according to Fischer [9, Lem. 20]. Furthermore, due to a result of Mifflin, the composition $g = g_1 \circ g_2$ of semismooth functions g_1, g_2 is again semismooth [9, p. 527]. Similarly, the composition of 1-order semismooth functions is again 1-order semismooth [9, Th. 19]. In particular, continuously differentiable functions are semismooth and functions having a locally Lipschitz continuous first derivative are 1-order semismooth. Consequently, the function ω in (2.11) is semismooth if the function c is continuously differentiable. Moreover, ω is 1-order semismooth if c' is locally Lipschitz continuous. With these remarks the semismoothness and the 1-order semismoothness of the superposition operator $F_2 : Z \to Y_2$ in (2.10) and (2.11) are established by the following lemma.

LEMMA 2.5. *Consider the operator*

$$g : L^\infty([0,1], \mathbb{R}^n) \to L^\infty([0,1], \mathbb{R}^m), \qquad z \mapsto g(z)(t) = \omega(z(t)).$$

It holds that

(i) *g is semismooth at z (in the sense of Remark 2.4) if $\omega : \mathbb{R}^n \to \mathbb{R}^m$ is semismooth at $z(t) \in \mathbb{R}^n$ for a.e. $t \in [0,1]$.*

(ii) *g is p-order semismooth at z (in the sense of Remark 2.4) if ω is uniformly p-order semismooth at z, i.e., there exists $C_z > 0$ such that for almost every $\bar{z} \in \{z(t) \in \mathbb{R}^n \mid t \in [0,1]\}$ it holds that*

$$\max_{V \in \partial\omega(\bar{z}+h)} \|\omega(\bar{z}+h) - \omega(\bar{z}) - Vh\| \le C_z \|h\|^{1+p} \qquad \text{as} \qquad \|h\| \to 0.$$

Proof. Define $\rho : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^m$ by

$$\rho(x,h) := \max_{V \in \partial\omega(x+h)} \|\omega(x+h) - \omega(x) - Vh\|.$$

(i) Owing to the semismoothness of ω at $z(t)$ for a.e. $t \in [0,1]$ it holds that

$$a(t) = \frac{\rho(z(t), d(t))}{\|d\|_\infty} = \frac{o(\|d(t)\|)}{\|d\|_\infty} \to 0$$

as $\|d(t)\| \to 0$ for a.e. $t \in [0,1]$. Since $\|d\|_\infty \to 0$ implies $\|d(t)\| \to 0$ a.e., it holds that $\|\rho(z(\cdot), d(\cdot))\|_\infty = \|a\|_\infty \cdot \|d\|_\infty = o(\|d\|_\infty)$.

(ii) The uniform p-order semismoothness of ω at z yields

$$\rho(z(t), d(t)) \le C_z \|d(t)\|^{1+p} \le C_z \|d\|_\infty^{1+p}$$

a.e. in $[0,1]$, where C_z does not depend on t. The assertion follows from $\|\rho(z(\cdot), d(\cdot))\|_\infty \le C_z \|d\|_\infty^{1+p}$. ∎

Application of the lemma and the previous considerations yield the following result.

THEOREM 2.6. *Let z_* be a zero of F. Suppose that there exist constants $\Delta > 0$ and $C > 0$ such that for every $z \in U_\Delta(z_*)$ the generalized Jacobian $\partial_* F(z)$ is nonsingular and $\|V^{-1}\|_{\mathcal{L}(Y,Z)} \le C$ for every $V \in \partial_* F(z)$.*

The nonsmooth Newton's method converges locally at a superlinear rate if f_0, f, c, ψ are twice continuously differentiable.

Proof. The assertion follows from Lemma 2.5 since the Fischer–Burmeister function is semismooth at every $(a,b)^\top \in \mathbb{R}^2$ and thus ω in (2.11) is semismooth everywhere provided that c is continuously differentiable. ∎

Remark 2.7. Unfortunately, the quadratic convergence of the method could not be established for the following reason. The Fischer–Burmeister function is twice continuously differentiable at every argument $(a,b)^\top \neq (0,0)^\top$, but the second derivative is getting unbounded for $(a,b) \to (0,0)$. Hence, the assumption in (ii) of Lemma 2.5 can be satisfied at z_* only if $\eta_i(t)$ and $c_i(x_*(t),u_*(t))$ do not approach zero simultaneously. Thus, a uniform strict complementarity condition

$$|\eta_{*,i}(t)| + |c_i(x_*(t),u_*(t))| \geq \alpha \tag{2.15}$$

has to hold a.e. in $[0,1]$ for all $i = 1,\dots,n_c$ and some $\alpha > 0$.

Unfortunately, as a rule for problems of type OCP, η_i or u is continuous and thus condition (2.15) is not satisfied, except for the trivial case that a constraint is nowhere active or active everywhere.

3. Computation of the search direction. For brevity we neglect the arguments whenever possible. The linear operator equation $V_k(d^k) = -F(z^k)$ in step 2 of Algorithm 2.2 reads as

$$\begin{aligned}&\begin{pmatrix} x' \\ \lambda' \end{pmatrix} - \begin{pmatrix} f'_x & 0 \\ -H''_{xx} & -H''_{x\lambda} \end{pmatrix}\begin{pmatrix} x \\ \lambda \end{pmatrix} - \begin{pmatrix} f'_u & 0 \\ -H''_{xu} & -H''_{x\eta} \end{pmatrix}\begin{pmatrix} u \\ \eta \end{pmatrix} \\ &\quad = -\begin{pmatrix} (x^k)' - f \\ (\lambda^k)' + (H'_x)^\top \end{pmatrix}\end{aligned} \tag{3.1}$$

and

$$\begin{aligned}&\begin{pmatrix} \psi'_{x_0} & 0 & 0 \\ (\psi'_{x_0}{}^\top \sigma^k)'_{x_0} & I & \psi'_{x_0}{}^\top \\ -(\psi'_{x_1}{}^\top \sigma^k)'_{x_0} & 0 & -\psi'_{x_1}{}^\top \end{pmatrix}\begin{pmatrix} x(0) \\ \lambda(0) \\ \sigma \end{pmatrix} + \begin{pmatrix} \psi'_{x_1} & 0 & 0 \\ (\psi'_{x_0}{}^\top \sigma^k)'_{x_1} & 0 & 0 \\ -(\psi'_{x_1}{}^\top \sigma^k)'_{x_1} & I & 0 \end{pmatrix}\begin{pmatrix} x(1) \\ \lambda(1) \\ \sigma \end{pmatrix} \\ &= -\begin{pmatrix} \psi(x^k(0),x^k(1)) \\ \lambda^k(0) + \psi'_{x_0}{}^\top \sigma^k \\ \lambda^k(1) - \psi'_{x_1}{}^\top \sigma^k \end{pmatrix}\end{aligned} \tag{3.2}$$

and

$$\mathcal{A}\begin{pmatrix} u \\ \eta \end{pmatrix} + \begin{pmatrix} H''_{ux} & H''_{u\lambda} \\ -Sc'_x & 0 \end{pmatrix}\begin{pmatrix} x \\ \lambda \end{pmatrix} = -\begin{pmatrix} H'_u \\ \omega(z^k(\cdot)) \end{pmatrix}, \tag{3.3}$$

where

$$\mathcal{A} := \begin{pmatrix} H''_{uu} & (c'_u)^\top \\ -Sc'_u & R \end{pmatrix}. \tag{3.4}$$

Herein, every function is evaluated at the current iterate z^k. If the inverse operator $\mathcal{A}^{-1}$ exists, (3.3) can be solved for u and η according to

$$\begin{pmatrix} u \\ \eta \end{pmatrix} = -\mathcal{A}^{-1}\left[\begin{pmatrix} H''_{ux} & H''_{u\lambda} \\ -Sc'_x & 0 \end{pmatrix}\begin{pmatrix} x \\ \lambda \end{pmatrix} + \begin{pmatrix} H'_u \\ \omega(z^k(\cdot)) \end{pmatrix}\right]. \tag{3.5}$$

A sufficient condition for the nonsingularity of $\mathcal{A}$ is given in Theorem 3.2. The constant σ in (3.2) can be viewed as a solution of the differential equation $\sigma' = 0$. Introducing (3.5) into the differential equation (3.1), augmenting this system by $\sigma' = 0$,

and taking into account the boundary conditions (3.2), yields the linear boundary value problem for $\xi = (x, \lambda, \sigma)^\top$:

$$\xi' = B\xi + b, \qquad E_0\xi(0) + E_1\xi(1) = q, \tag{3.6}$$

where

$$B = \begin{pmatrix} f'_x & 0 & 0 \\ -H''_{xx} & -H''_{x\lambda} & 0 \\ 0 & 0 & 0 \end{pmatrix} - \begin{pmatrix} f'_u & 0 \\ -H''_{xu} & -H''_{x\eta} \\ 0 & 0 \end{pmatrix} \mathcal{A}^{-1} \begin{pmatrix} H''_{ux} & H''_{u\lambda} & 0 \\ -Sc'_x & 0 & 0 \end{pmatrix},$$

$$b = -\left[\begin{pmatrix} (x^k)' - f \\ (\lambda^k)' + H'_x{}^\top \\ 0 \end{pmatrix} + \begin{pmatrix} f'_u & 0 \\ -H''_{xu} & -H''_{x\eta} \\ 0 & 0 \end{pmatrix} \mathcal{A}^{-1} \begin{pmatrix} H'_u \\ \omega(z^k(\cdot)) \end{pmatrix} \right],$$

$$E_0 = \begin{pmatrix} \psi'_{x_0} & 0 & 0 \\ (\psi'_{x_0}{}^\top \sigma^k)'_{x_0} & I & \psi'_{x_0}{}^\top \\ -(\psi'_{x_1}{}^\top \sigma^k)'_{x_0} & 0 & -\psi'_{x_1}{}^\top \end{pmatrix},$$

$$E_1 = \begin{pmatrix} \psi'_{x_1} & 0 & 0 \\ (\psi'_{x_0}{}^\top \sigma^k)'_{x_1} & 0 & 0 \\ -(\psi'_{x_1}{}^\top \sigma^k)'_{x_1} & I & 0 \end{pmatrix},$$

$$q = -\begin{pmatrix} \psi(x^k(0), x^k(1)) \\ \lambda^k(0) + \psi'_{x_0}{}^\top \sigma^k \\ \lambda^k(1) - \psi'_{x_1}{}^\top \sigma^k \end{pmatrix}.$$

Hence, in each iteration of Algorithm 2.2 we have to solve the linear boundary value problem (3.6).

If the operator $\mathcal{A}$ is not invertible, the situation becomes more involved. In this case, (3.3) imposes algebraic constraints and (3.1) and (3.3) form a DAE with an index of at least 2. Actually, the case when $\mathcal{A}$ is invertible corresponds to the index 1 case. We will not go into detail here and leave this problem open for future research.

We state a sufficient condition for the existence and boundedness of the inverse operator of $\mathcal{A}$ in (3.4). The proof of this condition uses the Banach lemma [21, Th. 3].

LEMMA 3.1 (Banach lemma). *Let X_1 and X_2 be Banach spaces and $M, \Delta : X_1 \to X_2$ linear and continuous operators. Let M^{-1} exist and let $\|M^{-1}\Delta\| < 1$. Then, the operator $M + \Delta$ possesses an inverse $(M + \Delta)^{-1}$ and*

$$\|(M + \Delta)^{-1}\| \le \frac{1}{1 - \|M^{-1}\Delta\|} \|M^{-1}\|.$$

The following sufficient conditions for the boundedness of $\mathcal{A}^{-1}$ aim at the formulation of conditions that do not assume that the underlying process z satisfies the first-order necessary optimality conditions. This is important in view of globalization of the method as the iterate z^k may be arbitrary.

THEOREM 3.2. *Let $z = (x, u, \lambda, \eta, \sigma) \in Z$ be given. Define the index sets*

$$I_>(t) := \{i \in \{1, \dots, n_c\} \mid c_i[t] = 0,\ \eta_i(t) > 0\},$$
$$J_\gamma(t) := \{i \in \{1, \dots, n_c\} \mid |c_i[t]| \le \gamma\eta_i(t),\ \eta_i(t) \ge 0\}, \qquad \gamma > 0.$$

Let the following assumptions hold at z:

(i) *Let there exist constants* C_1, C_2, C_3 *such that a.e. in* $[0,1]$ *it holds that*

$$\|H''_{uu}[t]\| \le C_1, \quad \|c'_u[t]^\top\| \le C_2, \quad \|c'_u[t]\| \le C_3.$$

(ii) *(Coercivity) Let there exist a constant* $\alpha > 0$ *such that a.e. in* $[0,1]$ *it holds that*

$$d^\top H''_{uu}[t]d \ge \alpha\|d\|^2 \quad \forall \quad d \in \mathbb{R}^{n_u} \ : \ c'_{I_>(t),u}[t]d = 0.$$

(iii) *(Linear independence) Let there exist constants* $\gamma > 0$ *and* $\beta > 0$ *such that a.e. in* $[0,1]$ *it holds that*

$$\|c'_{J_\gamma(t),u}[t]^\top\zeta\| \ge \beta\|\zeta\| \quad \forall\ \zeta \text{ of appropriate dimension.}$$

Then, a.e. in $[0,1]$ *the inverse operator* $\mathcal{A}^{-1}(t)$ *exists and it holds that* $\|\mathcal{A}^{-1}(t)\| \le C$ *for some constant* C.

Proof. In what follows we will make use of the following notation. For an index set $I \subseteq \{1,\dots,n_c\}$ let $S_I := \operatorname{diag}(s_i \mid i \in I)$, $R_I := \operatorname{diag}(r_i \mid i \in I)$, and $A_I := \big(c'_{i,u} \mid i \in I\big)$. Moreover, $I^c := \{1,\dots,n_c\} \setminus I$ denotes the complementary index set of the index set I and $Q := H''_{uu}$.

Without loss of generality, using row and column permutations the operator $\mathcal{A}$ in (3.4) can be partitioned as

$$\mathcal{A}(t) = \begin{pmatrix} Q(t) & A_{I_\varepsilon(t)}(t)^\top & A_{I^c_\varepsilon(t)}(t)^\top \\ -S_{I_\varepsilon(t)}(t)A_{I_\varepsilon(t)}(t) & R_{I_\varepsilon(t)}(t) & 0 \\ -S_{I^c_\varepsilon(t)}(t)A_{I^c_\varepsilon(t)}(t) & 0 & R_{I^c_\varepsilon(t)}(t) \end{pmatrix},$$

where $I_\varepsilon(t)$ is a suitable index set depending on a constant $0 < \varepsilon < 1$. The idea behind this partition is to collect all indices i with $-\varepsilon \le r_i(t) \le 0$ in the set $I_\varepsilon(t)$, i.e.,

$$I_\varepsilon(t) := \{i \in \{1,\dots,n_c\} \mid \ -\varepsilon \le r_i(t) \le 0\}.$$

Consequently, the index set $I^c_\varepsilon(t)$ is given by

$$I^c_\varepsilon(t) = \{i \in \{1,\dots,n_c\} \mid \ -2 \le r_i(t) < -\varepsilon\}.$$

Recall that a.e. in $[0,1]$ we have $(s_i(t), r_i(t)) \in \{(s,r) \in \mathbb{R}^2 \mid (s+1)^2+(r+1)^2 \le 1\}$ and hence a.e. in $[0,1]$ it holds $-2 \le s_i(t) \le 0$ and $-2 \le r_i(t) \le 0$ for all $i \in \{1,\dots,n_c\}$.

Owing to these considerations, a.e. in $[0,1]$ the matrices $R_{I^c_\varepsilon}$ and S_{I_ε} are nonsingular and the following estimates hold (w.r.t. the spectral norm):

$$\|R_{I_\varepsilon}\| \le \varepsilon, \quad \varepsilon < \|R_{I^c_\varepsilon}\| \le 2, \quad \frac{1}{2} \le \|R^{-1}_{I^c_\varepsilon}\| < \frac{1}{\varepsilon},$$

$$0 \le \|S_{I^c_\varepsilon}\| \le 2, \quad 1 - \sqrt{\varepsilon(2-\varepsilon)} \le \|S_{I_\varepsilon}\| \le 2, \quad \frac{1}{2} \le \|S^{-1}_{I_\varepsilon}\| \le \frac{1}{1-\sqrt{\varepsilon(2-\varepsilon)}}.$$

Herein, and in what follows as well, we omit the explicit dependence on t for brevity.

In order to show the nonsingularity of $\mathcal{A}$ we investigate the linear equation

$$\begin{pmatrix} Q & A^\top_{I_\varepsilon} & A^\top_{I^c_\varepsilon} \\ -S_{I_\varepsilon}A_{I_\varepsilon} & R_{I_\varepsilon} & 0 \\ -S_{I^c_\varepsilon}A_{I^c_\varepsilon} & 0 & R_{I^c_\varepsilon} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}$$

and we will show that $\|\mathcal{A}w\| \geq C\|w\|$ holds for all $w = (w_1, w_2, w_3)^\top$ and some $C > 0$. Since S_{I_ε} and $R_{I_\varepsilon^c}$ are nonsingular, we obtain

$$\begin{pmatrix} Q + A_{I_\varepsilon^c}^\top R_{I_\varepsilon^c}^{-1} S_{I_\varepsilon^c} A_{I_\varepsilon^c} & A_{I_\varepsilon}^\top \\ A_{I_\varepsilon} & -S_{I_\varepsilon}^{-1} R_{I_\varepsilon} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} e_1 - A_{I_\varepsilon^c}^\top R_{I_\varepsilon^c}^{-1} e_3 \\ -S_{I_\varepsilon}^{-1} e_2 \end{pmatrix},$$

$$w_3 = R_{I_\varepsilon^c}^{-1} \left(e_3 + S_{I_\varepsilon^c} A_{I_\varepsilon^c} w_1 \right). \tag{3.7}$$

We will now show that the operator

$$M_\varepsilon := \begin{pmatrix} Q + T & A_{I_\varepsilon}^\top \\ A_{I_\varepsilon} & 0 \end{pmatrix}, \qquad T = A_{I_\varepsilon^c}^\top R_{I_\varepsilon^c}^{-1} S_{I_\varepsilon^c} A_{I_\varepsilon^c},$$

is nonsingular for $\varepsilon > 0$ sufficiently small and that there exists a constant $K > 0$ independent of ε with $\|M_\varepsilon^{-1}\| \leq K$. Notice that T is symmetric and positive semidefinite as $R_{I_\varepsilon^c}^{-1} S_{I_\varepsilon^c}$ is a diagonal matrix with nonnegative entries.

We need to specify the index set I_ε in more detail. It holds that

$$\begin{aligned} I_\varepsilon = \{ & i \in \{1, \ldots, n_c\} \mid |c_i| \leq \delta \eta_i,\ \eta_i > 0 \} \\ & \cup \{ i \in \{1, \ldots, n_c\} \mid c_i = 0,\ \eta_i = 0,\ r_i \geq -\varepsilon \}, \end{aligned} \tag{3.8}$$

where $\delta = \frac{\sqrt{\varepsilon(2-\varepsilon)}}{1-\varepsilon}$. This can be seen as follows. If $|c_i| \leq \delta \eta_i$ and $\eta_i > 0$, then

$$r_i = \frac{\eta_i}{\sqrt{c_i^2 + \eta_i^2}} - 1 \geq \frac{\eta_i}{\sqrt{\delta^2 \eta_i^2 + \eta_i^2}} - 1 = \frac{1}{\sqrt{1+\delta^2}} - 1 = -\left(1 - \frac{1}{\sqrt{1+\delta^2}}\right) = -\varepsilon.$$

Notice that for those indices with $c_i = 0 = \eta_i$ the corresponding values (s_i, r_i) can be chosen arbitrarily from the set $\{(s, r) \in \mathbb{R}^2 \mid (s+1)^2 + (r+1)^2 \leq 1\}$. This explains the second set on the right-hand side of (3.8). On the other hand, if $\eta_i < 0$, then $r_i < -1$. If $\eta_i = 0$ and $c_i \neq 0$, then $r_i = -1$. If $|c_i| > \delta \eta_i$ and $\eta_i > 0$, then as above $r_i < -\varepsilon$. Finally, if $c_i = 0 = \eta_i$ and $r_i < -\varepsilon$, then evidently $r_i \notin I_\varepsilon$.

Notice that $I_> \subseteq I_\varepsilon$ for every $\varepsilon > 0$ as $c_i = 0$ and $\eta_i > 0$ implies $r_i = 0$. Hence,

$$\{ d \in \mathbb{R}^{n_u} \mid A_{I_\varepsilon} d = 0 \} \subseteq \{ d \in \mathbb{R}^{n_u} \mid A_{I_>} d = 0 \} \qquad \forall \varepsilon > 0$$

and (ii) implies

$$d^\top (Q + T) d \geq d^\top Q d \geq \alpha \|d\|^2 \quad \forall \quad d \in \mathbb{R}^{n_u} \ :\ A_{I_\varepsilon} d = 0. \tag{3.9}$$

Now, choose $\varepsilon > 0$ such that $\frac{\sqrt{\varepsilon(2-\varepsilon)}}{1-\varepsilon} \leq \gamma$. Then, $I_\varepsilon \subseteq J_\gamma$ and assumption (iii) imply

$$\|A_{I_\varepsilon}^\top \tilde{\zeta}\| = \|A_{I_\varepsilon}^\top \tilde{\zeta} + A_{J_\gamma \setminus I_\varepsilon}^\top \cdot 0\| \geq \beta \|(\tilde{\zeta}, 0)\| = \beta \|\tilde{\zeta}\| \tag{3.10}$$

for every $\tilde{\zeta}$ of appropriate dimension.

Using (3.9), (3.10), assumption (i), and the same arguments as Hager [14] in the proof of Lemma 3.2, it can be shown that the matrix M_ε is nonsingular and there exists a constant K with $\|M_\varepsilon^{-1}\| \leq K$ for every $\varepsilon > 0$ satisfying $\frac{\sqrt{\varepsilon(2-\varepsilon)}}{1-\varepsilon} \leq \gamma$. It is important to point out that the constant K depends on the constants $C_1, C_2, C_3, \alpha, \beta$ but not on ε.

The operator

$$\Gamma = \begin{pmatrix} Q + T & A_{I_\varepsilon}^\top \\ A_{I_\varepsilon} & -S_{I_\varepsilon}^{-1} R_{I_\varepsilon} \end{pmatrix} = M_\varepsilon + \begin{pmatrix} 0 & 0 \\ 0 & -S_{I_\varepsilon}^{-1} R_{I_\varepsilon} \end{pmatrix} =: M_\varepsilon + \Delta_\varepsilon$$

can be viewed as a perturbation of M_ε with

$$\|\Delta_\varepsilon\| = \|S_{I_\varepsilon}^{-1} R_{I_\varepsilon}\| \le \|S_{I_\varepsilon}^{-1}\| \cdot \|R_{I_\varepsilon}\| \le \frac{\varepsilon}{1-\sqrt{\varepsilon(2-\varepsilon)}}.$$

Let $\varepsilon > 0$ be such that

$$\frac{\varepsilon}{1-\sqrt{\varepsilon(2-\varepsilon)}} < \frac{1}{K}, \qquad \frac{\sqrt{\varepsilon(2-\varepsilon)}}{1-\varepsilon} \le \gamma.$$

Then, $\|\Delta_\varepsilon\| < \frac{1}{\|M_\varepsilon^{-1}\|}$ and according to the Banach lemma, Lemma 3.1, Γ^{-1} exists and there is a constant $\tilde{K}$ with

$$\|\Gamma^{-1}\| \le \tilde{K}.$$

Equation (3.7) yields the estimates

$$\begin{aligned}\|(w_1, w_2)\| &\le \|\Gamma^{-1}\| \left(\|e_1\| + \|A_{I_\varepsilon^c}^\top\| \cdot \|R_{I_\varepsilon^c}^{-1}\| \cdot \|e_3\| + \|S_{I_\varepsilon}^{-1}\| \cdot \|e_2\| \right) \\ &\le \tilde{K} \left(1 + \frac{C_2}{\varepsilon} + \frac{1}{1-\sqrt{\varepsilon(2-\varepsilon)}} \right) \|e\| \\ &=: \tilde{C}\|e\|\end{aligned}$$

and

$$\|w_3\| \le \|R_{I_\varepsilon^c}^{-1}\| \left(\|e_3\| + \|S_{I_\varepsilon^c}\| \cdot \|A_{I_\varepsilon}\| \cdot \|w_1\| \right) \le \frac{1}{\varepsilon} \left(1 + 2C_3\tilde{C} \right) \|e\|.$$

The triangle inequality yields $\|w\| \le \|(w_1, w_2)\| + \|w_3\| \le C\|e\|$, where $C = \tilde{C} + \frac{1}{\varepsilon}(1 + 2C_3\tilde{C})$, and the assertion follows with Ljusternik and Sobolew [21, Th. 1]. □

Remark 3.3. Assumptions (ii) and (iii) of Theorem 3.2 are related to the linear independence condition and the Legendre–Clebsch condition which were imposed in assumptions (A1) and (B) in [25] and in assumptions (A3) and (A4) in [24]. However, they differ in some details. In particular, as the proof indicates, the region of validity of the uniform linear independance condition in (iii) has to be coupled to the value of the multiplier η.

It remains to establish the nonsingularity and the boundedness of the inverse of the linear operator defining the boundary value problem (3.6). This operator $G : W^{1,\infty}([0,1], \mathbb{R}^{2n_x+n_\psi}) \to L^\infty([0,1], \mathbb{R}^{2n_x+n_\psi}) \times \mathbb{R}^{2n_x+n_\psi} =: \Omega$ is defined by

$$G(\xi)(t) = \begin{pmatrix} \xi'(t) - B(t)\xi(t) \\ E_0\xi(0) + E_1\xi(1) \end{pmatrix}$$

with $\|(\omega_1, \omega_2)\|_\Omega = \max\{\|\omega_1\|_\infty, \|\omega_2\|\}$.

THEOREM 3.4. *Let the following assumptions be satisfied:*

(i) *Let there exist a constant C such that a.e. in $[0,1]$ it holds that $\|B(t)\| \le C$.*

(ii) *Let there exist $\kappa > 0$ such that for all $\zeta \in \mathbb{R}^{2n_x+n_\psi}$ it holds that*

$$\| \left(E_0\Phi(0) + E_1\Phi(1) \right) \zeta \| \ge \kappa\|\zeta\|,$$

where Φ is a fundamental solution with $\Phi'(t) = B(t)\Phi(t)$, $\Phi(0) = I$.

Then, the inverse operator G^{-1} exists and it holds $\|G^{-1}\| \le K$ for some constant K.

Proof. The proof uses a similar reasoning as Malanowski and Maurer [24, sect. 4]. Consider the boundary value problem

$$\begin{aligned} \xi'(t) - B(t)\xi(t) &= \omega_1(t), \\ E_0 x(0) + E_1 x(1) &= \omega_2. \end{aligned} \tag{3.11}$$

Since $\|G^{-1}\| = \frac{1}{\inf\{\|G(\xi)\| \mid \|\xi\|_{1,\infty}=1\}}$, we must show that $\|(\omega_1,\omega_2)\|_\Omega \geq \|\xi\|_{1,\infty}/K$ for all $(\omega_1,\omega_2) \in \Omega$ and ξ solving the above linear equation.

Consider the initial value problem

$$\tilde{\xi}'(t) = B(t)\tilde{\xi}(t) + \omega_1(t), \qquad \tilde{\xi}(0) = 0.$$

The solution is given implicitly by

$$\tilde{\xi}(t) = \int_0^t B(\tau)\tilde{\xi}(\tau) + \omega_1(\tau)d\tau.$$

Gronwall's lemma yields

$$\|\tilde{\xi}(t)\| \leq \int_0^t \|B\|_\infty \|\tilde{\xi}(\tau)\| + \|\omega_1(\tau)\| d\tau \leq \|\omega_1\|_\infty \exp(\|B\|_\infty) \leq \|\omega_1\|_\infty \exp(C).$$

Similarly, we find

$$\|\xi(t)\| \leq (\|\xi(0)\| + \|\omega_1\|_\infty)\exp(C).$$

For the fundamental system Φ we obtain

$$\|\Phi(t)\| \leq 1 + \|B\|_\infty \int_0^t \|\Phi(\tau)\| d\tau \leq \exp(\|B\|_\infty) \leq \exp(C).$$

Using the solution formula for linear differential equations we find

$$\xi(t) = \Phi(t)\left(\xi(0) + \int_0^t \Phi(\tau)^{-1}\omega_1(\tau)d\tau\right) = \Phi(t)\xi(0) + \tilde{\xi}(t).$$

Moreover,

$$\left(E_0\Phi(0) + E_1\Phi(1)\right)\xi(0) = \omega_2 - E_1\Phi(1)\int_0^1 \Phi(\tau)^{-1}\omega_1(\tau)d\tau = \omega_2 - E_1\tilde{\xi}(1).$$

It follows that

$$\kappa\|\xi(0)\| \leq \|\omega_2\| + \|E_1\|\|\tilde{\xi}(1)\| \leq \|\omega_2\| + \|E_1\|\|\omega_1\|_\infty \exp(C)$$

and thus

$$\begin{aligned} \|\xi(0)\| &\leq \frac{1}{\kappa}\left(\|\omega_2\| + \|E_1\|\|\omega_1\|_\infty \exp(C)\right) \\ &\leq \frac{1}{\kappa}\left(1 + \|E_1\|\exp(C)\right)\max\{\|\omega_2\|, \|\omega_1\|_\infty\} \\ &=: \kappa_1\|(\omega_1,\omega_2)\|_\Omega. \end{aligned}$$

Hence,

$$\|\xi(t)\| \leq (\|\xi(0)\| + \|\omega_1\|_\infty)\exp(C) \leq (\kappa_1 + 1)\exp(C)\|(\omega_1,\omega_2)\|_\Omega.$$

With $\kappa_2 := (\kappa_1 + 1)\exp(C)$ we proved $\|\xi\|_\infty \le \kappa_2 \|(\omega_1, \omega_2)\|_\Omega$. As

$$\|\xi'(t)\| \le \|B(t)\|\|\xi(t)\| + \|\omega_1(t)\| \le C\kappa_2 \|(\omega_1, \omega_2)\|_\Omega + \|\omega_1\|_\infty$$

holds, we have

$$\|\xi'\|_\infty \le (1 + C\kappa_2)\|(\omega_1, \omega_2)\|_\Omega.$$

With $K := \max\{\kappa_2, 1 + C\kappa_2\}$ we obtain $\|\xi\|_{1,\infty} \le K\|(\omega_1, \omega_2)\|_\Omega$, which shows the assertion. □

Remark 3.5. An alternative way to show the (unique) solvability of the boundary value problem (3.6) can be found in sections 3 and 4 of Malanowski and Maurer [24]. The idea is to interpret the boundary value problem (3.6) as the first-order necessary optimality conditions for a linear-quadratic accessory problem. Then, the unique solvability of the accessory problem is shown under a complete controllability condition and a coercivity condition for the objective function. The latter will be satisfied if—in addition to other assumptions—a suitably defined Riccati equation has a bounded solution.

A combination of Theorems 2.6, 3.2, and 3.4 leads to the following result.

THEOREM 3.6. *Let z_* be a zero of F. Suppose that there exists a constant $\Delta > 0$ such that for every $z \in U_\Delta(z_*)$ the assumptions of Theorems* 3.2 *and* 3.4 *hold with uniform constants. Then, the generalized Jacobian $\partial_* F(z)$ is nonsingular and there exists a constant $C > 0$ such that $\|V^{-1}\|_{\mathcal{L}(Y,Z)} \le C$ for every $V \in \partial_* F(z)$. Moreover, the assertions of Theorem* 2.6 *hold.*

4. Globalization. One reason that the Fischer–Burmeister function is appealing is that its square

$$\phi(a, b) := \varphi(a, b)^2 = \left(\sqrt{a^2 + b^2} - a - b\right)^2$$

is continuously differentiable with $\phi'(a, b) = 2\varphi(a, b)v$, where $v \in \partial\varphi(a, b)$ is arbitrary. Hence, the mappings

$$(\bar{x}, \bar{u}, \bar{\eta}) \in \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \times \mathbb{R}^{n_c} \mapsto \phi(-c_i(\bar{x}, \bar{u}), \bar{\eta}_i), \quad i = 1, \dots, n_c,$$

are continuously differentiable by the chain rule. This allows us to globalize the local nonsmooth Newton's method using the squared L^2-norm of F as a merit function:

$$\begin{aligned}
\Theta(z) &:= \frac{1}{2}\|F(z)\|_2^2 \\
&= \frac{1}{2}\int_0^1 \|x'(t) - f(x(t), u(t))\|^2 \, dt \\
&\quad + \frac{1}{2}\int_0^1 \left\|\lambda'(t) + H'_x(x(t), u(t), \lambda(t), \eta(t))^\top\right\|^2 dt \\
&\quad + \frac{1}{2}\int_0^1 \|H'_u(x(t), u(t), \lambda(t), \eta(t))\|^2 \, dt + \frac{1}{2}\sum_{i=1}^{n_c}\int_0^1 \phi(-c_i(x(t), u(t)), \eta_i(t)) dt \\
&\quad + \frac{1}{2}\|\psi(x(0), x(1))\|^2 + \frac{1}{2}\|\lambda(0) + \psi'_{x_0}(x(0), x(1))^\top \sigma\|^2 \\
&\quad + \frac{1}{2}\|\lambda(1) - \psi'_{x_1}(x(0), x(1))^\top \sigma\|^2.
\end{aligned}$$

Θ is Fréchet-differentiable in Z if f_0, f, c, ψ are twice continuously differentiable. An analysis of the derivative of Θ reveals that for d^k with $V_k(d^k) = -F(z^k)$ it holds

$$\Theta'(z^k)(d^k) = -2\Theta(z^k) = -\|F(z^k)\|_2^2. \tag{4.1}$$

As a consequence, d^k is a direction of descent of Θ at z^k and the line search in the following global version of the nonsmooth Newton's method is well defined unless z^k is a zero of F.

ALGORITHM 4.1. GLOBAL NONSMOOTH NEWTON'S METHOD.

(0) *Choose* z^0, $\beta \in (0,1)$, $\sigma \in (0, 1/2)$.

(1) *If some stopping criterion is satisfied, stop.*

(2) *Chose an arbitrary* $V_k \in \partial_* F(z^k)$ *and compute the search direction* d^k *from*

$$V_k(d^k) = -F(z^k).$$

(3) *Find smallest* $i_k \in \mathbb{N}_0$ *with*

$$\Theta(z^k + \beta^{i_k} d^k) \leq \Theta(z^k) + \sigma \beta^{i_k} \Theta'(z^k)(d^k)$$

and set $\alpha_k = \beta^{i_k}$.

(4) *Set* $z^{k+1} = z^k + \alpha_k d^k$, $k = k+1$, *and goto* (1).

The upcoming global convergence proof extends the proof presented in Jiang [18] for finite dimensions into infinite dimensions.

THEOREM 4.2. *Let the inverse operators* V_k^{-1} *exist for all* k *and let* $C > 0$ *be a constant such that* $\|V_k^{-1}\|_{\mathcal{L}(Y,Z)} \leq C$ *holds for all* k. *Let* z_* *be an accumulation point of the sequence* $\{z^k\}$ *generated by the global nonsmooth Newton method.*

Then, z_* *is a zero of* F.

Proof. Let $\{z^{k_j}\}_{j\in\mathbb{N}}$ be a subsequence with $z^{k_j} \to z_*$ and $F(z^{k_j}) \neq 0$. Then, $\Theta'(z^{k_j})(d^{k_j}) = -2\Theta(z^{k_j}) = -\|F(z^{k_j})\|_2^2 < 0$. The line search is well defined by the differentiability of Θ.

(i) Case 1: Assume

$$\alpha := \liminf_{j\to\infty} \alpha_{k_j} > 0.$$

Then

$$0 \leq \Theta(z^{k_j+1}) \leq \Theta(z^{k_j+1}) \leq \Theta(z^{k_j}) + \sigma\alpha_{k_j}\Theta'(z^{k_j})(d^{k_j}) = \Theta(z^{k_j})\left(1 - 2\sigma\alpha_{k_j}\right).$$

With $\sigma \in (0, 1/2)$ and $\alpha \leq \alpha_{k_j} \leq 1$ it follows that $0 < 1 - 2\sigma\alpha_{k_j} \leq 1 - 2\sigma\alpha < 1$, and repeated application yields

$$0 \leq \Theta(z^{k_j}) \leq \Theta(z^{k_0})\,(1 - 2\sigma\alpha)^j \to 0.$$

By the continuity of F, z_* is a zero of F.

(ii) Case 2: Assume that there is a subsequence $\{z^k\}_{k\in J}$, $J \subseteq \{k_j \mid j \in \mathbb{N}\}$ with $\alpha_k \to 0$, $k \in J$.

The sequence $\{d^k\}$ is bounded since $\{V_k^{-1}\}$ is bounded and

$$0 \leq \|d^k\|_Z = \|V_k^{-1}(F(z^k))\|_Z \leq C\|F(z^k)\|_Y \leq C\|F(z^0)\|_Y.$$

Unfortunately, the boundedness of $\{d^k\}$ in an infinite dimensional space does not imply that there exists a convergent subsequence. However, since d^k is

bounded in $Z = W^{1,\infty}([0,1],\mathbb{R}^{n_x}) \times L^\infty([0,1],\mathbb{R}^{n_u}) \times W^{1,\infty}([0,1],\mathbb{R}^{n_x}) \times L^\infty([0,1],\mathbb{R}^{n_c}) \times \mathbb{R}^{n_\psi}$, it is also bounded in the space $\hat{Z} := W^{1,2}([0,1],\mathbb{R}^{n_x}) \times L^2([0,1],\mathbb{R}^{n_u}) \times W^{1,2}([0,1],\mathbb{R}^{n_x}) \times L^2([0,1],\mathbb{R}^{n_c}) \times \mathbb{R}^{n_\psi}$. $\hat{Z}$ is a Hilbert space and thus reflexive. According to Theorem III.3.7 in [33], there exists a weakly convergent subsequence $\{d^k\}$, $k \in \hat{J} \subseteq J$. Hence, there exists some $d_* \in \hat{Z}$ such that for every element $g \in \hat{Z}^*$ it holds that

$$g(d^k) \to g(d_*). \tag{4.2}$$

Herein, $\hat{Z}^*$ denotes the topological dual space of $\hat{Z}$. The derivative $\Theta'(z_*)(\cdot)$ is an element of Z^* and an investigation reveals that it is essentially made up of linear functionals of type

$$g_1(z) = \int_0^1 h_1(z_*(t))z(t)dt, \qquad g_2(z) = \int_0^1 h_2(z_*(t))z'(t)dt$$

with essentially bounded functions $h_1(z_*(\cdot))$ and $h_2(z_*(\cdot))$. Thus, by application of the Cauchy–Schwartz inequality, the functionals g_1 and g_2 are also linear continuous functionals on $\hat{Z}$ and thus g_1, g_2, and in particular $\Theta'(z_*)(\cdot)$ can be viewed as elements of $\hat{Z}^*$.
Hence, (4.2) holds for $g(\cdot) = \Theta'(z_*)(\cdot)$:

$$\Theta'(z_*)(d^k) \to \Theta'(z_*)(d_*).$$

Furthermore, due to the continuity of $\Theta'(\cdot)$ (in Z) for every $\varepsilon > 0$ there exists $\delta > 0$ such that for every $\|z^k - z_*\|_Z \le \delta$ it holds that

$$\begin{aligned}
|\Theta'(z^k)(d^k) - \Theta'(z_*)(d^k)| &= \|d^k\|_Z \left|\Theta'(z^k)\left(\frac{d^k}{\|d^k\|_Z}\right) - \Theta'(z_*)\left(\frac{d^k}{\|d^k\|_Z}\right)\right| \\
&\le \|d^k\|_Z \cdot \sup_{\|d\|_Z=1} |\Theta'(z^k)(d) - \Theta'(z_*)(d)| \\
&= \|d^k\|_Z \cdot \|\Theta'(z^k) - \Theta'(z_*)\|_{\mathcal{L}(Z,\mathbb{R})} \le \varepsilon \|d^k\|_Z.
\end{aligned}$$

For arbitrary $\varepsilon > 0$ we find

$$\begin{aligned}
|\Theta'(z^k)(d^k) - \Theta'(z_*)(d_*)| &\le |\Theta'(z^k)(d^k) - \Theta'(z_*)(d^k)| \\
&\quad + |\Theta'(z_*)(d^k) - \Theta'(z_*)(d_*)| \\
&\le \varepsilon \|d^k\|_Z + |\Theta'(z_*)(d^k) - \Theta'(z_*)(d_*)|.
\end{aligned}$$

Since $\varepsilon > 0$ was arbitrary and since d^k is weakly convergent it holds that

$$\Theta'(z^k)(d^k) \to \Theta'(z_*)(d_*) \qquad \text{as } k \to \infty,\ k \in \hat{J}.$$

In a similar way, the Fréchet differentiability of Θ yields

$$\begin{aligned}
&\left|\frac{1}{\alpha_k}\left(\Theta(z^k + \alpha_k d^k) - \Theta(z^k)\right) - \Theta'(z_*)(d_*)\right| \\
&\le \left|\frac{1}{\alpha_k}\left(\Theta(z^k + \alpha_k d^k) - \Theta(z^k)\right) - \Theta'(z^k)(d^k)\right| + \left|\Theta'(z^k)(d^k) - \Theta'(z_*)(d_*)\right| \\
&\le \frac{1}{\alpha_k} o(\|\alpha_k d^k\|_Z) + \left|\Theta'(z^k)(d^k) - \Theta'(z_*)(d^k)\right| + \left|\Theta'(z_*)(d^k) - \Theta'(z_*)(d_*)\right| \\
&\le \|d^k\|_Z \frac{o(\alpha_k\|d^k\|_Z)}{\alpha_k\|d^k\|_Z} + \varepsilon\|d^k\|_Z + \left|\Theta'(z_*)(d^k) - \Theta'(z_*)(d_*)\right|.
\end{aligned}$$

Since d^k is weakly convergent it holds that

$$\frac{1}{\alpha_k}\left(\Theta(z^k+\alpha_k d^k)-\Theta(z^k)\right)\to\Theta'(z_*)(d_*)\qquad\text{as } k\to\infty,\ k\in\hat{J}.$$

The line search in step 3 of the algorithm yields

$$\begin{aligned}\frac{\Theta(z^k+\alpha_k d^k)-\Theta(z^k)}{\alpha_k}&\le\sigma\Theta'(z^k)(d^k),\\ \frac{\Theta(z^k+\frac{\alpha_k}{\beta}d^k)-\Theta(z^k)}{\frac{\alpha_k}{\beta}}&>\sigma\Theta'(z^k)(d^k).\end{aligned}$$

Passing to the limit and exploiting the previous considerations yields

$$\sigma\Theta'(z_*)(d_*)=\Theta'(z_*)(d_*).$$

Since $\sigma\in(0,1/2)$ this only holds for $\Theta'(z_*)(d_*)=0$. Thus, we have shown

$$-\|F(z^k)\|_2^2=\Theta'(z^k)(d^k)\to\Theta'(z_*)(d_*)=0.$$

By the continuity of F, z_* is a zero of F. □

The previous result shows only that each accumulation point is a zero of F. It would be nice to have also the fast local convergence properties of the local method. The locally superlinear convergence would follow from the local convergence theorem, Theorem 2.3, if we were able to show that $\alpha_k=1$ satisfies Armijo's rule for all sufficiently large k. But unfortunately, this leads to a two-norm discrepancy. The proof of Theorem 2.3 showed the superlinear convergence of the values $\|F(z^k)\|_Y$, i.e., for any $\varepsilon>0$ there exists $\delta>0$ such that for all $\|z-z_*\|\le\delta$ it holds that

$$\|z+d-z_*\|_Z\le\varepsilon\|z-z_*\|_Z,\qquad\|F(z+d)\|_Y\le\varepsilon\|F(z)\|_Y,$$

where $d=-V^{-1}F(z)$, $V\in\partial_*F(z)$. In particular, with $z=z^k$ and $d=d^k$ there exists $\delta>0$ such that for all $\|z^k-z_*\|_Z\le\delta$ it holds that

$$\|z^k+d^k-z_*\|_Z\le\frac{1}{2}\|z^k-z_*\|_Z,\qquad\|F(z^k+d^k)\|_Y\le\sqrt{1-2\sigma}\|F(z^k)\|_Y.$$

Unfortunately, we would need this property not for the norm $\|\cdot\|_Y$ but for the norm $\|\cdot\|_2$ since then

$$\Theta(z^k+d^k)=\frac{1}{2}\|F(z^k+d^k)\|_2^2\le\frac{1-2\sigma}{2}\|F(z^k)\|_2^2=(1-2\sigma)\Theta(z^k),$$

respectively,

$$\Theta(z^k+d^k)\le\Theta(z^k)-2\sigma\Theta(z^k)=\Theta(z^k)+\sigma\Theta'(z^k)(d^k),$$

i.e., Armijo's line search would accept $\alpha_k=1$ and $z^{k+1}=z^k+d^k$. Furthermore, $\|z^{k+1}-z_*\|_Z\le\frac{1}{2}\|z^k-z_*\|_Z\le\delta$ and we are in the same situation as above and the argument could be repeated.

Unfortunately, the superlinear convergence of the residual norms $\|F(z^k)\|_2$ could not be established by now. An additional assumption is needed to prove the fast local convergence.

THEOREM 4.3. *Let the assumptions of Theorem* 4.2 *be valid. Let, in addition, there exist a constant* $K > 0$ *such that*

$$\|F(z^k)\|_Y \le K\|F(z^k)\|_2$$

holds for $\{z^k\}$ *with* $z^k \to z_*$. *Then, for sufficiently large* k *the step length* $\alpha_k = 1$ *is accepted and the global method turns into the local one.*

Proof. Owing to the previous considerations it remains to show that

$$\lim_{k\to\infty} \frac{\|F(z^{k+1})\|_2}{\|F(z^k)\|_2} = 0.$$

Recall that $\|\cdot\|_Y$ is essentially the L^∞-norm. Hence, there exits a constant $C_1 > 0$ with $\|y\|_2 \le C\|y\|_Y$ for all $y \in Y$. Together with the superlinear convergence of the values $\|F(z^k)\|_Y$ in Theorem 2.3 for every $\varepsilon > 0$ and for sufficiently large k it holds that

$$\|F(z^k + d^k)\|_2 \le C\|F(z^k + d^k)\|_Y \le\ C\varepsilon\|F(z^k)\|_Y \le C \cdot K \cdot \varepsilon\|F(z^k)\|_2.$$

Since ε was arbitrary, this shows the superlinear convergence of the values $\|F(z^k)\|_2$. ∎

5. Numerical results. All computations were performed on a PC with 3 GHz processing speed. We used $\|F(z^k)\|^2 \le 10^{-15}$ as a stopping criterion in the nonsmooth Newton method.

5.1. Rayleigh problem, version 1. We illustrate the method for the Rayleigh problem [26, p. 39]. Minimize

$$\int_0^{4.5} u(t)^2 + x_1(t)^2 dt \tag{5.1}$$

subject to

$$\begin{aligned} x_1' &= x_2, & x_1(0) &= -5, \\ x_2' &= -x_1 + x_2\left(1.4 - 0.14x_2^2\right) + 4u, & x_2(0) &= -5, \end{aligned} \tag{5.2}$$

and

$$u + \frac{1}{6}x_1 \le 0.$$

With $x = (x_1, x_2)^\top$, $\lambda = (\lambda_1, \lambda_2)^\top$, $\sigma = (\sigma_1, \sigma_2)^\top$ the Hamilton function reads as

$$H(x, u, \lambda, \eta) = u^2 + x_1^2 + \lambda_1 x_2 + \lambda_2\left(-x_1 + x_2\left(1.4 - 0.14x_2^2\right) + 4u\right) + \eta\left(u + \frac{1}{6}x_1\right).$$

With $z = (x, u, \lambda, \eta, \sigma)$ the function F in (2.9) is given by

$$F(z) = \begin{pmatrix} x_1' - x_2 \\ x_2' - \left(-x_1 + x_2\left(1.4 - 0.14x_2^2\right) + 4u\right) \\ \lambda_1' + 2x_1 - \lambda_2 + \frac{1}{6}\eta \\ \lambda_2' + \lambda_1 + \lambda_2\left(1.4 - 0.42x_2^2\right) \\ x_1(0) + 5 \\ x_2(0) + 5 \\ \lambda_1(0) + \sigma_1 \\ \lambda_2(0) + \sigma_2 \\ \lambda_1(4.5) \\ \lambda_2(4.5) \\ 2u + 4\lambda_2 + \eta \\ \varphi\left(-\left(u + \frac{1}{6}x_1\right), \eta\right) \end{pmatrix}.$$

TABLE 5.1
Output of globalized nonsmooth Newton method for the first version of Rayleigh's problem for $N = 100$ subintervals and Euler discretization: local quadratic convergence.

ITER	ALPHA	||F||^2	||d^k||
0	0.000000E+00	0.245000E+04	0.173257E+04
1	0.531441E+00	0.173372E+04	0.316003E+04
2	0.717898E-01	0.170185E+04	0.897810E+03
3	0.185302E+00	0.155477E+04	0.653211E+03
...			
10	0.100000E+01	0.147905E-05	0.592231E-02
11	0.100000E+01	0.167034E-08	0.213155E-03
12	0.100000E+01	0.253557E-14	0.263768E-06
13	0.100000E+01	0.152582E-25	0.598877E-12

In each iteration of the nonsmooth Newton method we have to solve the linear boundary value problem (3.2), (3.6) for x, λ, σ. We leave the details of the boundary value problem (3.2), (3.6) and equation (3.5) to the reader. We note that for all $(s+1)^2 + (r+1)^2 \leq 1$ it holds

$$\det \mathcal{A} = \det \begin{pmatrix} 2 & 1 \\ -s & r \end{pmatrix} = 2r + s \neq 0$$

and thus the operator $\mathcal{A}$ in (3.4) is invertible. The differential equations are discretized on $[0, 4.5]$ using forward Euler's method with N equidistant subintervals. The occurring derivatives $(x^k)'$ and $(\lambda^k)'$ are approximated by finite forward differences. Moreover, it turned out that it is advisable to scale the boundary conditions and the transversality conditions in the merit function by the step size $h = 1/N$. The boundary value problem was solved by the single shooting method. Table 5.1 shows the output of the globalized nonsmooth Newton method, i.e., step size α, residual norm $\|F\|^2$, and $\|d^k\|$ during iteration. The iterations show the rapid quadratic convergence at the end of the iteration sequence. Recall that only a locally superlinear convergence rate was established in Theorem 2.6.

The following table summarizes results for different step sizes. The number of iterations differs only by one, which indicates—at least numerically—the mesh independence of the method. Furthermore, the CPU time grows at a linear rate with N.

N	CPU time [s]	Iterations
100	0.027	13
500	0.136	14
1000	0.271	14
2000	0.505	14
4000	1.083	14
8000	2.065	14

Figure 5.1 illustrates the iterates of the nonsmooth Newton's method. Notice the small inactive arc of the control-state constraint at the end of the time interval.

For comparison reasons the same optimal control problem was solved alternatively by a direct discretization method as in Gerdts [10] with Euler discretization and $N = 100$ subintervals. For this method the overall CPU time was 3.81 CPU seconds on the same processor. Furthermore, for the direct discretization method the CPU time grows nonlinearly with N. Hence, if all regularity assumptions are fulfilled, the nonsmooth Newton's method is an extremely efficient method.

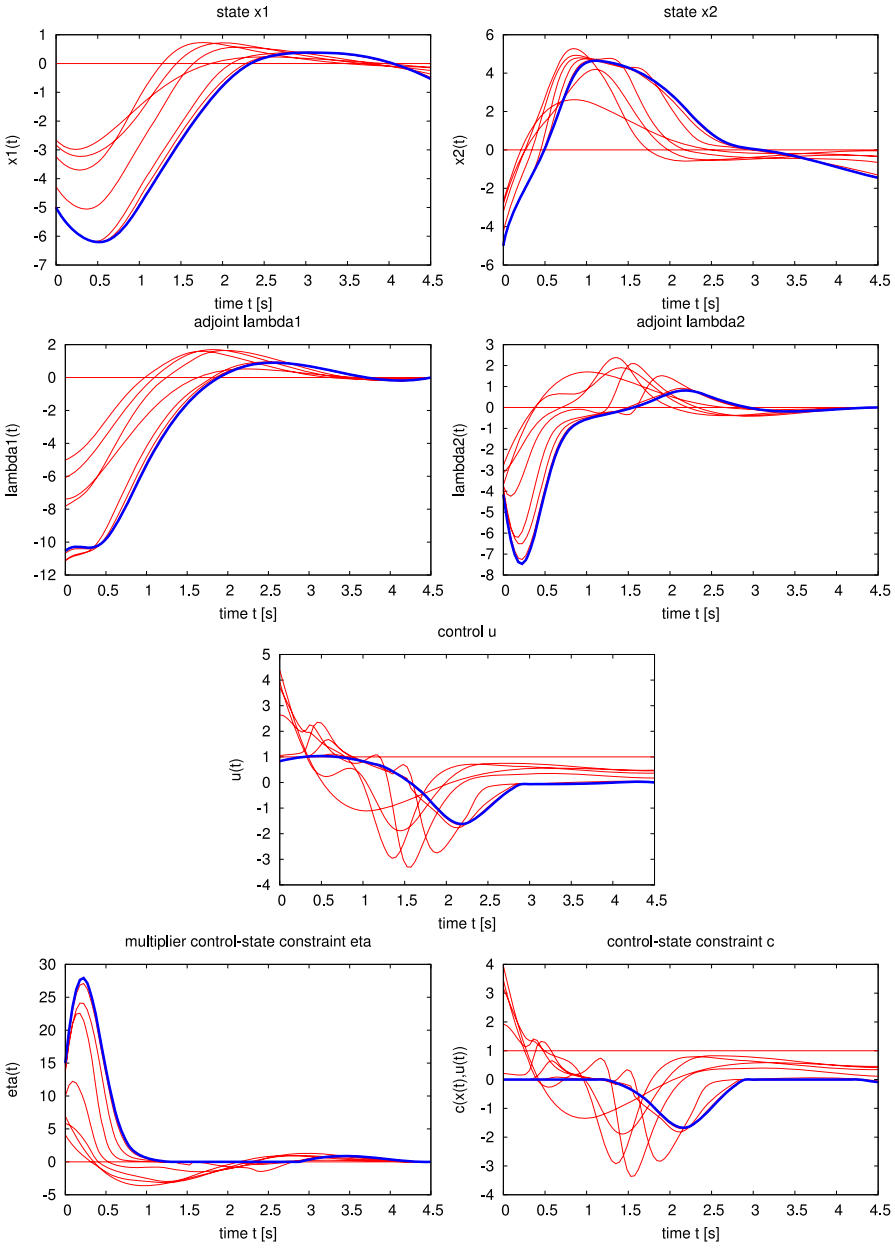


FIG. 5.1. *Numerical solution of the first version of Rayleigh's problem for* $N = 100$ *Euler steps: intermediate iterates (thin lines) and converged solution (thick lines).*

5.2. Rayleigh problem, version 2. We consider a slight variation of the Rayleigh problem where boundary conditions are added and the control-state constraint is replaced by box constraints for the control [26, p. 39]. Minimize (5.1) subject to (5.2) and $x_1(4.5) = 0$, $x_2(4.5) = 0$ and

$$-1 \le u \le 1.$$

With $x = (x_1, x_2)^\top$, $\lambda = (\lambda_1, \lambda_2)^\top$, $\sigma = (\sigma_1, \dots, \sigma_4)^\top$, $\eta = (\eta_1, \eta_2)^\top$ the Hamilton function reads as

$$H(x,u,\lambda,\eta) = u^2 + x_1^2 + \lambda_1 x_2 + \lambda_2\left(-x_1 + x_2\left(1.4 - 0.14x_2^2\right) + 4u\right) + \eta_1(u-1) + \eta_2(-u-1).$$

With $z = (x, u, \lambda, \eta, \sigma)$ the function F in (2.9) is given by

$$F(z) = \begin{pmatrix} x_1' - x_2 \\ x_2' - \left(-x_1 + x_2\left(1.4 - 0.14x_2^2\right) + 4u\right) \\ \lambda_1' + 2x_1 - \lambda_2 \\ \lambda_2' + \lambda_1 + \lambda_2\left(1.4 - 0.42x_2^2\right) \\ x_1(0) + 5 \\ x_2(0) + 5 \\ x_1(4.5) \\ x_2(4.5) \\ \lambda_1(0) + \sigma_1 \\ \lambda_2(0) + \sigma_2 \\ \lambda_1(4.5) - \sigma_3 \\ \lambda_2(4.5) - \sigma_4 \\ 2u + 4\lambda_2 + \eta_1 - \eta_2 \\ \varphi\left(-(u-1), \eta_1\right) \\ \varphi\left(-(-u-1), \eta_2\right) \end{pmatrix}.$$

Again, we leave the details of the linear boundary value problem (3.2), (3.6) and equation (3.5) to the reader. An investigation of the generalized differential of φ yields

$$\det \mathcal{A} = \det \begin{pmatrix} 2 & 1 & -1 \\ -s_1 & r_1 & 0 \\ s_2 & 0 & r_2 \end{pmatrix} = 2r_1r_2 + r_1s_2 + r_2s_1 \neq 0$$

for any $(s_1, r_1) \in \partial\varphi(-(u-1), \eta_1)$ and $(s_2, r_2) \in \partial\varphi(-(-u-1), \eta_2)$. Figure 5.2 illustrates the iterates of the nonsmooth Newton method for $N = 100$. Table 5.2 shows more detailed information about the iterations, i.e., step size α, residual norm $\|F\|^2$, and $\|d^k\|$. Again, the boundary conditions and the transversality conditions in the merit function were scaled by the step size $h = 1/N$. The iterations show the rapid quadratic convergence at the end of the iteration sequence. Recall that only a locally superlinear convergence rate was established in Theorem 2.6.

The number of iterations remains nearly constant, which indicates—at least numerically—the mesh independence of the method. Furthermore, the CPU time grows at a linear rate with N.

N	CPU time [s]	Iterations
100	0.049	17
500	0.204	15
1000	0.502	18
2000	0.848	16
4000	1.785	17
8000	3.713	17

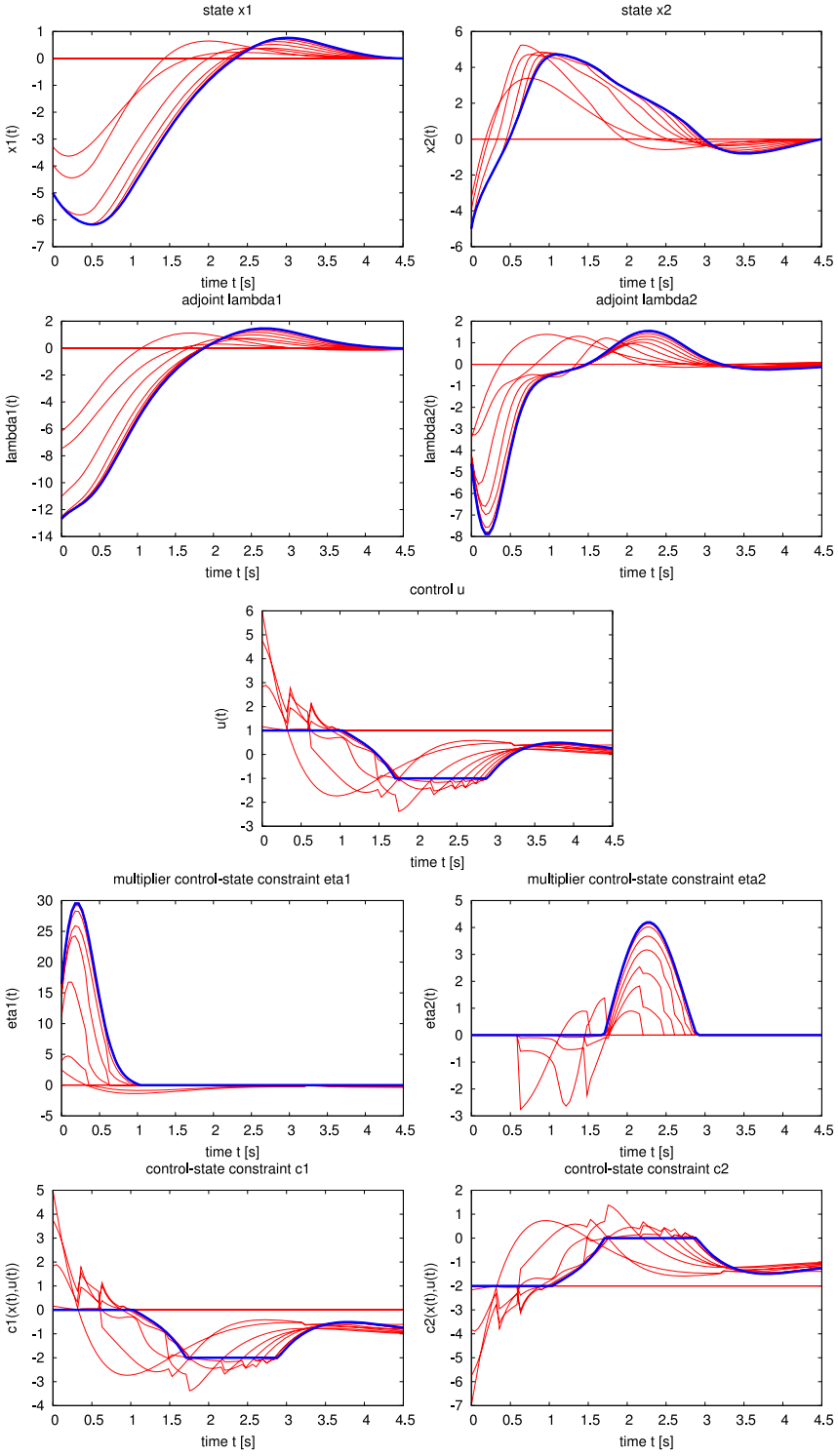


FIG. 5.2. *Numerical solution of the second version of Rayleigh's problem for* $N = 100$ *Euler steps: intermediate iterates (thin lines) and converged solution (thick lines).*

TABLE 5.2
Output of globalized nonsmooth Newton's method for the second version of Rayleigh's problem for $N = 100$ subintervals and Euler discretization: local quadratic convergence.

ITER	ALPHA	||F||^2	||d^k||
0	0.000000E+00	0.205000E+04	0.301353E+08
1	0.898145E-07	0.205000E+04	0.772399E+06
2	0.442969E-05	0.204999E+04	0.137827E+04
3	0.656100E+00	0.137884E+04	0.533635E+03
...			
14	0.100000E+01	0.485899E-04	0.165212E+00
15	0.100000E+01	0.710910E-07	0.678731E-02
16	0.100000E+01	0.108957E-12	0.842304E-05
17	0.100000E+01	0.271474E-24	0.130452E-10

Again, the same optimal control problem was solved alternatively by a direct discretization method as in Gerdts [10] with Euler discretization and $N = 100$ subintervals. Herein, for better comparableness the control constraints $-1 \leq u \leq 1$ are not viewed as simple box constraints but are treated algorithmically as two nonlinear mixed control-state constraints. For the direct method the overall CPU time was 2.41 CPU seconds. As mentioned before, the CPU time grows at a nonlinear rate with N. Again, if all regularity assumptions are fulfilled, the nonsmooth Newton method turns out to be extremely efficient.

Acknowledgments. The author thanks the anonymous referees for very detailed and helpful comments and suggestions that helped to improve the paper.

REFERENCES

[1] W. ALT AND K. MALANOWSKI, *The Lagrange-Newton method for nonlinear optimal control problems*, Comput. Optim. Appl., 2 (1993), pp. 77–100.
[2] W. ALT AND K. MALANOWSKI, *The Lagrange-Newton method for state constrained optimal control problems*, Comput. Optim. Appl., 4 (1995), pp. 217–239.
[3] C. BÜSKENS, *Optimierungsmethoden und Sensitivitätsanalyse für Optimale Steuerprozesse mit Steuer- und Zustandsbeschränkungen*, Ph.D. thesis, Fachbereich Mathematik, Westfälische Wilhems-Universität Münster, Münster, Germany, 1998.
[4] X. CHEN, Z. NASHED, AND L. QI, *Smoothing methods and semismooth methods for nondifferentiable operator equations*, SIAM J. Numer. Anal., 38 (2000), pp. 1200–1216.
[5] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley & Sons, New York, 1983.
[6] A. L. DONTCHEV, W. W. HAGER, AND K. MALANOWSKI, *Error bounds for Euler approximation of a state and control constrained optimal control problem*, Numer. Funct. Anal. Optim., 21 (2000), pp. 653–682.
[7] A. L. DONTCHEV, W. W. HAGER, AND V. M. VELIOV, *Second-order Runge–Kutta approximations in control constrained optimal control*, SIAM J. Numer. Anal., 38 (2000), pp. 202–226.
[8] A. FISCHER, *A special Newton-type optimization method*, Optimization, 24 (1992), pp. 269–284.
[9] A. FISCHER, *Solution of monotone complementarity problems with locally Lipschitzian functions*, Math. Programming., 76 (1997), pp. 513–532.
[10] M. GERDTS, *Direct shooting method for the numerical soluti-of higher index DAE optimal control problems*, J. Optim. Theory Appl., 117 (2003), pp. 267–294.
[11] M. GERDTS, *A nonsmooth Newton's method for control-state constrained optimal control problems*, in Proceedings of the 5th Annual Vienna Symposium on Mathematical Modelling, ARGESIM Reports 30, Vienna, 2006.
[12] M. GERDTS, *Optimal Control of Ordinary Differential Equations and Differential-Algebraic Equations*, Habilitation thesis, Universität Bayreuth, Bayreuth, 2006; also available online from http://web.mat.bham.ac.uk/M.Gerdts/habilitation.pdf.
[13] M. GRÖTSCHEL, S. O. KRUMKE, AND J. RAMBAU, EDS., *Online Optimization of Large Scale Systems*, Springer, Berlin, 2001.

[14] W. W. Hager, *The dual active set algorithm and its application to linear programming*, SIAM J. Control Optim., 17 (1979), pp. 321–338.
[15] W. W. Hager, *Runge–Kutta methods in optimal control and the transformed adjoint system*, Numer. Math., 87 (2000), pp. 247–282.
[16] R. F. Hartl, S. P. Sethi, and G. Vickson, *A survey of the maximum principles for optimal control problems with state constraints*, SIAM Rev., 37 (1995), pp. 181–218.
[17] A. D. Ioffe and V. M. Tihomirov, *Theory of Extremal Problems*, in Stud. Math. Appl. 6, North–Holland, Amsterdam, 1979.
[18] H. Jiang, *Global convergence analysis of the generalized Newton and Gauss–Newton methods of the Fischer–Burmeister equation for the complementarity problem*, Math. Oper. Res., 24 (1999), pp. 529–543.
[19] B. Kummer, *Newton's method for non-differentiable functions*, in Advances in Mathematical Optimization, J. Guddat et al., eds., Akademie-Verlag, Berlin, 1988, pp. 171–194.
[20] B. Kummer, *Newton's method based on generalized derivatives for nonsmooth functions: Convergence analysis*, in Advances in Optimization, W. Oettli and D. Pallaschke, eds., 1991, Springer, Berlin, pp. 171–194.
[21] L. A. Ljusternik and W. I. Sobolew, *Elemente der Funktionalanalysis*, Verlag Harri Deutsch, Zürich, 1976.
[22] K. Malanowski, *On normality of Lagrange multipliers for state constrained optimal control problems*, Optimization, 52 (2003), pp. 75–91.
[23] K. Malanowski, C. Büskens, and H. Maurer, *Convergence of approximations to nonlinear optimal control problems*, in Mathematical Programming with Data Perturbations, Anthony Fiacco, ed., Lecture Notes in Pure and Appl. Math. 195, Marcel Dekker, New York, 1997, pp. 253–284.
[24] K. Malanowski and H. Maurer, *Sensitivity analysis for parametric control problems with control-state constraints*, Comput. Optim. Appl., 5 (1996), pp. 253–283.
[25] K. Malanowski, H. Maurer, and S. Pickenhain, *Second-order sufficient conditions for state-constrained optimal control problems*, J. Optim. Theory Appl., 123 (2004), pp. 595–617.
[26] H. Maurer and D. Augustin, *Sensitivity Analysis and Real-Time Control of Parametric Optimal Control Problems Using Boundary Value Methods*, in Online Optimization of Large Scale Systems, M. Grötschel, S. O. Krumke, and J. Rambau, eds., Springer, New York, 2001, pp. 17–55.
[27] L. W. Neustadt, *Optimization: A Theory of Necessary Conditions*, Princeton University Press, Princeton, NJ, 1976.
[28] H. J. Oberle and W. Grimm, *Bndsco—A Program for the Numerical Solution of Optimal Control Problems*, Tech. Report Reihe B, Bericht 36, Hamburger Beiträge zur Angewandten Mathematik, Department of Mathematics, University of Hamburg, Hamburg, Germany; also available online from http://www.math.uni-hamburg.de/home/oberle/software.html, 2001.
[29] L. Qi, *Convergence analysis of some algorithms for solving nonsmooth equations*, Math. Oper. Res., 18 (1993), pp. 227–244.
[30] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Program., 58 (1993), pp. 353–367.
[31] M. Ulbrich, *Nonsmooth Newton-Like Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*, Habilitation, Technical University of Munich, Munich, Germany, 2002.
[32] M. Ulbrich, *Semismooth Newton methods for operator equations in function spaces*, SIAM J. Optim., 13 (2003), pp. 805–841.
[33] D. Werner, *Funktionalanalysis*, Springer, Berlin, 1995.
[34] V. Zeidan, *The Riccati equation for optimal control problems with mixed state-control constraints: Necessity and sufficiency*, SIAM J. Control Optim., 32 (1994), pp. 1297–1321.

SIAM J. OPTIM.
Vol. 19, No. 1, pp. 351–369

AN INEXACT SQP METHOD FOR EQUALITY CONSTRAINED OPTIMIZATION*

RICHARD H. BYRD†, FRANK E. CURTIS‡, AND JORGE NOCEDAL§

Abstract. We present an algorithm for large-scale equality constrained optimization. The method is based on a characterization of inexact sequential quadratic programming (SQP) steps that can ensure global convergence. Inexact SQP methods are needed for large-scale applications for which the iteration matrix cannot be explicitly formed or factored and the arising linear systems must be solved using iterative linear algebra techniques. We address how to determine when a given inexact step makes sufficient progress toward a solution of the nonlinear program, as measured by an exact penalty function. The method is globalized by a line search. An analysis of the global convergence properties of the algorithm and numerical results are presented.

Key words. large-scale optimization, constrained optimization, sequential quadratic programming, inexact linear system solvers, Krylov subspace methods

AMS subject classifications. 49M37, 65K05, 90C06, 90C30, 90C55

DOI. 10.1137/060674004

1. Introduction.

In this paper we discuss an algorithm for equality constrained optimization problems of the form

$$\begin{aligned} &\min_{x\in\mathbb{R}^n} \quad f(x) \\ &\text{subject to (s.t.)} \quad c(x) = 0, \end{aligned} \tag{1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $c : \mathbb{R}^n \to \mathbb{R}^t$ are smooth nonlinear functions. Our interest is in methods for very large problems with $t \leq n$ for which the exact computation of steps in contemporary methods can be prohibitively expensive. One class of problems of this type that demands algorithmic improvements are those where the constraint functions are defined by systems of partial differential equations (PDEs).

One of the leading methods for solving constrained optimization problems is sequential quadratic programming (SQP). (In fact, modern interior point methods reduce to SQP when inequality constraints are not present in the problem formulation [18].) Algorithms in this class enjoy global convergence guarantees and typically require few iterations and function evaluations to locate a solution point. A drawback of many contemporary SQP algorithms, however, is that they require explicit representations of exact derivative information and the solution of one or more linear systems during every iteration. The acquisition of these quantities is particularly cumbersome in large-scale settings and the factorization of large iteration matrices is often impractical.

*Received by the editors November 2, 2006; accepted for publication (in revised form) September 8, 2007; published electronically April 16, 2008.
http://www.siam.org/journals/siopt/19-1/67400.html

†Department of Computer Science, University of Colorado, Boulder, CO 80309 (richard@cs.colorado.edu). This author was supported by Army Research Office grant DAAD19-02-1-0407 and by National Science Foundation grants CCR-0219190 and CHE-0205170.

‡Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208-3118 (fecurt@gmail.com). This author was supported by Department of Energy grant DE-FG02-87ER25047-A004.

§Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208-3118 (nocedal@dario.ece.northwestern.edu). This author was supported by National Science Foundation grant CCF-0514772 and by a grant from the Intel Corporation.

One way to overcome these difficulties is to solve the SQP subproblems approximately using iterative linear algebra techniques. The main purpose of this paper is to determine the accuracy with which the SQP subproblems must be solved in order to ensure global convergence in the context of a practical algorithm for problem (1.1). We both propose such a method and analyze its global behavior.

Our method resembles those in the class of inexact Newton methods for solving nonlinear systems of equations. There are, however, important differences between the two approaches. Inexact Newton methods for systems of equations are controlled by forcing parameters that ensure that the norm of the entire residual of the Newton equations decreases at every iteration [8]. Our approach, on the other hand, is based on a requirement that the step decreases a local approximation of a merit function while also satisfying bounds on the primal and dual components of the residual. We present sets of easily calculable conditions that handle these two components of the residual as separate quantities when determining if a given inexact solution is appropriate for the algorithm to follow. Such a solution may, for example, allow for an increase in the residual corresponding to primal feasibility provided it yields a substantial decrease in dual feasibility, or vice versa. The behavior of these components also helps determine when it is appropriate to increase the penalty parameter in the merit function.

A variety of methods for constrained optimization with inexactness in step computations have been proposed. Jäger and Sachs [14] describe an inexact reduced SQP method in Hilbert space. Lalee, Nocedal, and Plantenga [16], Byrd, Hribar, and Nocedal [5], and Heinkenschloss and Vicente [13] propose composite step trust region approaches where the step is computed as an approximate solution to an SQP subproblem. Similarly, Walther [22] provides a composite step method that allows incomplete constraint Jacobian information. Leibfritz and Sachs [17] analyze an interior point method that benefits from a reformulation of the quadratic programming subproblems as mixed linear complementarity problems. Our approach has some features in common with the algorithms of Biros and Ghattas [1, 2], Haber and Ascher [11], and Prudencio, Byrd, and Cai [20] as we follow a full space SQP method and perform a line search to promote convergence. Unlike these papers, however, we present conditions that guarantee the global convergence of inexact SQP steps.

This paper is organized as follows. In section 2 we provide an overview of our approach and globalization strategy. Section 3 contains details about the most crucial aspect of our algorithm, namely, the sets of conditions used to determine if a given inexact SQP solution is considered an acceptable step. The well-posedness of our approach is also discussed, the accountability of which allows us to present global convergence guarantees under common conditions in section 4. Section 5 provides numerical results to illustrate the robustness of our method. We focus on problems for which overall algorithm performance has been seen to be sensitive to the quality of inexact subproblem solutions. Closing remarks and issues related to extensions of this work are presented in section 6.

2. Outline of the algorithm. Let us formalize a basic SQP approach before clarifying the novelties of our algorithm. The Lagrangian function corresponding to problem (1.1) is

$$\mathcal{L}(x, \lambda) \triangleq f(x) + \lambda^T c(x), \tag{2.1}$$

where $\lambda \in \mathbb{R}^t$ are Lagrange multipliers. If f and c are continuously differentiable, then the first-order optimality conditions for x^* to be an optimal solution to problem (1.1)

state that there exist multipliers λ^* such that (x^*, λ^*) is a solution to the nonlinear system of equations

$$\nabla \mathcal{L}(x, \lambda) = \begin{bmatrix} g(x) + A(x)^T \lambda \\ c(x) \end{bmatrix} = 0, \tag{2.2}$$

where $g(x)$ is the gradient of the objective function and $A(x)$ is the Jacobian of $c(x)$. The components in (x, λ) are referred to as the primal and dual variables, respectively.

An SQP algorithm defines an appropriate displacement d_k in the primal space from the kth iterate x_k as the minimizer of a quadratic model of the objective subject to a linearization of the constraints. The quadratic program can be defined as

$$\begin{aligned} \min_{d \in \mathbb{R}^n} \quad & f(x_k) + g(x_k)^T d + \tfrac{1}{2} d^T W(x_k, \lambda_k) d \\ \text{s.t.} \quad & c(x_k) + A(x_k) d = 0, \end{aligned} \tag{2.3}$$

where

$$W(x, \lambda) \approx \nabla^2_{xx} \mathcal{L}(x, \lambda) = \nabla^2_{xx} f(x) + \sum_{i=1}^{t} \lambda^i \nabla^2_{xx} c^i(x)$$

is equal to, or is a symmetric approximation for, the Hessian of the Lagrangian. Here, $c^i(x)$ and λ^i denote the ith constraint function and its corresponding dual variable, respectively. If the constraint Jacobian $A(x_k)$ has full row rank and $W(x_k, \lambda_k)$ is positive definite on the null space of $A(x_k)$, then a solution to (2.3) is well defined in this context. An alternative characterization of the SQP step d_k is given by the fact that it can equivalently be obtained under similar assumptions as part of the solution to the primal-dual system (see [18])

$$\begin{bmatrix} W(x_k, \lambda_k) & A(x_k)^T \\ A(x_k) & 0 \end{bmatrix} \begin{bmatrix} d_k \\ \delta_k \end{bmatrix} = - \begin{bmatrix} g(x_k) + A(x_k)^T \lambda_k \\ c(x_k) \end{bmatrix} \tag{2.4}$$

constructed by applying Newton's method to (2.2).

An explicit representation of the primal-dual matrix

$$\begin{bmatrix} W(x_k, \lambda_k) & A(x_k)^T \\ A(x_k) & 0 \end{bmatrix} \tag{2.5}$$

and an exact solution of (2.4) can be expensive to obtain, particularly when the factors of (2.5) are not very sparse. We are interested, therefore, in identifying inexact solutions of (2.4) that can also be considered appropriate steps for the algorithm to accept during a given iteration. Such inexact solutions can be obtained in a variety of ways, such as by applying an iterative linear system solver to the primal-dual system. Regardless of the method chosen, for an inexact solution (d_k, δ_k) we define the residual vectors (ρ_k, r_k) by the equation

$$\begin{bmatrix} W(x_k, \lambda_k) & A(x_k)^T \\ A(x_k) & 0 \end{bmatrix} \begin{bmatrix} d_k \\ \delta_k \end{bmatrix} = - \begin{bmatrix} g(x_k) + A(x_k)^T \lambda_k \\ c(x_k) \end{bmatrix} + \begin{bmatrix} \rho_k \\ r_k \end{bmatrix}. \tag{2.6}$$

The step can then be appraised based on properties of the residual vector and other quantities related to the SQP subproblem formulation (2.3). For convex problems, an inexact Newton method intended for nonlinear equations will suffice, provided

that $W(x_k, \lambda_k)$ is the exact Hessian of the Lagrangian [8]. That is, the norm of the right-hand-side vector in (2.4) can serve as a merit function, and convergence can be guaranteed by systematically decreasing this value. For nonconvex problems, however, a step that decreases the first-order optimality error may move away from a minimizer, or may be trapped near a stationary point of the Lagrangian. Thus, merit functions more appropriate to constrained optimization should be considered.

We now outline the algorithm and globalization strategy that will be developed in detail in the following sections. An integral part of the approach is the mechanism used to determine if a trial primal-dual solution (d, δ) to (2.4) is acceptable during a given iteration. For this purpose, we make use of the merit function

$$\phi(x; \pi) \triangleq f(x) + \pi \|c(x)\|, \tag{2.7}$$

where $\pi > 0$ is known as the penalty parameter and $\|\cdot\|$ denotes a norm on $\mathbb{R}^t$. We observe that $\phi(x; \pi)$ is not continuously differentiable, but it is exact in the sense that if π is greater than a certain threshold, then a first-order optimal point of (1.1) is a stationary point of $\phi(x; \pi)$. That is, the directional derivative of $\phi(x; \pi)$ in a direction d, denoted by $D\phi(d; \pi)$, is nonnegative at x^* for all $d \in \mathbb{R}^n$. The challenge is to compute inexact SQP steps and a value for π that ensure progress in the merit function $\phi(x; \pi)$ during every iteration.

Upon the calculation and acceptance of the search direction d_k for a particular value π_k of the penalty parameter, we perform a backtracking line search to compute a steplength coefficient α_k satisfying the Armijo condition

$$\phi(x_k + \alpha_k d_k; \pi_k) \leq \phi(x_k; \pi_k) + \eta \alpha_k D\phi(d_k; \pi_k) \tag{2.8}$$

for some $0 < \eta < 1$. Accordingly, a primal-dual step will be accepted only if its primal component is a descent direction for the merit function.

In summary, our approach follows a standard line search SQP framework. During each iteration, a step is computed as an inexact solution to the primal-dual system (2.6) satisfying appropriate conditions that deem the step acceptable. The penalty parameter is then set based on properties of the computed step, after which a backtracking line search is performed to compute a steplength coefficient α_k satisfying the Armijo condition (2.8). Finally, the iterate is updated along with function and derivative information at the new point. The novelty of our approach, i.e., the precise definition of what constitutes an acceptable step, and the convergence properties of this algorithm are considered in the remainder of this paper.

We drop functional notation throughout the rest of the paper when values are clear from the context and delimit iteration number information for functions as with variables; i.e., we denote $g_k \triangleq g(x_k)$ and similarly for other quantities. All norms are considered Euclidean (or l_2) norms unless otherwise indicated, though much of our analysis will apply for any norm.

3. Step computation and selection. An intuitive condition that one may impose on an inexact SQP step is that the directional derivative of the merit function along the primal component d_k must be sufficiently negative. Such a condition could be used in the development of a globally convergent SQP approach, but quantifying an appropriate steepness of the directional derivative is a difficult task in practice.

As an alternative, let us borrow from an approach commonly employed in trust region methods that begins by considering a local model of the merit function $\phi(x; \pi)$ around the current iterate x_k and the changes in the merit function it predicts for

steps in the primal space. The model has the form

$$m_k(d;\pi) \triangleq f_k + g_k^T d + \max\{\tfrac{1}{2} d^T W_k d, 0\} + \pi \|c_k + A_k d\|,$$

where the max term yields a quadratic model of the objective or a linear one depending on the curvature of W_k along d. With this approximation, we can estimate the reduction in the merit function given by a step d_k by evaluating

$$\begin{aligned} \Delta m_k(d_k;\pi_k) &\triangleq m_k(0;\pi_k) - m_k(d_k;\pi_k) \\ &= -g_k^T d_k - \max\{\tfrac{1}{2} d_k^T W_k d_k, 0\} + \pi_k(\|c_k\| - \|c_k + A_k d_k\|) \\ &= -g_k^T d_k - \max\{\tfrac{1}{2} d_k^T W_k d_k, 0\} + \pi_k(\|c_k\| - \|r_k\|), \end{aligned} \tag{3.1}$$

where the residual $r_k = c_k + A_k d_k$ as in (2.6).

At the heart of our approach is the claim that a given primal-dual step is often beneficial for the algorithm to follow provided the following condition is satisfied.

Model Reduction Condition. *A step (d_k, δ_k) computed in an inexact SQP algorithm must satisfy*

$$\Delta m_k(d_k;\pi_k) \geq \sigma \pi_k \max\{\|c_k\|, \|r_k\| - \|c_k\|\} \tag{3.2}$$

for some $0 < \sigma < 1$ and appropriate $\pi_k > 0$.

We will see the effects of this condition below and in section 4. In particular, (3.2) will indeed ensure that the directional derivative of the merit function is sufficiently negative along the primal step component d_k while also providing a mechanism for determining appropriate values of the penalty parameter. We note that conditions similar to the model reduction condition (3.2) are presented in the context of the inexact composite-step SQP algorithm proposed by Heinkenschloss and Vicente [13]. However, their conditions are applicable only to a step that has been decomposed into basic and nonbasic components, as accuracy is imposed on the components separately. Their approach also differs from the one treated here in that they use a trust region and assume that an approximate reduced Hessian is available.

3.1. Step acceptance conditions. An acceptable step will be required to satisfy one of two sets of conditions. We refer to the conditions as termination tests in reference to algorithms that apply an iterative solver to the primal-dual system (2.4), as in this framework the conditions are used to determine when to terminate the iteration. Each termination test will allow us to ensure that the step satisfies (3.2) for an appropriate value of the penalty parameter and enforces requirements on the residuals (ρ_k, r_k) to ensure convergence to a local solution of (1.1). In addition, the tests impose restrictions on when the algorithm is allowed to increase the penalty parameter in order to satisfy the model reduction condition (3.2).

The first termination test addresses those steps providing a sufficiently large reduction in the model of the merit function for the most recent value of the penalty parameter. We assume that an initial value $\pi_{-1} > 0$ is given.

Termination Test I. *Let $0 < \sigma, \kappa < 1$ be given constants. A step (d_k, δ_k) computed in an inexact SQP algorithm is acceptable if the model reduction condition* (3.2) *holds for $\pi_k = \pi_{k-1}$ and*

$$\left\| \begin{bmatrix} \rho_k \\ r_k \end{bmatrix} \right\| \leq \kappa \left\| \begin{bmatrix} g_k + A_k^T \lambda_k \\ c_k \end{bmatrix} \right\| \tag{3.3}$$

for the residuals (ρ_k, r_k) defined by (2.6).

We claim that Termination Test I allows for productive steps to be taken that may have been computed in a relatively cheap manner, say, after only a few iterations of an iterative solver applied to the primal-dual system (2.4). For steps satisfying this test, given that a sufficient reduction in the model of the merit function has been obtained we need only enforce a generally loose bound on the residual vector. For even greater flexibility one can in fact choose $\kappa \geq 1$ in Termination Test I if the additional condition

$$\|\rho_k\| \leq \max\{\kappa_1 \|g_k + A_k^T \lambda_k\|, \kappa_2 \|c_k\|\}, \quad 0 < \kappa_1 < 1, \ 0 < \kappa_2, \tag{3.4}$$

is enforced. This may be useful, say, when applying our step acceptance criteria when steps are not computed directly via (2.4) or when the use of a left preconditioner for (2.4) produces steps corresponding to residuals larger in norm than the right-hand-side vector $(g_k + A_k^T \lambda_k, c_k)$. All the results in the following sections hold if Termination Test I has $\kappa < 1$ or if (3.4) is included when $\kappa \geq 1$.

The second termination test addresses those steps providing a sufficiently large reduction in the linear model of the constraints.

Termination Test II. *Let $0 < \epsilon < 1$ and $0 < \beta$ be given constants. A step (d_k, δ_k) computed in an inexact SQP algorithm is acceptable if*

$$\|r_k\| \leq \epsilon \|c_k\| \tag{3.5a}$$

$$\text{and} \quad \|\rho_k\| \leq \beta \|c_k\|, \tag{3.5b}$$

where the residuals (ρ_k, r_k) are defined by (2.6).

A step satisfying Termination Test II may not satisfy the model reduction condition (3.2) for $\pi_k = \pi_{k-1}$. Thus, for such steps we require that the penalty parameter be increased to satisfy

$$\pi_k \geq \frac{g_k^T d_k + \max\{\frac{1}{2} d_k^T W_k d_k, 0\}}{(1-\tau)(\|c_k\| - \|r_k\|)} \triangleq \pi_k^{trial} \tag{3.6}$$

for a given $0 < \tau < 1$. Notice from (3.5a) and $0 < \epsilon < 1$ that the denominator in the above expression is positive and along with (3.1) the rule (3.6) implies

$$\Delta m_k(d_k; \pi_k) \geq \tau \pi_k (\|c_k\| - \|r_k\|) \geq \tau(1-\epsilon)\pi_k \|c_k\|. \tag{3.7}$$

Therefore, when (3.5a) is satisfied, the model reduction condition (3.2) holds with $\sigma = \tau(1-\epsilon)$.

In summary, a step (d_k, δ_k) will be required to satisfy Termination Test I or II. In each case, the model reduction condition (3.2) will hold; Termination Test I demands it explicitly and the rule (3.6) is used to enforce it when Termination Test II is satisfied. For consistency between Termination Test I and II and (3.6), one should set $\sigma = \tau(1-\epsilon)$ for Termination Test I.

The complete algorithm is the following. We refer to our step acceptance criteria as SMART tests because they can be characterized as sufficient merit function approximation reduction termination tests.

Algorithm A: Inexact SQP with SMART Tests.
Given parameters $0 < \kappa, \epsilon, \tau, \sigma, \eta < 1$ and $\beta > 0$
Initialize x_0, λ_0, and $\pi_{-1} > 0$
for $k = 0, 1, 2, \ldots$, until a convergence test for (1.1) is satisfied
Compute f_k, g_k, c_k, W_k, and A_k and set $\pi_k \leftarrow \pi_{k-1}$ and $\alpha_k \leftarrow 1$

Compute a step (d_k, δ_k) satisfying Termination Test I or II
if Termination Test II is satisfied and (3.6) does not hold, set $\pi_k \leftarrow \pi_k^{trial} + 10^{-4}$
Perform a backtracking line search to obtain α_k satisfying (2.8)
Set $(x_{k+1}, \lambda_{k+1}) \leftarrow (x_k, \lambda_k) + \alpha_k(d_k, \delta_k)$
endfor

In practice, the step can be computed by producing a sequence of candidate steps $\{(d, \delta)\}$ via the application of an iterative solver to (2.4). The corresponding residuals $\{(\rho, r)\}$ can then be computed and Termination Tests I and II can be evaluated during each iteration or after a few steps of the iterative solver. The constants $(\kappa, \epsilon, \beta)$ should be tuned for a specific application and can significantly influence the practical performance of the algorithm. In particular, the value for β should be chosen to reflect the relationship between the scales of the primal and dual feasibility measures. The scale dependence of such a parameter is not ideal, but a bound similar to (3.5b) is used to ensure the boundedness of the penalty parameter π_k (as we show in Lemma 4.7) if the rule (3.6) is enforced. Since such a method for setting the penalty parameter has proved to work well in practice [23], we employ this update rule in the algorithms in this paper and define β and (3.5b) as given. The constants (τ, σ, η) generally can be set to default values, or, in the case of σ, to promote consistency between Termination Tests I and II. Further discussion of appropriate values for the constants and an example implementation of Algorithm A are given in section 5.

3.2. Well-posedness of the algorithm. It is important to verify that the iterates specified by Algorithm A can be always be computed in practice.

Suppose that (x_k, λ_k) is an iterate that does not satisfy the optimality conditions (2.2). We argue here that whenever A_k has full row rank and W_k is positive definite on the null space of A_k, a sufficiently accurate solution to (2.4) will satisfy either Termination Test I or II. If $c(x_k) \neq 0$, then for (ρ_k, r_k) sufficiently small we have that (3.5), and thus Termination Test II, will be satisfied. Otherwise, if $c(x_k) = 0$, then (3.3) will be satisfied for (ρ_k, r_k) sufficiently small. Then, since W_k is positive definite on the null space of A_k, the solution of (2.4) is the solution to problem (2.3), which means that the solution lies in the null space of A_k and corresponds to a nonpositive objective value of (2.3) (since $d = 0$ is feasible). Therefore, by computing a step with (ρ_k, r_k) sufficiently small, it can easily be seen that (3.2), and thus Termination Test I, will be satisfied.

Once an acceptable step is obtained, we must ensure that a positive steplength parameter α_k can be calculated to satisfy the Armijo condition (2.8). We consider this issue by first presenting the following result.

LEMMA 3.1. *The directional derivative of the merit function $\phi(x; \pi)$ along a step d satisfies*

$$D\phi(d; \pi) \leq g^T d - \pi(\|c\| - \|r\|).$$

Proof. Applying Taylor's theorem, we find for some constant $\gamma_1 > 0$

$$\begin{aligned}
\phi(x + \alpha d; \pi) - \phi(x; \pi) &= f(x + \alpha d) - f(x) + \pi(\|c(x + \alpha d)\| - \|c(x)\|) \\
&\leq \alpha g^T d + \gamma_1 \pi \alpha^2 \|d\|^2 + \pi(\|c(x) + \alpha A d\| - \|c(x)\|) \\
&= \alpha g^T d + \gamma_1 \pi \alpha^2 \|d\|^2 + \pi(\|(1 - \alpha)c(x) + \alpha r\| - \|c(x)\|) \\
&\leq \alpha(g^T d - \pi(\|c(x)\| - \|r\|)) + \gamma_1 \pi \alpha^2 \|d\|^2,
\end{aligned}$$

where $r = c(x) + Ad$ as in (2.6). Dividing both sides by α and taking the limit as $\alpha \to 0$ yields the result. ∎

Given this result, we present the following consequence of our model reduction condition. (A stronger result will be given as Lemma 4.6.)

LEMMA 3.2. *If the model reduction condition* (3.2) *holds for a step* (d_k, δ_k) *and penalty parameter* π_k*, then the directional derivative of the merit function satisfies* $D\phi(d_k; \pi_k) \leq 0$.

Proof. Observe from (3.1) that the inequality (3.2) can be rewritten as

$$g_k^T d_k - \pi_k(\|c_k\| - \|r_k\|) \leq -\max\{\tfrac{1}{2} d_k^T W_k d_k, 0\} - \sigma \pi_k \max\{\|c_k\|, \|r_k\| - \|c_k\|\},$$

so, by Lemma 3.1, a step (d_k, δ_k) satisfying (3.2) yields

$$\begin{aligned} D\phi(d_k; \pi_k) &\leq g_k^T d_k - \pi_k(\|c_k\| - \|r_k\|) \\ &\leq -\max\{\tfrac{1}{2} d_k^T W_k d_k, 0\} - \sigma \pi_k \max\{\|c_k\|, \|r_k\| - \|c_k\|\}, \end{aligned} \tag{3.8}$$

which yields the result. ∎

We have shown under common conditions that an acceptable inexact SQP step (d_k, δ_k) can always be computed by Algorithm A and that steps satisfying the model reduction condition (3.2) correspond to directions of nonincrease for the merit function $\phi(x; \pi_k)$. These results allow us to show that the Armijo condition (2.8) is satisfied by some positive α_k (see Lemma 4.8), and so Algorithm A is well-posed.

We mention in passing that, as a corollary to Lemma 3.1, we may avoid the exact computation of the directional derivative of the merit function along a step d by defining the estimate

$$\tilde{D}\phi(d; \pi) \triangleq g^T d - \pi(\|c\| - \|r\|). \tag{3.9}$$

As such, the Armijo condition (2.8) can be substituted by

$$\phi(x_k + \alpha_k d_k; \pi_k) \leq \phi(x_k; \pi_k) + \eta \alpha_k \tilde{D}\phi(d_k; \pi_k). \tag{3.10}$$

All the analysis in this paper holds when either (2.8) or (3.10) is observed in the line search procedure of Algorithm A. For convenience, we choose to use (3.10).

4. Global analysis. Let us begin our investigation of the global behavior of Algorithm A by making the following assumptions about the problem and the set of computed iterates.

ASSUMPTIONS 4.1. *The sequence* $\{x_k, \lambda_k\}$ *generated by Algorithm* A *is contained in a convex set* Ω *and the following properties hold:*

(a) *The functions* f *and* c *and their first and second derivatives are bounded on* Ω.

(b) *The sequence* $\{\lambda_k\}$ *is bounded.*

(c) *The constraint Jacobians* A_k *have full row rank and their smallest singular values are bounded below by a positive constant.*

(d) *The sequence* $\{W_k\}$ *is bounded.*

(e) *There exists a positive constant* $\mu > 0$ *such that for any* $u \in \mathbb{R}^n$ *with* $u \neq 0$ *and* $A_k u = 0$ *we have* $u^T W_k u \geq \mu \|u\|^2$.

These assumptions are fairly standard for a line search method [12, 19]. Assumption 4.1(a) is a little weaker than the common assumption that the iterates are contained in a compact set. Assumptions 4.1(b) and (c) are strong; we use them to simplify the analysis in order to focus on the issues related to inexactness. It would be of interest in future studies of inexact SQP methods to relax these assumptions. Assuming that W_k is positive definite on the null space of the constraints is natural

for line search algorithms, for otherwise there would be no guarantee of descent. We comment further on the validity of Assumption 4.1(b) in section 6.

We now assume that during iteration k we have obtained an acceptable step (d_k, δ_k) with residuals (ρ_k, r_k) defined by (2.6). We consider the decomposition

$$d_k = u_k + v_k, \tag{4.1}$$

where u_k lies in the null space of the constraint Jacobian A_k and v_k lies in the range space of A_k^T. We do not intend to compute the components explicitly; the decomposition is only for analytical purposes [4, 6]. We refer to u_k, which by definition satisfies $A_k u_k = 0$, as the tangential component and v_k as the normal component of the step.

Our analysis hinges on our ability to classify the effects of two types of steps: those lying sufficiently in the null space of the constraints and those sufficiently orthogonal to the linearized feasible region. We show that such a distinction can be made by observing the relative magnitudes of the normal and tangential components of the primal component d_k.

We first present a result related to the magnitude of the normal step.

LEMMA 4.2. *For all k, the normal component v_k is bounded in norm and for some $\gamma_2 > 0$ satisfies*

$$\|v_k\|^2 \le \gamma_2 \max\{\|c_k\|, \|r_k\|\}. \tag{4.2}$$

Furthermore, for all k such that Termination Test II *is satisfied, there exists $\gamma_3 > 0$ such that*

$$\|v_k\| \le \gamma_3(\|c_k\| - \|r_k\|). \tag{4.3}$$

Proof. From $A_k v_k = -c_k + r_k$ and the fact that v_k lies in the range space of A_k^T, it follows that

$$v_k = A_k^T(A_k A_k^T)^{-1}(-c_k + r_k),$$

and so

$$\|v_k\| \le \|A_k^T(A_k A_k^T)^{-1}\|(\|c_k\| + \|r_k\|). \tag{4.4}$$

This, along with (3.3), the fact that Assumptions 4.1(a) and (b) imply that $\|c_k\|$ and $\|g_k + A_k^T\lambda_k\|$ are bounded, and the fact that Assumptions 4.1(a) and (c) imply that $\|A_k^T(A_k A_k^T)^{-1}\|$ is bounded, implies v_k is bounded in norm for all k. The inequality (4.4) also yields

$$\begin{aligned} \|v_k\|^2 &\le \left(\|A_k^T(A_k A_k^T)^{-1}\|(\|c_k\| + \|r_k\|)\right)^2 \\ &\le \left(2\|A_k^T(A_k A_k^T)^{-1}\| \max\{\|c_k\|, \|r_k\|\}\right)^2 \\ &= \left[4\|A_k^T(A_k A_k^T)^{-1}\|^2 \max\{\|c_k\|, \|r_k\|\}\right] \max\{\|c_k\|, \|r_k\|\}, \end{aligned} \tag{4.5}$$

where (3.3) and Assumptions 4.1(a), (b), and (c) also imply that the bracketed expression in (4.5) is bounded. Thus, (4.2) holds. Finally, if Termination Test II is satisfied, then from (3.5a) and (4.4) we have

$$\begin{aligned} \|v_k\| &\le \|A_k^T(A_k A_k^T)^{-1}\|(1+\epsilon)\|c_k\| \\ &\le \|A_k^T(A_k A_k^T)^{-1}\| \left(\tfrac{1+\epsilon}{1-\epsilon}\right)(\|c_k\| - \|r_k\|), \end{aligned}$$

and so (4.3) holds. ∎

A similar result can be proved for the tangential component.

LEMMA 4.3. *The tangential components* $\{u_k\}$ *are bounded in norm.*

Proof. Assumption 4.1(e), the fact that u_k lies in the null space of the constraint Jacobian A_k, and the first block equation of (2.6) yield

$$\begin{aligned}\mu\|u_k\|^2 &\leq u_k^T W_k u_k \\ &= -g_k^T u_k + \rho_k^T u_k - u_k^T W_k v_k \\ &\leq (\|g_k\| + \|\rho_k\| + \|W_k v_k\|)\|u_k\|,\end{aligned}$$

and so

$$\|u_k\| \leq (\|g_k\| + \|\rho_k\| + \|W_k v_k\|)/\mu.$$

The result follows from the facts that Assumptions 4.1, Lemma 4.2, and the bounds (3.3) and (3.5b) imply that all terms in the right-hand side of this inequality are bounded. ∎

We now turn to the following result addressing the relative magnitudes of the normal and tangential components of a given step.

LEMMA 4.4. *There exists a constant* $\gamma_4 > 0$ *such that if* $\|u_k\|^2 \geq \gamma_4\|v_k\|^2$*, then* $\frac{1}{2}d_k^T W_k d_k \geq \frac{\mu}{4}\|u_k\|^2$.

Proof. Assumption 4.1(e) implies that for any γ_4 such that $\|u_k\|^2 \geq \gamma_4\|v_k\|^2$ we have

$$\begin{aligned}\tfrac{1}{2}d_k^T W_k d_k &= \tfrac{1}{2}u_k^T W_k u_k + u_k^T W_k v_k + \tfrac{1}{2}v_k^T W_k v_k \\ &\geq \tfrac{\mu}{2}\|u_k\|^2 - \|u_k\|\|W_k\|\|v_k\| - \tfrac{1}{2}\|W_k\|\|v_k\|^2 \\ &\geq \left(\frac{\mu}{2} - \frac{\|W_k\|}{\sqrt{\gamma_4}} - \frac{\|W_k\|}{2\gamma_4}\right)\|u_k\|^2.\end{aligned}$$

Thus, Assumption 4.1(d) implies the result holds for a sufficiently large $\gamma_4 > 0$. ∎

With the above results, we can now formalize a distinction between two types of steps. Let $\gamma_4 > 0$ be chosen large enough as described in Lemma 4.4 and consider the sets of indices

$$\begin{aligned}K_1 &\triangleq \{k : \|u_k\|^2 \geq \gamma_4\|v_k\|^2\} \\ \text{and}\quad K_2 &\triangleq \{k : \|u_k\|^2 < \gamma_4\|v_k\|^2\}.\end{aligned}$$

Most of the remainder of our analysis will be dependent on these sets and the corresponding quantity

$$\Theta_k \triangleq \begin{cases} \|u_k\|^2 + \|c_k\|, & k \in K_1, \\ \max\{\|c_k\|, \|r_k\|\}, & k \in K_2. \end{cases}$$

The relevance of Θ_k will be seen in the following three lemmas as a quantity that can be used for bounding the length of the primal step and the directional derivative of the merit function, which will then provide a lower bound for the sequence of steplength coefficients $\{\alpha_k\}$.

LEMMA 4.5. *There exists* $\gamma_5 > 1$ *such that for all* k*,*

$$\|d_k\|^2 \leq \gamma_5\Theta_k$$

and hence

$$\|d_k\|^2 + \|c_k\| \leq 2\gamma_5\Theta_k. \tag{4.6}$$

Proof. For $k \in K_1$, we find

$$\begin{aligned} \|d_k\|^2 &= \|u_k\|^2 + \|v_k\|^2 \\ &\le \left(1 + \tfrac{1}{\gamma_4}\right) \|u_k\|^2 \\ &\le \left(1 + \tfrac{1}{\gamma_4}\right) (\|u_k\|^2 + \|c_k\|). \end{aligned}$$

Similarly, Lemma 4.2 implies that for $k \in K_2$

$$\begin{aligned} \|d_k\|^2 &= \|u_k\|^2 + \|v_k\|^2 \\ &< (\gamma_4 + 1)\|v_k\|^2 \\ &\le (\gamma_4 + 1)\gamma_2 \max\{\|c_k\|, \|r_k\|\}. \end{aligned}$$

To establish (4.6) we note that $\Theta_k + \|c_k\| \le 2\Theta_k$ for all k. ∎

The directional derivative of the merit function can be bounded in a similar manner.

LEMMA 4.6. *There exists $\gamma_6 > 0$ such that for all k,*

$$\tilde{D}\phi(d_k; \pi_k) \le -\gamma_6 \Theta_k.$$

Proof. Recalling (3.8) and (3.9) we have

$$\tilde{D}\phi(d_k; \pi_k) \le -\max\{\tfrac{1}{2} d_k^T W_k d_k, 0\} - \sigma \pi_k \max\{\|c_k\|, \|r_k\| - \|c_k\|\}. \tag{4.7}$$

By Lemma 4.4, we have that $\frac{1}{2} d_k^T W_k d_k \ge \frac{\mu}{4} \|u_k\|^2$ for $k \in K_1$ and thus

$$\tilde{D}\phi(d_k; \pi_k) \le -\tfrac{\mu}{4} \|u_k\|^2 - \sigma \pi_k \|c_k\|.$$

Similarly, for $k \in K_2$ we have from (4.7) that

$$\begin{aligned} \tilde{D}\phi(d_k; \pi_k) &\le -\sigma \pi_k \max\{\|c_k\|, \|r_k\| - \|c_k\|\} \\ &\le -\tfrac{1}{2} \sigma \pi_k \max\{\|c_k\|, \|r_k\|\}. \end{aligned}$$

The result holds for $\gamma_6 = \min\{\frac{\mu}{4}, \frac{1}{2}\sigma\pi_k\}$, which is bounded away from zero as $\{\pi_k\}$ is nondecreasing. ∎

Another important property of Algorithm A is that under Assumptions 4.1 the penalty parameter remains bounded. We prove this result in the following lemma, illustrating the importance of the bound (3.5b).

LEMMA 4.7. *The sequence of penalty parameters $\{\pi_k\}$ is bounded above and $\pi_k = \pi_{\bar{k}}$ for all $k \ge \bar{k}$ for some $\bar{k} \ge 0$.*

Proof. Recall that the penalty parameter is increased during iteration k of Algorithm A only if Termination Test II is satisfied. Therefore, for the remainder of this proof let us assume that Termination Test II is satisfied and so the inequalities in (3.5) hold. By (3.6) the parameter π_k is chosen to satisfy the first inequality in (3.7), namely,

$$\Delta m_k(d_k; \pi_k) \ge \tau \pi_k (\|c_k\| - \|r_k\|), \tag{4.8}$$

where, according to the first block equation of (2.6), we can rewrite the model reduction as

$$\begin{aligned} \Delta m_k(d_k; \pi_k) &= \pi_k(\|c_k\| - \|r_k\|) - g_k^T d_k - \max\{\tfrac{1}{2} d_k^T W_k d_k, 0\} \\ &= \pi_k(\|c_k\| - \|r_k\|) \\ &\quad + \begin{cases} -g_k^T v_k - \frac{1}{2} v_k^T W_k v_k - \rho_k^T u_k + \frac{1}{2} u_k^T W_k u_k & \text{if } \frac{1}{2} d_k^T W_k d_k \ge 0 \\ -g_k^T v_k - (\rho_k - W_k v_k)^T u_k + u_k^T W_k u_k & \text{otherwise.} \end{cases} \end{aligned}$$

The result follows from our ability to bound the terms in the second line of this expression with respect to the constraint reduction.

If $\frac{1}{2}d_k^T W_k d_k \geq 0$, then under Assumptions 4.1 we have that Lemmas 4.2 and 4.3 and the bounds (3.5) on the residuals (ρ_k, r_k) imply that there exists $\gamma_7, \gamma_7' > 0$ such that

$$\begin{aligned} -g_k^T v_k - \tfrac{1}{2}v_k^T W_k v_k - \rho_k^T u_k + \tfrac{1}{2}u_k^T W_k u_k &\geq -\|g_k\|\|v_k\| - \tfrac{1}{2}\|W_k\|\|v_k\|^2 - \|\rho_k\|\|u_k\| \\ &\geq -\gamma_7(\|v_k\| + \|\rho_k\|) \\ &\geq -\gamma_7\left(\gamma_3 + \tfrac{\beta}{1-\epsilon}\right)(\|c_k\| - \|r_k\|) \\ &= -\gamma_7'(\|c_k\| - \|r_k\|). \end{aligned}$$

Similarly, if $\frac{1}{2}d_k^T W_k d_k < 0$, then under Assumptions 4.1 we again find that Lemmas 4.2 and 4.3 and the bounds (3.5) on the residuals (ρ_k, r_k) imply that there exists $\gamma_8, \gamma_8' > 0$ such that

$$\begin{aligned} -g_k^T v_k - (\rho_k - W_k v_k)^T u_k + u_k^T W_k u_k &\geq -\|g_k\|\|v_k\| - \|\rho_k\|\|u_k\| - \|v_k\|\|W_k\|\|u_k\| \\ &\geq -\gamma_8(\|v_k\| + \|\rho_k\|) \\ &\geq -\gamma_8\left(\gamma_3 + \tfrac{\beta}{1-\epsilon}\right)(\|c_k\| - \|r_k\|) \\ &= -\gamma_8'(\|c_k\| - \|r_k\|). \end{aligned}$$

These results together imply

$$\Delta m_k(d_k; \pi_k) \geq (\pi_k - \max\{\gamma_7', \gamma_8'\})(\|c_k\| - \|r_k\|),$$

and so (4.8) is always satisfied for

$$\pi_k \geq \max\{\gamma_7', \gamma_8'\}/(1-\tau).$$

Therefore, if $\pi_{\bar{k}} \geq \max\{\gamma_7', \gamma_8'\}/(1-\tau)$ for some $\bar{k} \geq 0$, then $\pi_k = \pi_{\bar{k}}$ for $k \geq \bar{k}$. This, along with the fact that whenever Algorithm A increases the penalty parameter it does so by at least a positive finite amount, proves the result. ∎

The previous three lemmas can be used to bound the sequence of steplength coefficients.

LEMMA 4.8. *The sequence $\{\alpha_k\}$ is bounded below and away from zero.*

Proof. Recall that the line search requires (3.10), which we rewrite for convenience as

$$\phi(x_k + \alpha_k d_k; \pi_k) - \phi(x_k; \pi_k) \leq \eta\alpha_k \tilde{D}\phi(d_k; \pi_k).$$

Suppose that the line search fails for some $\bar{\alpha} > 0$, so

$$\phi(x_k + \bar{\alpha} d_k; \pi_k) - \phi(x_k; \pi_k) > \eta\bar{\alpha} \tilde{D}\phi(d_k; \pi_k).$$

From the proof of Lemma 3.1 and (3.9) we have

$$\phi(x_k + \bar{\alpha} d_k; \pi_k) - \phi(x_k; \pi_k) \leq \bar{\alpha}\tilde{D}\phi(d_k; \pi_k) + \bar{\alpha}^2\gamma_1\pi_k\|d_k\|^2,$$

so

$$(\eta - 1)\tilde{D}\phi(d_k; \pi_k) \leq \bar{\alpha}\gamma_1\hat{\pi}\|d_k\|^2.$$

Here, $\hat{\pi}$ is a finite upper bound for the sequence $\{\pi_k\}$ whose existence follows from Lemma 4.7. Lemmas 4.5 and 4.6 then yield

$$(1-\eta)\gamma_6\Theta_k < \bar{\alpha}\gamma_1\gamma_5\hat{\pi}\Theta_k,$$

so

$$\bar{\alpha} > (1-\eta)\gamma_6/(\gamma_1\gamma_5\hat{\pi}).$$

Thus, α_k need never be set below $(1-\eta)\gamma_6/(\gamma_1\gamma_5\hat{\pi})$ for (3.10) to be satisfied. ∎

We can now present the following result related to the lengths of the primal components of the steps computed in Algorithm A and the convergence of the iterates toward the feasible region of problem (1.1).

LEMMA 4.9. *Algorithm* A *yields*

$$\lim_{k\to\infty}\|c_k\| = 0 \quad \textit{and} \quad \lim_{k\to\infty}\|d_k\| = 0.$$

Proof. For all k, it can easily be seen that

$$\phi(x_k;\pi_k) - \phi(x_k+\alpha_k d_k;\pi_k) \geq \gamma_9\Theta_k$$

for some $\gamma_9 > 0$ follows from (3.10) and Lemmas 4.6 and 4.8. By Lemma 4.7 the algorithm eventually computes, during iteration $\bar{k} \geq 0$, a finite value $\pi_{\bar{k}}$ beyond which the penalty parameter will never be increased. Therefore, the penalty parameter remains unchanged for $k \geq \bar{k}$, and for all $k > \bar{k}$ (4.6) implies

$$\begin{aligned}\phi(x_{\bar{k}};\pi_{\bar{k}}) - \phi(x_k;\pi_{\bar{k}}) &= \sum_{j=\bar{k}}^{k-1}(\phi(x_j;\pi_{\bar{k}}) - \phi(x_{j+1};\pi_{\bar{k}})) \\ &\geq \gamma_9\sum_{j=\bar{k}}^{k-1}\Theta_j \\ &\geq \frac{\gamma_9}{2\gamma_5}\sum_{j=\bar{k}}^{k-1}(\|d_j\|^2 + \|c_j\|).\end{aligned}$$

The result follows from the above and the fact that Assumption 4.1(a) implies $\phi(x;\pi_{\bar{k}})$ is bounded below. ∎

We are now ready to present the main result of this section.

THEOREM 4.10. *Algorithm* A *yields*

$$\lim_{k\to\infty}\left\|\begin{bmatrix} g_k + A_k^T\lambda_k \\ c_k \end{bmatrix}\right\| = 0.$$

Proof. Recall that $\alpha_k \leq 1$ for all k and from Lemma 4.8 we have that $\{\alpha_k\}$ is bounded below and away from zero. An expansion of the first block of the optimality conditions (2.2) yields

$$\|g_{k+1} + A_{k+1}^T\lambda_{k+1}\| \leq \|g_k + A_k^T\lambda_k + \alpha_k(\nabla_{xx}^2\mathcal{L}_k d_k + A_k^T\delta_k)\| + \alpha_k^2 E(d_k,\delta_k),$$

where

$$E(d_k,\delta_k) = O(\|d_k\|^2 + \|d_k\cdot\delta_k\|).$$

This, along with the first block equation in (2.6) and Assumptions 4.1, implies

$$\begin{aligned}
&\|g_{k+1} + A_{k+1}^T \lambda_{k+1}\| \\
&\quad \leq \|g_k + A_k^T \lambda_k + \alpha_k (W_k d_k + A_k^T \delta_k) + \alpha_k (\nabla_{xx}^2 \mathcal{L}_k - W_k) d_k\| + \alpha_k^2 E(d_k, \delta_k) \\
&\quad \leq \|g_k + A_k^T \lambda_k + \alpha_k (\rho_k - g_k - A_k^T \lambda_k)\| + \alpha_k \|(\nabla_{xx}^2 \mathcal{L}_k - W_k) d_k\| + \alpha_k^2 E(d_k, \delta_k) \\
&\quad \leq (1 - \alpha_k)\|g_k + A_k^T \lambda_k\| + \alpha_k \|\rho_k\| + \alpha_k E'(d_k, \delta_k),
\end{aligned} \tag{4.9}$$

where

$$E'(d_k, \delta_k) = O(\|d_k\| + \|d_k\|^2 + \|d_k \cdot \delta_k\|). \tag{4.10}$$

The bounds (3.3) and (3.5b) and the triangle inequality imply

$$\|\rho_k\| \leq \max\{\kappa(\|g_k + A_k^T \lambda_k\| + \|c_k\|), \beta \|c_k\|\},$$

which, along with (4.9) and the boundedness of $\{\alpha_k\}$, implies that for some $0 < \gamma_{10} < 1$ and $\gamma_{11} > 0$ we have

$$\|g_{k+1} + A_{k+1}^T \lambda_{k+1}\| \leq \max\{(1 - \gamma_{10})\|g_k + A_k^T \lambda_k\|, \gamma_{11}\|c_k\|\} + \alpha_k E'(d_k, \delta_k). \tag{4.11}$$

The boundedness of $\{\alpha_k\}$, Lemma 4.9, and the fact that Assumption 4.1(b) implies δ_k is bounded in norm imply, along with (4.10), that

$$\lim_{k \to \infty} \alpha_k E'(d_k, \delta_k) = 0. \tag{4.12}$$

Consider an arbitrary $\hat{\gamma} > 0$. Lemma 4.9 and the limit (4.12) imply that there exists $k' \geq 0$ such that for all $k \geq k'$ we have

$$\gamma_{11}\|c_k\| < (1 - \gamma_{10})\hat{\gamma} \quad \text{and} \quad \alpha_k E'(d_k, \delta_k) < \tfrac{1}{2}\gamma_{10}\hat{\gamma}. \tag{4.13}$$

Suppose $k \geq k'$ and $\|g_k + A_k^T \lambda_k\| > \hat{\gamma}$. We find from (4.11) that

$$\begin{aligned}
\|g_{k+1} + A_{k+1}^T \lambda_{k+1}\| &\leq (1 - \gamma_{10})\|g_k + A_k^T \lambda_k\| + \tfrac{1}{2}\gamma_{10}\hat{\gamma} \\
&\leq \|g_k + A_k^T \lambda_k\| - \tfrac{1}{2}\gamma_{10}\hat{\gamma}.
\end{aligned}$$

Therefore, $\{\|g_k + A_k^T \lambda_k\|\}$ decreases monotonically by at least a constant amount for $k \geq k'$ while $\{\|g_k + A_k^T \lambda_k\|\} > \hat{\gamma}$, so we eventually find $\|g_k + A_k^T \lambda_k\| \leq \hat{\gamma}$ for some $k = k'' \geq k'$. Then, for $k \geq k''$ we find from (4.11) and (4.13) that

$$\begin{aligned}
\|g_{k+1} + A_{k+1}^T \lambda_{k+1}\| &\leq (1 - \gamma_{10})\hat{\gamma} + \tfrac{1}{2}\gamma_{10}\hat{\gamma} \\
&\leq (1 - \tfrac{1}{2}\gamma_{10})\hat{\gamma},
\end{aligned}$$

so $\|g_k + A_k^T \lambda_k\| \leq \hat{\gamma}$ for all $k \geq k''$. Since the above holds for any $\hat{\gamma} > 0$, we have

$$\lim_{k \to \infty} \|g_k + A_k^T \lambda_k\| = 0,$$

and so the result follows with the above and the result of Lemma 4.9. ∎

5. An implementation. This section contains a description of a particular implementation of Algorithm A and corresponding numerical results to illustrate the robustness of our approach. Note that, for the greatest level of generality within our framework, we implemented Termination Test I with $\kappa \geq 1$ and (3.4) included. A study of the efficiency of the new algorithm in realistic applications is devoted to a separate study [7].

We developed a Matlab implementation of Algorithm A in which the generalized minimum residual (GMRES) method [21] was used for the step computation, for which we adapted the implementation by Kelley [15]. The GMRES method does not exploit the symmetry of the matrix (2.5) in the primal-dual system (2.6), but the stability of the approach is ideal for illustrating the robustness of Algorithm A.

In terms of the input parameters defined throughout the paper, we make the following general comments on their practical effects. First, the values $(\kappa, \kappa_1, \kappa_2)$ and (ϵ, β) should receive special attention as they may greatly affect the ease with which Termination Tests I and II, and therefore the model reduction condition (3.2), are satisfied; larger values for these constants allow for more steps to satisfy at least one of the tests at a given point. In general, looser bounds in Termination Tests I and II will result in cheaper step computations, but these savings must be balanced against possible increases in the number of outer iterations required to find a solution. These parameters and (σ, τ) may also affect the number of iterations required until the penalty parameter stabilizes, an important phenomenon in the analysis of section 4; e.g., larger values of (ϵ, β) may lead to more increases or larger values of π_k or both. In general, however, we claim that the parameters (σ, τ) can be set to default values or to promote consistency between the two termination tests, as we do in (5.3) below.

The stopping condition for the overall nonlinear program is given by

$$\|g_k + A_k^T \lambda_k\|_\infty \leq \max\{\|g_k\|_\infty, 1\}\epsilon_{opt}, \tag{5.1}$$

$$\|c_k\|_\infty \leq \max\{\|c_0\|_\infty, 1\}\epsilon_{feas}, \tag{5.2}$$

where $0 < \epsilon_{opt}, \epsilon_{feas} < 1$ and x_0 is the starting point (e.g., see [23]).

The following algorithm was implemented in Matlab and will be referred to as `isqp`. The `termination` variable is used to indicate the successful or unsuccessful termination of the solver near a local solution of problem (1.1).

ALGORITHM B: INEXACT SQP WITH GMRES AND SMART TESTS.
Given parameters $0 < \epsilon_{feas}, \epsilon_{opt}, \kappa_1, \epsilon, \tau, \sigma, \eta, \alpha_{\min} < 1$ and $0 < k_{\max}, \beta, \kappa, \kappa_2$
Initialize x_0, λ_0, and $\pi_{-1} > 0$
Set `termination` $\leftarrow$ `success`
for $k = 0, 1, 2, \ldots, k_{max}$, or until (5.1) and (5.2) are satisfied
  Compute f_k, g_k, c_k, W_k, and A_k and set $\pi_k \leftarrow \pi_{k-1}$ and $\alpha_k \leftarrow 1$
  for $j = 0, 1, 2, \ldots, n + t$, or until Termination Test I or II is satisfied
    Set (d_k, δ_k) as the jth GMRES solution
  endfor
  if $\tilde{D}\phi(d_k; \pi) > 0$ for all $\pi \geq \pi_k$, set `termination` $\leftarrow$ `failure` and **break**
  if Termination Test II is satisfied and (3.6) does not hold, set $\pi_k \leftarrow \pi_k^{trial} + 10^{-4}$
  while (3.10) is not satisfied and $\alpha_k \geq \alpha_{\min}$, set $\alpha_k \leftarrow \alpha_k/2$
  if $\alpha_k < \alpha_{\min}$, set `termination` $\leftarrow$ `failure` and **break**
  Set $(x_{k+1}, \lambda_{k+1}) \leftarrow (x_k, \lambda_k) + \alpha_k(d_k, \delta_k)$
endfor
if (5.1) or (5.2) is not satisfied, set `termination` $\leftarrow$ `failure`
return `termination`

TABLE 5.1
Input parameter values used for Algorithm B.

Parameter	Value	Parameter	Value
ϵ_{feas}	10^{-6}	η	10^{-8}
ϵ_{opt}	10^{-6}	$\alpha_{\min}$	10^{-8}
κ_1	0.1	$k_{\max}$	1000
ϵ	0.1	κ	1
τ	0.1	π_{-1}	1

We recognize three types of failures in the above approach. First, due to the iteration limit $(n + t)$ imposed on the inner **for** loop, or if the positive definiteness of Assumption 4.1(e) is violated, GMRES may not provide a solution satisfying Termination Test I or II. In this case, we will try to use the step d_k anyway, and, if necessary, we will try increasing π_k to yield a positive value for the directional derivative $D\phi(d_k; \pi_k)$. However, if the directional derivative is nonnegative for any value $\pi \geq \pi_{k-1}$ of the penalty parameter, then the step is an ascent direction for the merit function and the algorithm terminates. Second, if the steplength coefficient must be cut below a given $\alpha_{\min}$ in order to obtain a step satisfying the Armijo condition, then the search direction is deemed unsuitable and the algorithm fails. Since we have a descent direction, this failure can occur only due to finite precision arithmetic errors or if $\alpha_{\min}$ is too large relative to the curvature of the functions. Finally, if the algorithm terminates without satisfying the nonlinear program stopping conditions (5.1) and (5.2), then the maximum number of iterations has been reached. Though there exist techniques for continuing a stagnated run of the algorithm when an ascent direction for the merit function or a short steplength coefficient is computed, we implement naïve failure tests in Algorithm B to aggressively challenge the robustness of our approach.

Table 5.1 contains a listing of the input parameters implemented in our code. For the remaining parameters, we set, as is generally appropriate,

$$\sigma \leftarrow \tau(1 - \epsilon) \tag{5.3}$$

$$\text{and} \quad \kappa_2 \leftarrow \beta \leftarrow \max\left\{ \frac{\|g_0 + A_0^T \lambda_0\|}{\|c_0\| + 1}, 1 \right\}.$$

As previously mentioned, this value for σ promotes consistency between Termination Tests I and II and (3.6). Such a value for κ_2 and β aims to reflect the relationship in scale between the primal and dual feasibility measures.

We compare Algorithm B with an inexact method that only enforces a reduction in the entire primal-dual residual. Our implementation of this approach, also done in Matlab, is identical to Algorithm B except that the GMRES stopping test

for $j = 0, 1, 2, \ldots, n + t$, *or until Termination Test* I *or* II *is satisfied*

is replaced by

for $j = 0, 1, 2, \ldots, n + t$, *or until* (3.3) *is satisfied*,

where $0 < \kappa < 1$ for (3.3) is a given constant. We performed multiple runs of this algorithm, which we call `ires`, for each problem in the test set and will refer to each run by the particular value of κ used.

The algorithms described above were run for 44 equality constrained problems from the CUTEr [3, 10] and COPS [9] collections. Problems from the CUTEr set for

TABLE 5.2

Algorithm success rates; comparison between an inexact SQP method based on the entire residual of the Newton equations and `isqp`*, the algorithm proposed in this paper.*

Algorithm	`ires`										`isqp`
κ	2^{-1}	2^{-2}	2^{-3}	2^{-4}	2^{-5}	2^{-6}	2^{-7}	2^{-8}	2^{-9}	2^{-10}	-
% Solved	45%	66%	68%	80%	80%	77%	82%	82%	86%	86%	100%

which AMPL models were available were selected based on size—fewer than 10,000 variables—and two moderately sized COPS problems were chosen. We note that W_k was set to the exact Hessian of the Lagrangian and that a multiple of the identity matrix was added to W_k, when necessary, to satisfy the positive definiteness of Assumption 4.1(e). Also, as the results provided in this section are intended only as a simple illustration of the robustness of our approach, we did not implement a preconditioner for the primal-dual system for our numerical experiments and, in fact, this was not an issue as many of the problems are relatively small in size. We stress, however, that preconditioning is an essential part of any implementation for many large-scale problems.

Table 5.2 provides the percentage of problems successfully solved for each of the solvers. All the failures for the `ires` algorithm occurred because either the directional derivative $D\phi(d_k;\pi)$ of the merit function was nonnegative for all allowable values of the penalty parameter $\pi \geq \pi_{k-1}$ or the backtracking line search reduced the steplength coefficient α_k below the given tolerance $\alpha_{\min}$. Thus, we find that even for relatively small values of the tolerance parameter κ, the primal component d_k provided by GMRES can yield a value for the directional derivative $D\phi(d_k;\pi)$ of the merit function that is not sufficiently negative for any $\pi \geq \pi_{k-1}$. In other words, `ires` runs the risk of computing near-exact solutions of the primal-dual system (2.4) that correspond to directions of insufficient decrease for the merit function $\phi(x;\pi_k)$.

6. Final remarks. In this paper we have developed an inexact SQP algorithm for equality constrained optimization that is globally convergent under common conditions. The novelties of the approach are centered around a pair of SMART tests for controlling the level of inexactness in the step computation procedure. We close with some remarks about the assumptions used in the paper, the rate of convergence of our approach, and possible extensions of this work.

First, let us recall the boundedness of the multipliers stated in Assumption 4.1(b). Our analysis does not guarantee that the multipliers remain bounded in general; in fact, Algorithm A does not exert direct control over them. We can ensure that $\{\lambda_k\}$ remains bounded, however, by adding to Termination Test I a requirement of the form

$$\|\rho_k\| \leq \kappa' \max\{\|g_k\|, \|A_k\|\}$$

for a constant $\kappa' > 0$. Such a condition ensures that ρ_k is bounded independently of the multipliers λ_k, so then (2.6) and Assumptions 4.1 will imply that $\{\lambda_k\}$ is bounded. An alternative would be to include a safeguard in the algorithm by which the multiplier estimate λ_k is set to a nominal value, say, $\lambda_k = 0$, if $\|g_k + A_k^T\lambda_k\|$ is larger than a given constant.

Second, the rate of convergence of Algorithm A may be slow for a given problem. One can ensure a fast convergence rate, however, by imposing at each step a

requirement of the form

$$\left\| \begin{bmatrix} \rho_k \\ r_k \end{bmatrix} \right\| \leq \kappa_k \left\| \begin{bmatrix} g_k + A_k^T \lambda_k \\ c_k \end{bmatrix} \right\|, \tag{6.1}$$

where $0 < \{\kappa_k\} < 1$ [8]. Then, tightening the values of κ_k during any point of a run of Algorithm A will influence the convergence rate if unit steplengths are taken. For example, if $\kappa_k \leq \hat{\kappa} < 1$ for all large k, then the rate of convergence is linear with rate $\hat{\kappa}$. If, in addition, $\kappa_k \to 0$, then the rate of convergence is superlinear [8]. In practice, the exact penalty function (2.7) can reject unit steps even close to the solution, but this difficulty can be overcome by the use of a second-order correction or nonmonotone techniques [18]. In this manner, we can be sure that the rate of convergence of Algorithm A will be fast once the penalty parameter is stabilized.

Incidentally, by implementing such an approach, where we require the step provided by the iterative linear system solver to satisfy both (6.1) and Termination Test I or II, one can directly observe the extra cost associated with evolving the `ires` algorithm described in the previous section into a robust method. In our experiments we found this extra cost to be minimal for the problems in our test set. For example, let us define a third algorithm, call it `isqp-ires`, that imposes inequality (6.1) along with our termination tests within the step computation of Algorithm B, where $\kappa = \kappa_k = 2^{-5}$ for all k. Note that the key differences between `isqp-ires` and `isqp` are that we have now implemented $\kappa < 1$ for (3.3) and that an inequality of the form (3.3)/(6.1) is also enforced in Termination Test II. Now, if we compare `isqp-ires` with `ires` (with $\kappa = 2^{-5}$), we can observe the extra cost required to satisfy our termination tests beyond simply attaining an accurate solution to the primal-dual system (2.4). It turns out that for the 35 problems solved by both these algorithms, an average of only 0.5 extra total GMRES iterations over the entire run of the algorithm were required by `isqp-ires`. Moreover, by observing the termination tests for the iterative solver, the 9 problems left unsolved by `ires` (approximately 20% of the total number of 44 problems) were all solved successfully by `isqp-ires`. Indeed, the extra cost is minimal with respect to the added robustness.

In addition it is worth noting that imposing condition (6.1) with sufficiently small κ_k implies that the bound (3.4) would automatically be satisfied, and the bounds (3.5) of Termination Test II are satisfied in the case where $\|c_k\|$ is greater than some constant times $\|g_k + A_k^T \lambda_k\|$.

Finally, it would be of interest to analyze the behavior of inexact SQP methods in the presence of Jacobian singularities and when $W_k = \nabla_{xx}^2 \mathcal{L}_k$ for some k with W_k not positive definite in the null space of the constraint Jacobian A_k. However, such an analysis can be complex and would have taken the focus away from the intended scope of this paper. Therefore, we chose to discuss the design of inexact SQP methods in the benign context of Assumptions 4.1.

Acknowledgments. The authors are thankful to Eldad Haber and Nick Gould for productive discussions on this work.

REFERENCES

[1] G. Biros and O. Ghattas, *Parallel Lagrange–Newton–Krylov–Schur methods for PDE-constrained optimization. Part* I*: The Krylov–Schur solver*, SIAM J. Sci. Comput., 27 (2005), pp. 687–713.

[2] G. Biros and O. Ghattas, *Parallel Lagrange–Newton–Krylov–Schur methods for PDE-constrained optimization. Part* II*: The Lagrange–Newton solver, and its application to optimal control of steady viscous flows*, SIAM J. Sci. Comput., 27 (2005), pp. 714–739.

[3] I. Bongartz, A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *CUTE: Constrained and Unconstrained Testing Environment*, ACM Trans. Math. Software, 21 (1995), pp. 123–160.

[4] R. H. Byrd, J.-Ch. Gilbert, and J. Nocedal, *A trust region method based on interior point techniques for nonlinear programming*, Math. Program., 89 (2000), pp. 149–185.

[5] R. H. Byrd, M. E. Hribar, and J. Nocedal, *An interior point algorithm for large scale nonlinear programming*, SIAM J. Optim., 9 (1999), pp. 877–900.

[6] R. H. Byrd and J. Nocedal, *An analysis of reduced Hessian methods for constrained optimization*, Math. Program., 49 (1991), pp. 285–323.

[7] F. E. Curtis and E. Haber, *Numerical experience with an inexact SQP method for PDE-constrained optimization*, in preparation.

[8] R. S. Dembo, S. C. Eisenstat, and T. Steihaug, *Inexact-Newton methods*, SIAM J. Numer. Anal., 19 (1982), pp. 400–408.

[9] E. D. Dolan, J. J. Moré, and T. S. Munson, *Benchmarking Optimization Software with COPS* 3.0, Tech. report ANL/MCS-TM-273, Argonne National Laboratory, Argonne, IL, 2004.

[10] N. I. M. Gould, D. Orban, and Ph. L. Toint, *CUTEr and sifdec: A Constrained and Unconstrained Testing Environment, revisited*, ACM Trans. Math. Software, 29 (2003), pp. 373–394.

[11] E. Haber and U. M. Ascher, *Preconditioned all-at-once methods for large, sparse parameter estimation problems*, Inverse Prob., 17 (2001), pp. 1847–1864.

[12] S. P. Han, *A globally convergent method for nonlinear programming*, J. Optim. Theory Appl., 22 (1977), pp. 297–309.

[13] M. Heinkenschloss and L. N. Vicente, *Analysis of inexact trust-region SQP algorithms*, SIAM J. Optim., 12 (2001), pp. 283–302.

[14] H. Jäger and E. W. Sachs, *Global convergence of inexact reduced SQP methods*, Optim. Methods Softw., 7 (1996), pp. 83–110.

[15] C. T. Kelley, *Iterative methods for linear and nonlinear equations: MATLAB codes*, 1994, http://www4.ncsu.edu/~ctk/matlab_roots.html.

[16] M. Lalee, J. Nocedal, and T. D. Plantenga, *On the implementation of an algorithm for large-scale equality constrained optimization*, SIAM J. Optim., 8 (1998), pp. 682–706.

[17] F. Leibfritz and E. W. Sachs, *Inexact SQP interior point methods and large scale optimal control problems*, SIAM J. Control Optim., 38 (1999), pp. 272–293.

[18] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed., Springer Ser. Oper. Res., Springer, New York, 2006.

[19] M. J. D. Powell, *Variable Metric Methods for Constrained Optimization*, in Mathematical Programming: The State of the Art, Bonn 1982, A. Bachem, M. Grötschel, and B. Korte, eds., Springer-Verlag, Berlin, 1983.

[20] E. E. Prudencio, R. Byrd, and X. C. Cai, *Parallel full space SQP Lagrange–Newton–Krylov–Schwarz algorithms for PDE-constrained optimization problems*, SIAM J. Sci. Comput., 27 (2006), pp. 1305–1328.

[21] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[22] A. Walther, *A First-Order Convergence Analysis of Trust-Region Methods with Inexact Jacobians*, Tech. report MATH-WR-01-2005, Institute of Scientific Computing, Technische Universitat Dresden, Dresden, Germany, 2005.

[23] R. A. Waltz, J. L. Morales, J. Nocedal, and D. Orban, *An interior algorithm for nonlinear optimization that combines line search and trust region steps*, Math. Program., Ser. A, 107 (2006), pp. 391–408.

SIAM J. OPTIM.
Vol. 19, No. 1, pp. 370–396

CONSTRAINT NONDEGENERACY, STRONG REGULARITY, AND NONSINGULARITY IN SEMIDEFINITE PROGRAMMING*

ZI XIAN CHAN† AND DEFENG SUN‡

Abstract. It is known that the Karush–Kuhn–Tucker (KKT) conditions of semidefinite programming can be reformulated as a nonsmooth system via the metric projector over the cone of symmetric and positive semidefinite matrices. We show in this paper that the primal and dual constraint nondegeneracies, the strong regularity, the nonsingularity of the B-subdifferential of this nonsmooth system, and the nonsingularity of the corresponding Clarke's generalized Jacobian, at a KKT point, are all equivalent. Moreover, we prove the equivalence between each of these conditions and the nonsingularity of Clarke's generalized Jacobian of the smoothed counterpart of this nonsmooth system used in several globally convergent smoothing Newton methods. In particular, we establish the quadratic convergence of these methods under the primal and dual constraint nondegeneracies, but without the strict complementarity.

Key words. semidefinite programming, constraint nondegeneracy, strong regularity, nonsingularity, variational analysis, quadratic convergence

AMS subject classifications. 90C22, 90C25, 90C31, 65K05, 65K10

DOI. 10.1137/070681235

1. Introduction.

The standard semidefinite programming (SDP) problem takes the form

$$\begin{array}{ll} \min & \langle C, X\rangle \\ \text{s.t.} & \mathcal{A}X = b\,, \\ & X \in \mathcal{S}^n_+\,, \end{array} \tag{1}$$

where $C \in \mathcal{S}^n$, the linear space of all $n \times n$ real symmetric matrices, $\langle\cdot,\cdot\rangle$ is the usual Frobenius inner product in $\mathcal{S}^n$, $\mathcal{A}$ is a linear operator from $\mathcal{S}^n$ to $\Re^m$, $b \in \Re^m$, and $\mathcal{S}^n_+$ is the cone of all $n \times n$ positive semidefinite matrices in $\mathcal{S}^n$. Let $\mathcal{A}^* : \Re^m \to \mathcal{S}^n$ be the adjoint of $\mathcal{A}$. The dual form of the SDP problem (1) is

$$\begin{array}{ll} \max & b^T y \\ \text{s.t.} & \mathcal{A}^* y + S = C\,, \\ & S \in \mathcal{S}^n_+\,. \end{array} \tag{2}$$

The Karush–Kuhn–Tucker (KKT) conditions, i.e., the first order optimality conditions, for the SDP problem (1) and its dual (2) are

$$\left\{ \begin{array}{l} \mathcal{A}^* y + S = C\,, \\ \mathcal{A}X = b\,, \\ \mathcal{S}^n_+ \ni X \perp S \in \mathcal{S}^n_+\,, \end{array} \right. \tag{3}$$

*Received by the editors January 29, 2007; accepted for publication (in revised form) September 15, 2007; published electronically April 16, 2008. This research was partially supported by the Academic Research Fund under grant R-146-000-104-112 of the National University of Singapore.
http://www.siam.org/journals/siopt/19-1/68123.html
†Department of Mathematics, National University of Singapore, Republic of Singapore (u0301479@alumni.nus.edu.sg).
‡Department of Mathematics and Risk Management Institute, National University of Singapore, Republic of Singapore (matsundf@nus.edu.sg).

where "$X \perp S$" means that X and S are perpendicular to each other, i.e., $\langle X, S\rangle = 0$. Any point $(\overline{X}, \bar{y}, \overline{S}) \in \mathcal{S}^n \times \Re^m \times \mathcal{S}^n$ satisfying (3) is called a KKT point.

Due to its mathematical elegance and wide applications, the research on SDP has been extremely active after the discovery of polynomial time interior point algorithms [1, 27] for solving this problem. For an excellent survey on this, see [47]. Our research in this paper is motivated by [42] on various characterizations of strong regularity, one of the most important concepts in sensitivity and perturbation analysis, introduced by Robinson in his seminal paper [32], for a local optimal solution of the general nonlinear SDP problem. The basic question we want to ask here is: *What does the strong regularity mean for the SDP problem* (1) *and its dual* (2)?

Certainly, all conditions equivalent to the strong regularity presented in [42] for the general nonlinear SDP problem apply to the SDP problem (1), too. However, due to the special structure of the SDP problem (1) and its dual, one may be able to obtain more insightful characterizations about the strong regularity. This is exactly the primary objective of this paper.

For the purpose of achieving this objective, we study the B-subdifferential and Clarke's generalized Jacobian of the nonsmooth system reformulated from (3). We show that the primal and dual constraint nondegeneracies, the strong regularity, the nonsingularity of the B-subdifferential of this nonsmooth system, and the nonsingularity of the corresponding Clarke's generalized Jacobian, at a KKT point $(\overline{X}, \bar{y}, \overline{S}) \in \mathcal{S}^n \times \Re^m \times \mathcal{S}^n$, are all equivalent. The equivalence of the nonsingularity of the B-subdifferential and the nonsingularity of Clarke's generalized Jacobian comes as a surprise, at least to the authors, as we know that the nonsingularity of the B-subdifferential is only a necessary condition for the strong regularity, while the nonsingularity of Clarke's generalized Jacobian is a sufficient condition for the strong regularity (for more discussions, see [15, 28]). It is true, by [42, Theorem 4.1], that the nonsingularity of Clarke's generalized Jacobian is also necessary for the strong regularity in the context of SDP problems. However, it is never known if the nonsingularity of the B-subdifferential is sufficient, too. Here, the unique structure exhibited in SDP problems (1) and (2) plays a key role for us in proving these conditions equivalent. Consequently, the quadratic convergence of some local nonsmooth Newton-type methods studied in [18, 14] follows from any one of these equivalent conditions. In fact, by combining the two papers [18, 19, 14], we know that the primal and dual constraint nondegeneracies are sufficient for the nonsingularity of the B-subdifferential. On the other hand, our equivalent results imply that they are also necessary for the nonsingularity of the B-subdifferential.

The second objective of this paper, largely motivated by the first, is to study under what conditions the globally convergent smoothing Newton methods studied in [9, 10, 20, 46] for solving SDP problems (1) and (2) possess local quadratic convergence, without assuming the strict complementary condition. We achieve this objective by showing that the nonsingularity of the B-subdifferential of one smoothed system used in [9, 10, 20, 46] and the nonsingularity of Clarke's generalized Jacobian of this smoothed system are both equivalent to any of the above-stated equivalent conditions, in particular, the primal and dual constraint nondegeneracies.

The organization of this paper is as follows. In section 2, we study some useful properties of the B-subdifferential and Clarke's generalized Jacobian for Lipschitz functions, particularly for the metric projector over $\mathcal{S}^n_+$ and its smoothed counterpart. The promised equivalent conditions are given in section 3. In section 4, we prove the quadratic convergence of some smoothing Newton methods under the primal and

dual constraint nondegenerate conditions, but without the strict complementarity condition. We give our conclusions in section 5.

2. Generalized Jacobians. Assume that $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ are three finite dimensional real vector spaces, each equipped with a scalar product $\langle\cdot,\cdot\rangle$ and its induced norm $\|\cdot\|$, $\mathcal{O}$ is an open set in $\mathcal{Y}$, and $\Xi : \mathcal{O} \subseteq \mathcal{Y} \to \mathcal{Z}$ is a locally Lipschitz continuous function on the open set $\mathcal{O}$. By the well-known Rademacher theorem [37, section 9.J], we know that Ξ is almost everywhere F(réchet)-differentiable in $\mathcal{O}$. Denote by $\mathcal{D}_\Xi$ the set of all points in $\mathcal{O}$ where Ξ is F-differentiable. Then Clarke's generalized Jacobian of Ξ at $y \in \mathcal{O}$ is defined as follows [12]:

$$\partial\Xi(y) := \mathrm{conv}\{\, \partial_B\Xi(y)\,\},$$

where "conv" denotes the convex hull and the B-subdifferential $\partial_B\Xi(y)$, a name coined by Qi in [29], of Ξ at y takes the form

$$\partial_B\Xi(y) := \{V \,:\, V = \lim_{k\to\infty} \Xi'(y^k)\,,\; y^k \to y\,,\; y^k \in \mathcal{D}_\Xi\}.$$

The next lemma, which is originally proven in [42, Lemma 2.1] under the additional assumption of directional differentiability, is a useful property about characterizing the B-subdifferential of composite functions. Here we drop the condition of directional differentiability and provide a self-contained proof as it may have applications in other places where the directional differentiability is not readily available.

LEMMA 1. *Let $\Psi : \mathcal{X} \to \mathcal{Y}$ be a continuously differentiable function on an open neighborhood $\widehat{N}$ of $\bar{x}$ and $\Xi : \mathcal{O} \subseteq \mathcal{Y} \to \mathcal{Z}$ be a locally Lipschitz continuous function on an open set $\mathcal{O}$ containing $\bar{y} := \Psi(\bar{x})$. Define $\Phi : \widehat{N} \to \mathcal{Z}$ by $\Phi(x) := \Xi(\Psi(x))$, $x \in \widehat{N}$. Suppose that $\Psi'(\bar{x}) : \mathcal{X} \to \mathcal{Y}$ is onto. Then there exists an open neighborhood of $\bar{x}$ such that Φ is F-differentiable at x in this neighborhood if and only if Ξ is F-differentiable at $\Psi(x)$ and*

$$\partial_B\Phi(\bar{x}) = \partial_B\Xi(\bar{y})\Psi'(\bar{x})\,. \tag{4}$$

Proof. Shrink $\widehat{N}$, if necessary, assume that $\Psi(\widehat{N}) \subseteq \mathcal{O}$, and for each $x \in \widehat{N}$, $\Psi'(x)$ is onto. Then Φ is Lipschitz continuous on $\widehat{N}$.

We shall first show that Φ is F-differentiable at $x \in \widehat{N}$ if and only if Ξ is F-differentiable at $\Psi(x)$, which, by the definition of the B-subdifferential, implies

$$\partial_B\Phi(\bar{x}) \subseteq \partial_B\Xi(\bar{y})\Psi'(\bar{x}).$$

By the definition of Φ, we know that if Ξ is F-differentiable at $\Psi(x)$, then Φ is F-differentiable at $x \in \widehat{N}$. Now, assume that Φ is F-differentiable at $x \in \widehat{N}$. Since $A := \Psi'(x)$ is onto, AA^* is invertible, where $A^* : \mathcal{Y} \to \mathcal{X}$ is the adjoint of A. For any $\Delta y \in \mathcal{Y}$, let

$$\Delta x := A^*(AA^*)^{-1}\Delta y\,.$$

Then, for any $\mathcal{Y} \ni \Delta y \to 0$, we have

$$\begin{aligned}
&\|\Xi(\Psi(x)+\Delta y)-\Xi(\Psi(x))-\Phi'(x)(A^*(AA^*)^{-1}\Delta y)\| \\
&=\|\Xi(\Psi(x)+A\Delta x)-\Phi(x)-\Phi'(x)(\Delta x)\| \\
&\le \|\Phi(x+\Delta x)-\Phi(x)-\Phi'(x)(\Delta x)\|+O(\|\Xi(\Psi(x+\Delta x))-\Xi(\Psi(x)+A\Delta x)\|) \\
&=o(\|\Delta x\|)+O(\|\Psi(x+\Delta x)-(\Psi(x)+A\Delta x)\|) \\
&=o(\|\Delta x\|)+O(\|\Psi(x+\Delta x)-\Psi(x)-\Psi'(x)(\Delta x)\|) \\
&=\ o(\|\Delta x\|)=o(\|\Delta y\|)\,,
\end{aligned}$$

which implies that Ξ is F-differentiable at $\Psi(x)$. This proves the first part of our conclusion.

Next, we show that the following inclusion holds:

$$\partial_B\Phi(\bar{x}) \supseteq \partial_B\Xi(\bar{y})\Psi'(\bar{x}).$$

This part's proof follows exactly the proof of the second part of Lemma 2.1 in [42]. Let $W \in \partial_B\Xi(\bar{y})$ be an arbitrary element. Then there exists a sequence $\{y^k\}$ in $\mathcal{O}$ converging to $\bar{y}$ such that Ξ is F-differentiable at y^k and $W = \lim_{k\to\infty}\Xi'(y^k)$. Let $\overline{A} := \Psi'(\bar{x})$. By applying the classical inverse function theorem to

$$\Psi\big(\bar{x}+\overline{A}^*(y-\bar{y})\big)-\Psi(\bar{x})=0,$$

we obtain that there exists a sequence $\{\tilde{y}^k\}$ in $\mathcal{O}$ converging to $\bar{y}$ such that

$$\Psi\big(\bar{x}+\overline{A}^*(\tilde{y}^k-\bar{y})\big)-\Psi(\bar{x})=y^k-\Psi(\bar{x})$$

for all k sufficiently large. Let $\tilde{x}^k := \bar{x}+\overline{A}^*(\tilde{y}^k-\bar{y})$. Then $y^k=\Psi(\tilde{x}^k)$ and Φ is F-differentiable at $\tilde{x}^k$ with

$$\Phi'(\tilde{x}^k)=\Xi'(y^k)\Psi'(\tilde{x}^k).$$

By using the fact that $\tilde{y}^k \to \bar{y}$ implies $\tilde{x}^k \to \bar{x}$, we know that there exists a $V \in \partial_B\Phi(\bar{x})$ such that

$$W\Psi'(\bar{x})=\lim_{k\to\infty}\Xi'(y^k)\lim_{k\to\infty}\Psi'(\tilde{x}^k)=\lim_{k\to\infty}\Phi'(\tilde{x}^k)=V\in\partial_B\Phi(\bar{x}).$$

The proof is completed. □

For any nonempty closed convex set $K \subseteq \mathcal{Z}$, let $\Pi_K : \mathcal{Z} \to \mathcal{Z}$ denote the metric projector over K. That is, for any $y \in \mathcal{Z}$, $\Pi_K(y)$ is the unique optimal solution to the convex programming problem

$$\begin{array}{ll} \min & \dfrac{1}{2}\langle z-y, z-y\rangle \\ \text{s.t.} & z \in K. \end{array} \tag{5}$$

Since the metric projector $\Pi_K(\cdot)$ is globally Lipschitz continuous with modulus 1 [49], $\Pi_K(\cdot)$ is F-differentiable almost everywhere in $\mathcal{Z}$. Thus, for any $y \in Z$, $\partial\Pi_K(y)$ is well defined. In particular, it is shown in [25, Proposition 1] that for any $y \in \mathcal{Z}$, $V \in \partial\Pi_K(y)$ is self-adjoint and satisfies

$$V \succeq V^2, \quad \text{i.e., } \langle d, Vd\rangle \ge \langle d, V^2 d\rangle \quad \forall\, d \in \mathcal{Z}\,. \tag{6}$$

In our subsequent analysis, we need a finer characterization about the B-subdifferential and Clarke's generalized Jacobian of $\Pi_{\mathcal{S}^n_+}(\cdot)$ and its smoothed counterpart. We write $A \succeq 0$ and $A \succ 0$ to mean that A is a symmetric positive semidefinite matrix and a symmetric positive definite matrix, respectively. For any $A \in \mathcal{S}^n$, let $A_+ := \Pi_{\mathcal{S}^n_+}(A)$ be the metric projection of A onto $\mathcal{S}^n_+$ under the usual Frobenius inner product in $\mathcal{S}^n$. Assume that A has the spectral decomposition

$$A = P\Lambda P^T, \tag{7}$$

where Λ is the diagonal matrix of eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ of A and P is a corresponding orthogonal matrix of orthonormal eigenvectors. Then

$$A_+ = P\Lambda_+ P^T,$$

where Λ_+ is the diagonal matrix whose diagonal entries are the nonnegative parts of the respective diagonal entries of Λ. The formula for A_+ has been used by statisticians for several decades, e.g., [38, Theorem 1]. Higham [16] and Tseng [48] brought it to the attention of the optimization community. Define three index sets of positive, zero, and negative eigenvalues of A, respectively, as

$$\alpha := \{i : \lambda_i > 0\}, \quad \beta := \{i : \lambda_i = 0\}, \quad \gamma := \{i : \lambda_i < 0\}.$$

Write

$$\Lambda = \begin{bmatrix} \Lambda_\alpha & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \Lambda_\gamma \end{bmatrix} \quad \text{and} \quad P = \begin{bmatrix} P_\alpha & P_\beta & P_\gamma \end{bmatrix}$$

with $P_\alpha \in \Re^{n\times|\alpha|}$, $P_\beta \in \Re^{n\times|\beta|}$, and $P_\gamma \in \Re^{n\times|\gamma|}$. For this eigenvalue vector $\lambda \in \Re^n$, define the corresponding symmetric matrix $U \in \mathcal{S}^n$ with entries

$$U_{ij} := \frac{\max\{\lambda_i, 0\} + \max\{\lambda_j, 0\}}{|\,\lambda_i\,| + |\,\lambda_j\,|}, \quad i, j = 1, \ldots, n, \tag{8}$$

where $0/0$ is defined to be 1.

We know from Bonnans, Cominetti, and Shapiro [5, 6] that $\Pi_{\mathcal{S}^n_+}$ is directionally differentiable everywhere in $\mathcal{S}^n$, and from Sun and Sun [43] that $\Pi_{\mathcal{S}^n_+}$ is strongly semismooth everywhere in $\mathcal{S}^n$ and the directional derivative $\Pi'_{\mathcal{S}^n_+}(A; H)$ of $\Pi_{\mathcal{S}^n_+}$ at A with direction $H \in \mathcal{S}^n$ is given by

$$\Pi'_{\mathcal{S}^n_+}(A; H) = P \begin{bmatrix} \widetilde{H}_{\alpha\alpha} & \widetilde{H}_{\alpha\beta} & U_{\alpha\gamma} \circ \widetilde{H}_{\alpha\gamma} \\ \widetilde{H}^T_{\alpha\beta} & \Pi_{\mathcal{S}^{|\beta|}_+}(\widetilde{H}_{\beta\beta}) & 0 \\ \widetilde{H}^T_{\alpha\gamma} \circ U^T_{\alpha\gamma} & 0 & 0 \end{bmatrix} P^T, \tag{9}$$

where $\widetilde{H} := P^T H P$ and "$\circ$" denotes the Hadamard product. For a general discussion on (strongly) semismooth functions, see [26, 29, 31]. The tangent cone of $\mathcal{S}^n_+$ at A_+, in the sense of convex analysis [36], can be characterized as

$$\mathcal{T}_{\mathcal{S}^n_+}(A_+) = \{B \in \mathcal{S}^n : B = \Pi'_{\mathcal{S}^n_+}(A_+; B)\} = \{B \in \mathcal{S}^n : [\,P_\beta\ P_\gamma\,]^T B\,[\,P_\beta\ P_\gamma\,] \succeq 0\}.$$

Note, however, that the characterization of $\mathcal{T}_{\mathcal{S}^n_+}(A_+)$ was first obtained by Arnold [3] without using the directional derivative $\Pi'_{\mathcal{S}^n_+}(A_+; H)$. The linearity space of $\mathcal{T}_{\mathcal{S}^n_+}(A_+)$,

i.e., the largest linear space in $\mathcal{T}_{\mathcal{S}^n_+}(A_+)$, denoted by $\mathrm{lin}\left(\mathcal{T}_{\mathcal{S}^n_+}(A_+)\right)$, then takes the following form:

$$\mathrm{lin}\left(\mathcal{T}_{\mathcal{S}^n_+}(A_+)\right) = \{B \in \mathcal{S}^n \,:\, P_\beta^T B P_\beta = 0,\ \ P_\beta^T B P_\gamma = 0,\ \ P_\gamma^T B P_\gamma = 0\}\,. \tag{10}$$

The critical cone of $\mathcal{S}^n_+$ at $A \in \mathcal{S}^n$, associated with the convex optimization problem (5) with $K = \mathcal{S}^n_+$, is defined as

$$\begin{array}{rcl} \mathcal{C}(A;\mathcal{S}^n_+): & = & \mathcal{T}_{\mathcal{S}^n_+}(A_+) \cap (A_+ - A)^\perp \\ & = & \{B \in \mathcal{S}^n \,:\, P_\beta^T B P_\beta \succeq 0,\ P_\beta^T B P_\gamma = 0,\ P_\gamma^T B P_\gamma = 0\}\,, \end{array} \tag{11}$$

where $(A_+ - A)^\perp := \{B \in \mathcal{S}^n \,:\, \langle B, A_+ - A\rangle = 0\}$. Therefore, the affine hull of $\mathcal{C}(A;\mathcal{S}^n_+)$, which we denote $\mathrm{aff}(\mathcal{C}(A;\mathcal{S}^n_+))$, can be written as

$$\mathrm{aff}\left(\mathcal{C}(A;\mathcal{S}^n_+)\right) = \left\{B \in \mathcal{S}^n \,:\, P_\beta^T B P_\gamma = 0,\ P_\gamma^T B P_\gamma = 0\right\}. \tag{12}$$

In the case that $\beta = \emptyset$ holds, i.e., the case that A is nonsingular, $\Pi_{\mathcal{S}^n_+}(\cdot)$ is F-differentiable at A and (9) reduces to the famous result of Löwner [22]:

$$\Pi'_{\mathcal{S}^n_+}(A)H = P\begin{bmatrix} \widetilde{H}_{\alpha\alpha} & U_{\alpha\gamma}\circ\widetilde{H}_{\alpha\gamma} \\ \widetilde{H}^T_{\alpha\gamma}\circ U^T_{\alpha\gamma} & 0 \end{bmatrix} P^T \quad \forall\, H \in \mathcal{S}^n\,. \tag{13}$$

From (13), one may compute the B-subdifferential and Clarke's generalized Jacobian of $\Pi_{\mathcal{S}^n_+}(\cdot)$ by their definitions.[1] This has been done by a number of authors [9, 20, 23, 24, 28]. One difficulty in obtaining good formulas for $\partial_B\Pi_{\mathcal{S}^n_+}(A)$ and $\partial\Pi_{\mathcal{S}^n_+}(A)$ is that they both depend on the orthogonal matrices P in the spectral decomposition of A. This difficulty can be overcome by employing the following link developed by Pang, Sun, and Sun [28] on $\partial_B\Pi_{\mathcal{S}^n_+}(A)$ and the B-subdifferential of $\Theta(\cdot) := \Pi'_{\mathcal{S}^n_+}(A;\cdot)$ at the origin

$$\partial_B\Pi_{\mathcal{S}^n_+}(A) = \partial_B\Theta(0)\,. \tag{14}$$

This link leads to the following useful result on $\partial_B\Pi_{\mathcal{S}^n_+}(A)$ and $\partial\Pi_{\mathcal{S}^n_+}(A)$. See Sun [42, Proposition 2.2] for a short proof.

PROPOSITION 2. *Suppose that $A \in \mathcal{S}^n$ has the spectral decomposition as in (7). Then a $V \in \partial_B\Pi_{\mathcal{S}^n_+}(A)$ (respectively, $\partial\Pi_{\mathcal{S}^n_+}(A)$) if and only if there exists a $V_{|\beta|} \in \partial_B\Pi_{\mathcal{S}^{|\beta|}_+}(0)$ (respectively, $\partial\Pi_{\mathcal{S}^{|\beta|}_+}(0)$) such that*

$$V(H) = P\begin{bmatrix} \widetilde{H}_{\alpha\alpha} & \widetilde{H}_{\alpha\beta} & U_{\alpha\gamma}\circ\widetilde{H}_{\alpha\gamma} \\ \widetilde{H}^T_{\alpha\beta} & V_{|\beta|}(\widetilde{H}_{\beta\beta}) & 0 \\ \widetilde{H}^T_{\alpha\gamma}\circ U^T_{\alpha\gamma} & 0 & 0 \end{bmatrix} P^T \quad \forall\, H \in \mathcal{S}^n, \tag{15}$$

where $\widetilde{H} := P^T H P$.

Proposition 2 simply says that in order to compute $\partial_B\Pi_{\mathcal{S}^n_+}(A)$ and $\partial\Pi_{\mathcal{S}^n_+}(A)$, one needs only to fix an arbitrary orthogonal matrix P satisfying (7) and compute the

[1] Note that in numerical computations it is generally impossible to compute exactly the spectral decomposition of A as in (7). Instead, the right-hand side of (7) is the true spectral decomposition of a nearby matrix of A [8]. Consequently, the numerically computed subdifferentials are actually for this nearby matrix. In this paper, we will not address this numerical issue further.

corresponding "caged" part $\partial_B \Pi_{\mathcal{S}_+^{|\beta|}}(0)$ (hence $\partial \Pi_{\mathcal{S}_+^{|\beta|}}(0)$), which is much easier to handle. To see this, let $\mathcal{Q}_{|\beta|}$ be the set of all orthogonal matrices of order $|\beta| \times |\beta|$ and

$$\Re_>^{|\beta|} := \{ z \in \Re^{|\beta|} \,:\, z_1 \geq \cdots \geq z_{|\beta|} \ \text{ and } z_i \neq 0 \ \forall\, i \,\}.$$

Let $p : \Re \to \Re$ be the "plus" function defined by $p(t) \equiv \max(0, t)$, $t \in \Re$. For any $z \in \Re_>^{|\beta|}$, let $p^{[1]}(z)$ represent the first divided difference matrix used in matrix analysis for $p(\cdot)$ at z [4]:

$$\big[\, p^{[1]}(z) \,\big]_{ij} = \begin{cases} \dfrac{p(z_i) - p(z_j)}{z_i - z_j} \in [0,1] & \text{if } z_i \neq z_j, \\ p'(z_i) \in \{0, 1\} & \text{if } z_i = z_j, \end{cases} \qquad i, j = 1, \ldots, n\,. \tag{16}$$

Then, by (9) and (13), one can readily draw the conclusion that $V_{|\beta|} \in \partial_B \Pi_{\mathcal{S}_+^{|\beta|}}(0)$ if and only if there exist $Q \in \mathcal{Q}_{|\beta|}$ and $\Omega \in \mathcal{U}_{|\beta|}$ such that

$$V_{|\beta|}(Z) = Q\,[\,\Omega \circ (Q^T Z Q)\,]\,Q^T \quad \forall\, Z \in \mathcal{S}^{|\beta|}\,, \tag{17}$$

where

$$\mathcal{U}_{|\beta|} := \{\, \Omega \,:\, \Omega = \lim_{k \to \infty} p^{[1]}(z^k), \ \ z^k \to 0, \ z^k \in \Re_>^{|\beta|} \,\}\,.$$

In [23], Malick and Sendov gave a detailed account on the structure of $\mathcal{U}_{|\beta|}$. In this paper, we do not need the exact structure of $\mathcal{U}_{|\beta|}$ except for the following fact that for any $\Omega \in \mathcal{U}_{|\beta|}$,

$$\Omega_{ij} \in [0, 1], \quad i, j = 1, \ldots, |\beta|\,.$$

Note that both the zero mapping $V_{|\beta|}^0 \equiv 0$ and the identity mapping $V_{|\beta|}^{\mathcal{I}} = \mathcal{I}$ from $\mathcal{S}^{|\beta|} \to \mathcal{S}^{|\beta|}$ are elements in $\partial_B \Pi_{\mathcal{S}_+^{|\beta|}}(0)$. Let V^0 and $V^{\mathcal{I}}$ be defined by (15) with $V_{|\beta|}$ being replaced by $V_{|\beta|}^0$ and $V_{|\beta|}^{\mathcal{I}}$, respectively. Define

$$\mathrm{ex}\,(\partial_B \Pi_{\mathcal{S}_+^n}(A)) := \{\, V^0, \ V^{\mathcal{I}} \,\}. \tag{18}$$

Using the fact that both V^0 and $V^{\mathcal{I}}$ are elements in $\partial_B \Pi_{\mathcal{S}_+^n}(A)$, we have

$$\mathrm{ex}\,(\partial_B \Pi_{\mathcal{S}_+^n}(A)) \subseteq \partial_B \Pi_{\mathcal{S}_+^n}(A)\,.$$

Since $\Pi_{\mathcal{S}_+^n}(\cdot)$ is not differentiable everywhere, several papers [9, 10, 20, 46] on smoothing Newton methods, for solving the SDP problem and beyond, consider the following smoothed counterpart of $\Pi_{\mathcal{S}_+^n}(\cdot)$:

$$\Phi(\varepsilon, A) := \big[\, A + \sqrt{\varepsilon^2 I + A^2}\,\big]/2, \quad (\varepsilon, A) \in \Re \times \mathcal{S}^n\,, \tag{19}$$

where we use I to represent the identity matrix of appropriate dimension. Note that the function $\Phi(\cdot, \cdot)$ is continuously differentiable around any $(\varepsilon, A) \in \Re \times \mathcal{S}^n$ if $\varepsilon^2 I + A^2$ is nonsingular and when $\varepsilon = 0$, $\Phi(0, A) = \Pi_{\mathcal{S}_+^n}(A)$. Furthermore, $\Phi(\cdot, \cdot)$ is globally Lipschitz continuous and strongly semismooth at any $(0, A) \in \Re \times \mathcal{S}^n$ [46]. For some extensions on these properties, see [44].

Let $\phi : \Re^2 \to \Re$ be defined by

$$\phi(\varepsilon, t) = \left[\, t + \sqrt{\varepsilon^2 + t^2}\,\right]/2, \quad (\varepsilon, t) \in \Re \times \Re\,.$$

Let A have the spectral decomposition in (7). Then, by matrix analysis [4, 17], we have

$$\Phi(\varepsilon, A) = P \begin{bmatrix} \phi(\varepsilon, \lambda_1) & & \\ & \ddots & \\ & & \phi(\varepsilon, \lambda_n) \end{bmatrix} P^T\,.$$

For any $(\varepsilon, x) \in \Re \times \Re^n$ such that $\varepsilon^2 + x_i^2 > 0$ for all i, we use $\widehat{U}(\varepsilon, x) \in \mathcal{S}^n$ to represent the first divided difference matrix for $\phi(\varepsilon, \cdot)$ at x given by

$$\left[\,\widehat{U}(\varepsilon, x)\,\right]_{ij} = \begin{cases} \dfrac{\phi(\varepsilon, x_i) - \phi(\varepsilon, x_j)}{x_i - x_j} \in [0, 1] & \text{if } x_i \neq x_j, \\ \phi'_{x_i}(\varepsilon, x_i) \in [0, 1] & \text{if } x_i = x_j, \end{cases} \quad i, j = 1, \ldots, n\,. \tag{20}$$

Then, according to Lemma 2.3 in [46], we know that for any $\varepsilon \in \Re$ such that $\varepsilon^2 + \lambda_i^2 > 0$ for all i (i.e., $\varepsilon^2 I + A^2$ is nonsingular), and any $(\tau, H) \in \Re \times \mathcal{S}^n$, we have

$$\Phi'(\varepsilon, A)(\tau, H) = P\,[\,\widehat{U}(\varepsilon, \lambda) \circ (P^T H P) + \tau D(\varepsilon, \lambda)\,]\,P^T \tag{21}$$

and

$$\Phi'((0, A); (\tau, H)) = P \begin{bmatrix} \widetilde{H}_{\alpha\alpha} & \widetilde{H}_{\alpha\beta} & U_{\alpha\gamma} \circ \widetilde{H}_{\alpha\gamma} \\ \widetilde{H}_{\alpha\beta}^T & \Phi_{|\beta|}(\tau, \widetilde{H}_{\beta\beta}) & 0 \\ \widetilde{H}_{\alpha\gamma}^T \circ U_{\alpha\gamma}^T & 0 & 0 \end{bmatrix} P^T, \tag{22}$$

where $\widetilde{H} = P^T H P$, $D(\varepsilon, \lambda) \in \mathcal{S}^n$ is the diagonal matrix given by

$$D(\varepsilon, \lambda) = \begin{bmatrix} \phi'_\varepsilon(\varepsilon, \lambda_1) & & \\ & \ddots & \\ & & \phi'_\varepsilon(\varepsilon, \lambda_n) \end{bmatrix}, \tag{23}$$

$U \in \mathcal{S}^n$ is defined by (8), and for any $(t, Z) \in \Re \times \mathcal{S}^{|\beta|}$,

$$\Phi_{|\beta|}(t, Z) := \left[\, Z + \sqrt{t^2 I + Z^2}\,\right]/2\,. \tag{24}$$

Define $\Psi : \Re \times \mathcal{S}^n \to \Re \times \mathcal{S}^n$ by

$$\Psi(\tau, H) := (\tau, P^T H P), \quad (\tau, H) \in \Re \times \mathcal{S}^n,$$

and $\Xi : \Re \times \mathcal{S}^n \to \mathcal{S}^n$ by

$$\Xi(t, M) := P \begin{bmatrix} M_{\alpha\alpha} & M_{\alpha\beta} & U_{\alpha\gamma} \circ M_{\alpha\gamma} \\ M_{\alpha\beta}^T & \Phi_{|\beta|}(t, M_{\beta\beta}) & 0 \\ M_{\alpha\gamma}^T \circ U_{\alpha\gamma}^T & 0 & 0 \end{bmatrix} P^T\,, \tag{25}$$

where $(t, M) \in \Re \times \mathcal{S}^n$. Write $\Gamma(\cdot, \cdot) \equiv \Phi'((0, A); (\cdot, \cdot))$. Then, we have

$$\Gamma(\tau, H) = \Xi(\Psi(\tau, H)), \quad (\tau, H) \in \Re \times \mathcal{S}^n\,. \tag{26}$$

Since for any $(\tau, H) \in \Re \times \mathcal{S}^n$, $\Psi'(\tau, H) : \Re \times \mathcal{S}^n \to \Re \times \mathcal{S}^n$ is onto, we know from the first part of Lemma 1 that Γ is F-differentiable at $(\tau, H) \in \Re \times \mathcal{S}^n$ if and only if Ξ is F-differentiable at $\Psi(\tau, H)$, which is equivalent to the nonsingularity of $\tau^2 I + (\widetilde{H}_{\beta\beta})^2$, where $\widetilde{H} = P^T H P$. Thus, we have the following lemma.

LEMMA 3. *For any $(\tau, H) \in \Re \times \mathcal{S}^n$, let $\widetilde{H} = P^T H P$. Then $\Gamma(\cdot,\cdot) \equiv \Phi'((0, A); (\cdot,\cdot))$ is F-differentiable at $(\tau, H) \in \Re \times \mathcal{S}^n$ if and only if $\tau^2 I + (\widetilde{H}_{\beta\beta})^2$ is nonsingular.*

The following lemma establishes the equivalence between $\partial_B \Phi(0, A)$ and $\partial_B \Gamma(0, 0)$, which is analogous to (14) for operators $\Pi_{\mathcal{S}^n_+}$ and Θ. Its proof largely follows that given in [28, Lemma 11], but with new difficulties to overcome.

LEMMA 4. *Suppose that $A \in \mathcal{S}^n$ has the spectral decomposition in (7). For $\Gamma(\cdot,\cdot) \equiv \Phi'((0, A); (\cdot,\cdot))$, it holds that*

$$\partial_B \Phi(0, A) = \partial_B \Gamma(0, 0)\,. \tag{27}$$

Proof. Let $V \in \partial_B \Phi(0, A)$. Then, by (21), (22), and the definition of $\partial_B \Phi(0, A)$, there exists a sequence $\{(\varepsilon_k, A^k)\}$ in $\Re \times \mathcal{S}^n$ converging to $(0, A)$ with $\varepsilon_k^2 I + (A^k)^2$ being nonsingular such that $V = \lim_{k\to\infty} \Phi'(\varepsilon_k, A^k)$. Let $A^k \equiv P^k \Lambda^k (P^k)^T$ be the orthogonal decomposition of A^k, where Λ^k is the diagonal matrix whose diagonal entries are the eigenvalues $\lambda_1^k \geq \cdots \geq \lambda_n^k$ of A^k and P^k is a corresponding matrix of orthonormal eigenvectors. Writing each Λ^k in the same form as Λ,

$$\Lambda^k = \begin{bmatrix} \Lambda_\alpha^k & 0 & 0 \\ 0 & \Lambda_\beta^k & 0 \\ 0 & 0 & \Lambda_\gamma^k \end{bmatrix},$$

we have $\Lambda = \lim_{k\to\infty} \Lambda^k$, which implies that Λ_α^k and Λ_γ^k are nonsingular matrices for all k sufficiently large and $\lim_{k\to\infty} \Lambda_\beta^k = 0$. For each k, let $U^k \equiv \widehat{U}(\varepsilon_k, \lambda^k)$ be defined by (20) and $D^k \equiv D(\varepsilon_k, \lambda^k)$ be defined by (23), respectively. Then, for an arbitrarily chosen $(\tau, H) \in \Re \times \mathcal{S}^n$ with $\widetilde{H}^k = (P^k)^T H P^k$, we obtain from (21) that

$$\Phi'(\varepsilon_k, A^k)(\tau, H) = P^k \left[U^k \circ (P^k)^T H P^k + \tau D^k \right] (P^k)^T\,. \tag{28}$$

By taking a subsequence if necessary, we may assume that $\{P^k\}$ is a convergent sequence with limit $P^\infty \equiv \lim_{k\to\infty} P^k$. This matrix P^∞ will play the role of the matrix P in the spectral decomposition (7). Without causing any confusion, we will simply use P, rather than P^∞, in our subsequent analysis. Since both $\{U^k\}$ and $\{D^k\}$ are uniformly bounded, by further taking subsequences if necessary, we may assume that both sequences $\{U^k\}$ and $\{D^k\}$ converge. Taking limits on both sides of (28), we obtain

$$P^T V(\tau, H) P = \begin{bmatrix} \widetilde{H}_{\alpha\alpha} & \widetilde{H}_{\alpha\beta} & U_{\alpha\gamma} \circ \widetilde{H}_{\alpha\gamma} \\ \widetilde{H}_{\alpha\beta}^T & \lim\limits_{k\to\infty} U_{\beta\beta}^k \circ \widetilde{H}_{\beta\beta} & 0 \\ \widetilde{H}_{\alpha\gamma}^T \circ U_{\alpha\gamma}^T & 0 & 0 \end{bmatrix}$$

$$+ \tau \begin{bmatrix} 0 & 0 & 0 \\ 0 & \lim\limits_{k\to\infty} D_\beta^k & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

where

$$D^k_\beta = \begin{bmatrix} \phi'_\varepsilon(\varepsilon_k, \lambda^k_{|\alpha|+1}) & & \\ & \ddots & \\ & & \phi'_\varepsilon(\varepsilon_k, \lambda^k_{|\alpha|+|\beta|}) \end{bmatrix}.$$

For each k, define

$$M^k := P \begin{bmatrix} 0 & 0 & 0 \\ 0 & \Lambda^k_\beta & 0 \\ 0 & 0 & 0 \end{bmatrix} P^T.$$

Let $\widetilde{M}^k := P^T M^k P$. Because $\varepsilon_k^2 I + (\widetilde{M}^k_{\beta\beta})^2 = \varepsilon_k^2 I + (\Lambda^k_\beta)^2$ is nonsingular, Γ is F-differentiable at (ε_k, M^k) with

$$\begin{aligned}
\Gamma'(\varepsilon_k, M^k)(\tau, H) &= \lim_{t \downarrow 0} \left\{ \frac{\Gamma(\varepsilon_k + t\tau, M^k + tH) - \Gamma(\varepsilon_k, M^k)}{t} \right\} \\
&= P \begin{bmatrix} \widetilde{H}_{\alpha\alpha} & \widetilde{H}_{\alpha\beta} & U_{\alpha\gamma} \circ \widetilde{H}_{\alpha\gamma} \\ \widetilde{H}^T_{\alpha\beta} & \lim\limits_{t \downarrow 0} \dfrac{\Phi_{|\beta|}(\varepsilon_k + t\tau, \Lambda^k_\beta + t\widetilde{H}_{\beta\beta}) - \Phi_{|\beta|}(\varepsilon_k, \Lambda^k_\beta)}{t} & 0 \\ \widetilde{H}^T_{\alpha\gamma} \circ U^T_{\alpha\gamma} & 0 & 0 \end{bmatrix} P^T \\
&= P \begin{bmatrix} \widetilde{H}_{\alpha\alpha} & \widetilde{H}_{\alpha\beta} & U_{\alpha\gamma} \circ \widetilde{H}_{\alpha\gamma} \\ \widetilde{H}^T_{\alpha\beta} & U^k_{\beta\beta} \circ \widetilde{H}_{\beta\beta} + \tau D^k_\beta & 0 \\ \widetilde{H}^T_{\alpha\gamma} \circ U^T_{\alpha\gamma} & 0 & 0 \end{bmatrix} P^T,
\end{aligned}$$

where we have applied (21) to $\Phi_{|\beta|}$ defined by (24) at $(\varepsilon_k, \Lambda^k_\beta)$. Thus,

$$V(\tau, H) = \lim_{k \to \infty} \Gamma'(\varepsilon_k, M^k)(\tau, H).$$

Since $(\tau, H) \in \Re \times \mathcal{S}^n$ is arbitrary, it follows that $V \in \partial_B \Gamma(0, 0)$.

Conversely, let $V \in \partial_B \Gamma(0, 0)$. Since, from Lemma 3, Γ is F-differentiable at $(\varepsilon, M) \in \Re \times \mathcal{S}^n$ if and only if $\varepsilon^2 I + (\widetilde{M}_{\beta\beta})^2$ is nonsingular with $\widetilde{M} = P^T M P$, there exists a sequence $\{(\varepsilon_k, M^k)\} \in \Re \times \mathcal{S}^n$ converging to $(0, 0)$ such that $\varepsilon_k^2 I + (\widetilde{M}^k_{\beta\beta})^2$ is nonsingular for each k and $V = \lim_{k \to \infty} \Gamma'(\varepsilon_k, M^k)$, where $\widetilde{M}^k = P^T M^k P$. Let $\widetilde{M}^k_{\beta\beta}$ have the spectral decomposition

$$\widetilde{M}^k_{\beta\beta} = Q^k \widetilde{\Lambda}^k_\beta (Q^k)^T,$$

where $Q^k \in \mathcal{Q}_{|\beta|}$ is an orthogonal matrix in $\mathcal{S}^{|\beta|}$ and $\widetilde{\Lambda}^k_\beta$ is the diagonal matrix whose diagonal entries are the eigenvalues $\tilde{z}^k_1 \geq \cdots \geq \tilde{z}^k_{|\beta|}$ of $\widetilde{M}^k_{\beta\beta}$. Let $\tilde{\lambda}^k \in \Re^n$ be such that if $i \in \alpha \cup \gamma$, then $\tilde{\lambda}^k_i = \lambda_i$ and if $i \in \beta$, $\tilde{\lambda}^k_i$ is the $(i - |\alpha|)$th eigenvalue of $\widetilde{M}^k_{\beta\beta}$, i.e.,

$\tilde{z}^k_{(i-|\alpha|)}$. Then, by (22), for any $(\tau, H) \in \Re \times \mathcal{S}^n$ we have

(29)

$$\Gamma'(\varepsilon_k, M^k)(\tau, H) = \lim_{t\downarrow 0} \left\{ \frac{\Gamma(\varepsilon_k + t\tau, M^k + tH) - \Gamma(\varepsilon_k, M^k)}{t} \right\}$$

$$= P \begin{bmatrix} \widetilde{H}_{\alpha\alpha} & \widetilde{H}_{\alpha\beta} & U_{\alpha\gamma} \circ \widetilde{H}_{\alpha\gamma} \\ \widetilde{H}^T_{\alpha\beta} & \lim\limits_{t\downarrow 0} \dfrac{\Phi_{|\beta|}(\varepsilon_k + t\tau, \widetilde{M}^k_{\beta\beta} + t\widetilde{H}_{\beta\beta}) - \Phi_{|\beta|}(\varepsilon_k, \widetilde{M}^k_{\beta\beta})}{t} & 0 \\ \widetilde{H}^T_{\alpha\gamma} \circ U^T_{\alpha\gamma} & 0 & 0 \end{bmatrix} P^T$$

with $\widetilde{H} = P^T H P$ and

$$\lim_{t\downarrow 0} \quad \frac{\Phi_{|\beta|}(\varepsilon_k + t\tau, \widetilde{M}^k_{\beta\beta} + t\widetilde{H}_{\beta\beta}) - \Phi_{|\beta|}(\varepsilon_k, \widetilde{M}^k_{\beta\beta})}{t}$$

$$= Q^k \left(\lim_{t\downarrow 0} \frac{\Phi_{|\beta|}(\varepsilon_k + t\tau, \widetilde{\Lambda}^k_\beta + t(Q^k)^T \widetilde{H}_{\beta\beta} Q^k) - \Phi_{|\beta|}(\varepsilon_k, \widetilde{\Lambda}^k_\beta)}{t} \right) (Q^k)^T$$

$$= Q^k \left[\, \widetilde{\Omega}^k \circ \left(\, (Q^k)^T \widetilde{H}_{\beta\beta} Q^k \, \right) + \tau \widetilde{S}^k \, \right] (Q^k)^T \,, \tag{30}$$

where we have used (21) for $\Phi_{|\beta|}$ and the fact that $\Phi_{|\beta|}$ is F-differentiable at $(\varepsilon_k, \widetilde{\Lambda}^k_\beta)$ because $\varepsilon_k^2 I + (\widetilde{\Lambda}^k_\beta)^2$ is nonsingular,

$$(\widetilde{\Omega}^k)_{ij} = \begin{cases} \dfrac{\phi(\varepsilon_k, \tilde{z}^k_i) - \phi(\varepsilon_k, \tilde{z}^k_j)}{\tilde{z}^k_i - \tilde{z}^k_j} & \text{if } \tilde{z}^k_i \neq \tilde{z}^k_j, \\ \phi'_{\tilde{z}^k_i}(\varepsilon, \tilde{z}^k_i) & \text{if } \tilde{z}^k_i = \tilde{z}^k_j, \end{cases} \qquad i, j = 1, \ldots, |\beta| \,,$$

and

$$\widetilde{S}^k = \begin{bmatrix} \phi'_\varepsilon(\varepsilon_k, \tilde{z}^k_1) & & \\ & \ddots & \\ & & \phi'_\varepsilon(\varepsilon_k, \tilde{z}^k_{|\beta|}) \end{bmatrix}.$$

Define

$$A^k = A + P \begin{bmatrix} 0 & 0 & 0 \\ 0 & \widetilde{M}^k_{\beta\beta} & 0 \\ 0 & 0 & 0 \end{bmatrix} P^T \quad \text{and} \quad \widetilde{A}^k = P^T A^k P = \begin{bmatrix} \Lambda_\alpha & 0 & 0 \\ 0 & \widetilde{M}^k_{\beta\beta} & 0 \\ 0 & 0 & \Lambda_\gamma \end{bmatrix}.$$

Since, for each k, $\varepsilon_k^2 I + (\widetilde{M}^k_{\beta\beta})^2$ is nonsingular, the matrix $\varepsilon_k^2 I + (A^k)^2 = P[\varepsilon_k^2 I + (\widetilde{A}^k)^2]P^T$ is also nonsingular. Thus, Φ is F-differentiable at (ε_k, A^k). Let

$$P^k \equiv [\; P^k_\alpha \quad P^k_\beta \quad P^k_\gamma \;] = [\; P_\alpha \quad P_\beta Q^k \quad P_\gamma \;]$$

and $\widetilde{\Lambda}^k$ be the diagonal matrix whose diagonal entries are components of $\tilde{\lambda}^k$. Then

$$A^k = P^k \widetilde{\Lambda}^k (P^k)^T \,,$$

which, together with (21), implies that for any $(\tau, H) \in \Re \times \mathcal{S}^n$, we have

$$\Phi'(\varepsilon_k, A^k)(\tau, H) = P^k \left[\widetilde{U}^k \circ ((P^k)^T H P^k) + \tau \widetilde{D}^k \right] (P^k)^T , \tag{31}$$

where $\widetilde{U}^k \equiv \widehat{U}(\varepsilon_k, \tilde{\lambda}^k)$ and $\widetilde{D}^k \equiv D(\varepsilon_k, \tilde{\lambda}^k)$. Since $\{Q^k\}$, $\{\widetilde{U}^k\}$, and $\{\widetilde{D}^k\}$ are all uniformly bounded, by taking subsequences if necessary, we may assume that all these three sequences converge. By simple computations, we obtain

$$\lim_{k\to\infty} \widetilde{U}^k_{ij} = \begin{cases} 1 & \text{if } i \in \alpha,\ j \in \alpha \cup \beta, \\ U_{ij} & \text{if } i \in \alpha,\ j \in \gamma, \\ \lim\limits_{k\to\infty} (\widetilde{\Omega}^k)_{(i-|\alpha|)(j-|\alpha|)} & \text{if } i \in \beta,\ j \in \beta, \\ 0 & \text{if } i \in \beta \cup \gamma,\ j \in \gamma, \end{cases} \qquad i, j = 1, \ldots, n,$$

and

$$\lim_{k\to\infty} \widetilde{D}^k = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \lim\limits_{k\to\infty} \widetilde{S}^k & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

which, together with (31), (29), and (30), imply that for any $(\tau, H) \in \Re \times \mathcal{S}^n$,

$$\lim_{k\to\infty} (P^k)^T \left[\Gamma'(\varepsilon_k, M^k)(\tau, H) - \Phi'(\varepsilon_k, A^k)(\tau, H) \right] P^k = 0 .$$

Consequently, we can conclude $V(\tau, H) = \lim_{k\to\infty} \Phi'(\varepsilon_k, A^k)(\tau, H)$ for all $(\tau, H) \in \Re \times \mathcal{S}^n$, which implies $V \in \partial_B \Phi(0, A)$. Hence, (27) holds. □

Lemma 4 allows us to completely characterize $\partial_B \Phi(0, A)$ (hence, $\partial \Phi(0, A)$).

PROPOSITION 5. *Suppose that $A \in \mathcal{S}^n$ has the spectral decomposition in* (7)*. Then a $V \in \partial_B \Phi(0, A)$ (respectively, $\partial \Phi(0, A)$) if and only if there exists a $V_{|\beta|} \in \partial_B \Phi_{|\beta|}(0, 0)$ (respectively, $\partial \Phi_{|\beta|}(0, 0)$) such that*

$$V(\tau, H) = P \begin{bmatrix} \widetilde{H}_{\alpha\alpha} & \widetilde{H}_{\alpha\beta} & U_{\alpha\gamma} \circ \widetilde{H}_{\alpha\gamma} \\ \widetilde{H}^T_{\alpha\beta} & V_{|\beta|}(\tau, \widetilde{H}_{\beta\beta}) & 0 \\ \widetilde{H}^T_{\alpha\gamma} \circ U^T_{\alpha\gamma} & 0 & 0 \end{bmatrix} P^T \tag{32}$$

for all $(\tau, H) \in \Re \times \mathcal{S}^n$, where $\widetilde{H} = P^T H P$.

Proof. We need only to prove that (32) holds for $V \in \partial_B \Phi(0, A)$ and $V_{|\beta|} \in \partial_B \Phi_{|\beta|}(0, 0)$ as the case for Clarke's generalized Jacobian can be proved similarly.

Let $\Psi(\tau, H) := (\tau, P^T H P)$ for any $(\tau, H) \in \Re \times \mathcal{S}^n$, and let $\Xi : \Re \times \mathcal{S}^n \to \mathcal{S}^n$ be defined by (25). Then, since $\Psi'(\tau, H) : \Re \times \mathcal{S}^n \to \Re \times \mathcal{S}^n$ is onto, we know from Lemma 1 that

$$\partial_B \Gamma(0, 0) = \partial_B \Xi(0, 0) \Psi'(0, 0) ,$$

which, together with (27) in Lemma 4, completes the proof. □

Just as in the case for the metric projector $\Pi_{\mathcal{S}^n_+}$, Proposition 5 says that in order to compute $\partial_B \Phi(0, A)$ and $\partial \Phi(0, A)$, one needs only to fix P and compute the corresponding easy part $\partial_B \Phi_{|\beta|}(0, 0)$ (hence, $\partial \Phi_{|\beta|}(0, 0)$). For any $(\varepsilon, z) \in \Re \times \Re^{|\beta|}$ with $\varepsilon^2 + z_i^2 > 0$ for all i, let $\widehat{\Omega}(\varepsilon, z)$ be defined by (20) with n and x being replaced by

$|\beta|$ and z, respectively. Then, by (22) and (21), one can readily draw the conclusion that $V_{|\beta|} \in \partial_B \Phi_{|\beta|}(0,0)$ if and only if there exist $Q \in \mathcal{Q}_{|\beta|}$ and $\Omega \in \widehat{\mathcal{U}}_{|\beta|}$ such that

$$V_{|\beta|}(0,Z) = Q\,[\,\Omega \circ (Q^T Z Q)\,]\,Q^T \quad \forall\, Z \in \mathcal{S}^{|\beta|}\,, \tag{33}$$

where

$$\widehat{\mathcal{U}}_{|\beta|} := \big\{\Omega \,:\, \Omega = \lim_{k\to\infty} \widehat{\Omega}(\varepsilon_k, z^k),\ (\varepsilon_k, z^k) \to (0,0),$$

$$(z^k)_1 \geq \cdots \geq (z^k)_{|\beta|},\ \varepsilon_k^2 + (z_i^k)^2 > 0\ \forall\, i\big\}.$$

Note that for any $\Omega \in \widehat{\mathcal{U}}^{\beta}$, it holds that $\Omega_{ij} \in [0,1], \quad i,j = 1,\ldots,|\beta|$.

The next proposition establishes a link between $\partial_B \Pi_{\mathcal{S}^n_+}(A)$ and $\partial_B \Phi(0,A)$, and so a link between $\partial \Pi_{\mathcal{S}^n_+}(A)$ and $\partial \Phi(0,A)$.

Proposition 6. *For any $V_0 \in \partial_B \Pi_{\mathcal{S}^n_+}(A)$, there exists $V \in \partial_B \Phi(0,A)$ such that*

$$V_0(H) = V(0,H) \quad \forall\, H \in \mathcal{S}^n\,. \tag{34}$$

Proof. By comparing Proposition 2, together with (17), with Proposition 5, together with (33), we can derive the conclusion directly. □

We conclude this section by presenting a useful inequality for elements in $\partial \Phi(0,A)$, which is analogous to (6) for the metric projector Π_K with $K = \mathcal{S}^n_+$.

Proposition 7. *For any $V \in \partial \Phi(0,A)$, it holds that*

$$\langle H - V(0,H), V(0,H)\rangle \geq 0 \quad \forall\, H \in \mathcal{S}^n\,. \tag{35}$$

Proof. Let $V \in \partial \Phi(0,A)$. Then, by Carathéodory's theorem, there exist a positive integer κ and $V^i \in \partial_B \Phi(0,A)$, $i = 1,\ldots,\kappa$, such that V is the convex combination of $V^1,\ldots,V^\kappa$. Let $t_1,\ldots,t_\kappa$ be such that $V = \sum_{i=1}^{\kappa} t_i V^i$, where $t_i \geq 0$, $i = 1,\ldots,\kappa$, and $\sum_{i=1}^{\kappa} t_i = 1$.

From Sun, Sun, and Qi [46, Proposition 3.1], we know that for each $i \in \{1,\ldots,\kappa\}$,

$$\langle H - V^i(0,H), V^i(0,H)\rangle \geq 0 \quad \forall\, H \in \mathcal{S}^n\,. \tag{36}$$

In order to prove that (35) holds for V, let $\theta(X) := \langle X, X\rangle$, $X \in \mathcal{S}^n$. By the convexity of θ, we have for any $H \in \mathcal{S}^n$ that

$$\theta(V(0,H)) = \theta\left(\sum_{i=1}^{\kappa} t_i V^i(0,H)\right) \leq \sum_{i=1}^{\kappa} t_i\,\theta(V^i(0,H)) = \sum_{i=1}^{\kappa} t_i \langle V^i(0,H), V^i(0,H)\rangle\,,$$

which, together with (36) and the definition of θ, implies

$$\langle V(0,H), V(0,H)\rangle \leq \sum_{i=1}^{\kappa} t_i \langle H, V^i(0,H)\rangle = \left\langle H, \sum_{i=1}^{\kappa} t_i V^i(0,H)\right\rangle = \langle H, V(0,H)\rangle\,.$$

Thus, (35) holds. □

3. Equivalent conditions. Let $\mathcal{X}$ and $\mathcal{Y}$ be two finite dimensional real vector spaces each equipped with a scalar product $\langle \cdot, \cdot \rangle$ and its induced norm $\| \cdot \|$. Let $g : \mathcal{X} \to \mathcal{Y}$ be a continuously differentiable function and K be a nonempty and closed convex set in $\mathcal{Y}$. Consider the following feasible problem:

$$g(x) \in K\,, \quad x \in \mathcal{X}\,. \tag{37}$$

Assume that $\bar{x} \in \mathcal{X}$ is a feasible solution to (37). Let $\mathcal{T}_K(g(\bar{x}))$ be the tangent cone of K and $\mathcal{N}_K(g(\bar{x}))$ be the normal cone of K at $g(\bar{x})$, respectively. We write $\mathrm{lin}(\mathcal{T}_K(g(\bar{x})))$ for the linearity space of $\mathcal{T}_K(g(\bar{x}))$. Then we can define the following nondegeneracy condition for problem (37).

DEFINITION 8. *We say that a feasible point $\bar{x}$ to problem* (37) *is constraint nondegenerate if*

$$g'(\bar{x})\mathcal{X} + \mathrm{lin}\big(\mathcal{T}_K(g(\bar{x}))\big) = \mathcal{Y}\,. \tag{38}$$

The concept of nondegeneracy for the abstract problem (37) first appeared in Robinson [33, 34]. The name "constraint nondegeneracy" was coined by Robinson in [35]. The nondegenerate constraint condition (38) including its various equivalent forms was extensively used in [7, 40] for sensitivity and stability analysis in optimization and variational inequalities. If $\mathcal{Y}$ is the Euclidean space $\Re^m$ and $K = \{0\}^{m_1} \times \Re^{m_2}_+$ with $m_1 + m_2 = m$, then the constraint nondegenerate condition (38) is equivalent to the well-known linear independence constraint qualification [33, 40]. Here we shall apply Definition 8 to both the SDP problem (1) and its dual (2) to define the primal constraint nondegeneracy and the dual constraint nondegeneracy, respectively.

DEFINITION 9. *We say that the primal constraint nondegeneracy holds at a feasible solution $\overline{X} \in \mathcal{S}^n_+$ to the SDP problem* (1) *if*

$$\begin{bmatrix} \mathcal{A} \\ \mathcal{I} \end{bmatrix} \mathcal{S}^n + \begin{bmatrix} \{0\} \\ \mathrm{lin}\big(\mathcal{T}_{\mathcal{S}^n_+}(\overline{X})\big) \end{bmatrix} = \begin{bmatrix} \Re^m \\ \mathcal{S}^n \end{bmatrix} \tag{39}$$

or, equivalently,

$$\mathcal{A}\,\mathrm{lin}\big(\mathcal{T}_{\mathcal{S}^n_+}(\overline{X})\big) = \Re^m\,, \tag{40}$$

where $\mathcal{I}$ is the identity mapping from $\mathcal{S}^n$ to $\mathcal{S}^n$. Similarly, we say that the dual constraint nondegeneracy holds at a feasible solution $(\bar{y}, \overline{S}) \in \Re^m \times \mathcal{S}^n_+$ to the dual problem (2) *if*

$$\begin{bmatrix} \mathcal{A}^* & \mathcal{I} \\ 0 & \mathcal{I} \end{bmatrix} \begin{pmatrix} \Re^m \\ \mathcal{S}^n \end{pmatrix} + \begin{bmatrix} \{0\} \\ \mathrm{lin}\big(\mathcal{T}_{\mathcal{S}^n_+}(\overline{S})\big) \end{bmatrix} = \begin{bmatrix} \mathcal{S}^n \\ \mathcal{S}^n \end{bmatrix} \tag{41}$$

or, equivalently,

$$\mathcal{A}^*\Re^m + \mathrm{lin}\big(\mathcal{T}_{\mathcal{S}^n_+}(\overline{S})\big) = \mathcal{S}^n\,. \tag{42}$$

Note that in the literature constraint nondegeneracy is called different names. Shapiro and Fan [41] and Shapiro [39] termed it transversality. Primal constraint nondegeneracy and dual constraint nondegeneracy are better known as primal nondegeneracy and dual nondegeneracy, respectively, in the interior point methods community. See, for example, Alizadeh, Haeber, and Overton [2]. To avoid potential

confusion, we will stick to Robinson's terminology here and interpret different usages of constraint nondegeneracy in terms of Definition 9.

Let $\overline{Z} \equiv (\overline{X}, \bar{y}, \overline{S}) \in \mathcal{S}^n \times \Re^m \times \mathcal{S}^n$ be a KKT point satisfying the KKT conditions (3). Since $\mathcal{S}^n_+$ is a self-dual cone, from [13] we know that

$$\begin{aligned} \mathcal{S}^n_+ \ni X \perp S \in \mathcal{S}^n_+ &\Longleftrightarrow -X \in \mathcal{N}_{\mathcal{S}^n_+}(S) \\ &\Longleftrightarrow S - \Pi_{\mathcal{S}^n_+}[S - X] = X - \Pi_{\mathcal{S}^n_+}[X - S] = 0\,. \end{aligned} \tag{43}$$

Therefore, $(\overline{X}, \bar{y}, \overline{S}) \in \mathcal{S}^n \times \Re^m \times \mathcal{S}^n$ satisfies (3) if and only if $(\overline{X}, \bar{y}, \overline{S})$ is a solution to the nonsmooth system of equations

$$F(X, y, S) \equiv \begin{bmatrix} C - \mathcal{A}^* y - S \\ \mathcal{A}X - b \\ S - \Pi_{\mathcal{S}^n_+}[S - X] \end{bmatrix} = \begin{bmatrix} C - \mathcal{A}^* y - S \\ \mathcal{A}X - b \\ X - \Pi_{\mathcal{S}^n_+}[X - S] \end{bmatrix} = 0\,, \tag{44}$$

where $(X, y, S) \in \mathcal{S}^n \times \Re^m \times \mathcal{S}^n$.

Note that both the KKT conditions (3) and the nonsmooth system (44) can be written as the following special generalized equation:

$$0 \in \begin{bmatrix} C - \mathcal{A}^* y - S \\ \mathcal{A}X - b \\ X \end{bmatrix} + \begin{bmatrix} \mathcal{N}_{\mathcal{S}^n}(X) \\ \mathcal{N}_{\Re^m}(y) \\ \mathcal{N}_{\mathcal{S}^n_+}(S) \end{bmatrix}. \tag{45}$$

In [32], Robinson introduced an important concept called strong regularity for a solution of generalized equations. Here we define only the strong regularity for (45) rather than for the general problems.

DEFINITION 10. *Let $\mathcal{Z} \equiv \mathcal{S}^n \times \Re^m \times \mathcal{S}^n$. We say that a KKT point $\overline{Z} \equiv (\overline{X}, \bar{y}, \overline{S}) \in \mathcal{Z}$ is a strongly regular solution of the generalized equation* (45) *if there exist neighborhoods $\mathcal{B}$ of the origin $0 \in \mathcal{Z}$ and $\mathcal{V}$ of $\overline{Z}$ such that for every $\delta \in \mathcal{B}$, the generalized equation*

$$\delta \in \begin{bmatrix} C - \mathcal{A}^* y - S \\ \mathcal{A}X - b \\ X \end{bmatrix} + \begin{bmatrix} \mathcal{N}_{\mathcal{S}^n}(X) \\ \mathcal{N}_{\Re^m}(y) \\ \mathcal{N}_{\mathcal{S}^n_+}(S) \end{bmatrix} \tag{46}$$

has a unique solution in $\mathcal{V}$, denoted by $Z_{\mathcal{V}}(\delta)$, and the mapping $Z_{\mathcal{V}} : \mathcal{B} \to \mathcal{V}$ is Lipschitz continuous.

Recall that F is said to be a locally Lipschitz homeomorphism near $\overline{Z}$ if there exists an open neighborhood $\mathcal{V}$ of $\overline{Z}$ such that the restricted mapping $F|_{\mathcal{V}}: \mathcal{V} \to F(\mathcal{V})$ is Lipschitz continuous and bijective, and its inverse is also Lipschitz continuous. The following result, which holds in a more general framework, shows that F is a locally Lipschitz homeomorphism near $\overline{Z}$ if and only if $\overline{Z}$ is a strongly regular solution of the generalized equation (45). This is almost intuitively true. For the sake of completeness, however, we include a short proof.

LEMMA 11. *Let $\mathcal{Z} \equiv \mathcal{S}^n \times \Re^m \times \mathcal{S}^n$. Let $F : \mathcal{Z} \to \mathcal{Z}$ be defined by* (44) *and $\overline{Z}$ be a KKT point of the SDP problem. Then, it holds that F is a locally Lipschitz homeomorphism near $\overline{Z}$ if and only if $\overline{Z}$ is a strongly regular solution of the generalized equation* (45).

Proof. "$\Longrightarrow$" Assume that F is a locally Lipschitz homeomorphism near $\overline{Z}$. Then, there exists an open neighborhood $\mathcal{V}$ of $\overline{Z}$ such that $F(\mathcal{V})$ is an open neighborhood of the origin $0 \in \mathcal{Z}$, and for any $\hat{\delta} \in F(\mathcal{V})$, the equation $F(Z) = \hat{\delta}$ has a unique solution $\widehat{Z}_{\mathcal{V}}(\hat{\delta})$ in $\mathcal{V}$ and $\widehat{Z}_{\mathcal{V}} : F(\mathcal{V}) \to \mathcal{V}$ is Lipschitz continuous.

For any $\delta = (\delta^1, \delta^2, \delta^3) \in \mathcal{B} \equiv \frac{1}{2}F(\mathcal{V})$, let $Z(\delta) = (X(\delta), y(\delta), S(\delta))$ be a solution, if one exists, to (46). Write $\delta \equiv (\delta^1, \delta^2, \delta^3) \in \mathcal{S}^n \times \Re^m \times \mathcal{S}^n$. Then we have

$$\begin{bmatrix} C - \mathcal{A}^* y(\delta) - S(\delta) \\ \mathcal{A}X(\delta) - b \\ (S(\delta) + \delta^3) - \Pi_{\mathcal{S}^n_+}[\,(S(\delta) + \delta^3) - X(\delta)\,] \end{bmatrix} = \begin{bmatrix} \delta^1 \\ \delta^2 \\ \delta^3 \end{bmatrix},$$

i.e.,

$$F(X(\delta), y(\delta), S(\delta) + \delta^3) = \begin{bmatrix} \delta^1 - \delta^3 \\ \delta^2 \\ \delta^3 \end{bmatrix}.$$

Then $Z(\delta)$ uniquely exists in $\mathcal{V}$ and

$$Z(\delta) = \widehat{Z}_{\mathcal{V}}(\delta^1 - \delta^3, \delta^2, \delta^3) - \begin{bmatrix} 0 \\ 0 \\ \delta^3 \end{bmatrix}.$$

Hence, $Z(\cdot)$ is Lipschitz continuous on $\mathcal{B}$.

"$\Longleftarrow$" Assume that $\overline{Z}$ is a strongly regular solution of the generalized equation (45). Then, there exist neighborhoods $\mathcal{B}$ of the origin $0 \in \mathcal{Z}$ and $\mathcal{V}$ of $\overline{Z}$, and a locally Lipschitz function $Z_{\mathcal{V}} : \mathcal{B} \to \mathcal{V}$ such that for any $\delta \in \mathcal{B}$, $Z_{\mathcal{V}}(\delta)$ is the unique solution in $\mathcal{V}$ to (46). By reversing the arguments in the first part of the proof, we can conclude that for any $\hat{\delta} \equiv (\hat{\delta}^1, \hat{\delta}^2, \hat{\delta}^3) \in (\frac{1}{2}\mathcal{B}) \cap (\,\mathcal{S}^n \times \Re^m \times \mathcal{S}^n\,)$, $F(Z) = \hat{\delta}$ has a unique solution

$$\widehat{Z}(\hat{\delta}) \in \mathcal{V} \text{ given by } \widehat{Z}(\hat{\delta}) = Z_{\mathcal{V}}(\hat{\delta}^1 + \hat{\delta}^3, \hat{\delta}^2, \hat{\delta}^3) + \begin{bmatrix} 0 \\ 0 \\ \hat{\delta}^3 \end{bmatrix},$$

which implies that $\widehat{Z}(\cdot)$ is Lipschitz continuous on $\frac{1}{2}\mathcal{B}$. Thus, F is Lipschitz homeomorphism near $\overline{Z}$. □

The concept of strong regularity for general nonlinear semidefinite programming is closely related to another concept called the strong second order sufficient condition as shown by Sun in [42]. Here we will only present the strong second order sufficient condition in terms of the SDP problem (1). First, for any $B \in \mathcal{S}^n$, we define a linear-quadratic function $\Upsilon_B : \mathcal{S}^n \times \mathcal{S}^n \to \Re$.

DEFINITION 12 ([42, Definition 2.1]). *For any given $B \in \mathcal{S}^n$, define the linear-quadratic function $\Upsilon_B : S^n \times \mathcal{S}^n \to \Re$, which is linear in the first argument and quadratic in the second argument, by*

$$\Upsilon_B(S, H) := 2\langle S, HB^\dagger H\rangle, \quad (S, H) \in \mathcal{S}^n \times \mathcal{S}^n,$$

where $B^\dagger$ is the Moore–Penrose pseudoinverse of B.

Let $\overline{X} \in \mathcal{S}^n_+$ be an optimal solution to the SDP problem (1). Denote $\mathcal{M}(\overline{X})$ by the set of points $(y, S) \in \Re^m \times \mathcal{S}^n$ such that $(\overline{X}, y, S)$ is a KKT point, i.e., for any

$(y,S)\in\mathcal{M}(\overline{X})$, $(\overline{X},y,S)$ satisfies the KKT conditions (3). Let $(\bar{y},\overline{S})\in\mathcal{M}(\overline{X})$. Write $\overline{A}\equiv\overline{X}-\overline{S}$. By using the fact that $\mathcal{S}^n_+\ni\overline{X}\perp\overline{S}\in\mathcal{S}^n_+$, we may assume that $\overline{A}$ has the spectral decomposition as in (7) by replacing A with $\overline{A}$ and

(47)
$$\overline{A}=P\begin{bmatrix}\Lambda_\alpha & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & \Lambda_\gamma\end{bmatrix}P^T,\ \overline{X}=P\begin{bmatrix}\Lambda_\alpha & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}P^T,\ \overline{S}=P\begin{bmatrix}0 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & -\Lambda_\gamma\end{bmatrix}P^T.$$

Write $P=[\ P_\alpha\quad P_\beta\quad P_\gamma\ ]$. Then, according to (10) and (12), we have

$$\text{(48)}\qquad \mathrm{lin}\left(\mathcal{T}_{\mathcal{S}^n_+}(\overline{X})\right)=\{B\in\mathcal{S}^n\,:\,P_\beta^TBP_\beta=0,\ \ P_\beta^TBP_\gamma=0,\ \ P_\gamma^TBP_\gamma=0\},$$

$$\text{(49)}\qquad \mathrm{lin}\left(\mathcal{T}_{\mathcal{S}^n_+}(\overline{S})\right)=\{B\in\mathcal{S}^n\,:\,P_\alpha^TBP_\alpha=0,\ \ P_\alpha^TBP_\beta=0,\ \ P_\beta^TBP_\beta=0\},$$

and

$$\mathrm{aff}\left(\mathcal{C}(\overline{A};\mathcal{S}^n_+)\right)=\left\{B\in\mathcal{S}^n\,:\,P_\beta^TBP_\gamma=0,\ P_\gamma^TBP_\gamma=0\right\}.$$

Define

$$\begin{aligned}\mathrm{app}(\bar{y},\overline{S}) &:= \left\{B\in\mathcal{S}^n\,:\,\mathcal{A}B=0,\ \ B\in\mathrm{aff}\left(\mathcal{C}(\overline{A};\mathcal{S}^n_+)\right)\right\}\\ &= \{B\in\mathcal{S}^n\,:\,\mathcal{A}B=0,\ \ P_\beta^TBP_\gamma=0,\ \ P_\gamma^TBP_\gamma=0\}.\end{aligned}\qquad\text{(50)}$$

Then we can state the strong second order sufficient condition for the SDP problem tailored from Sun [42] for the general nonlinear SDP problem.

DEFINITION 13. *Let $\overline{X}\in\mathcal{S}^n_+$ be an optimal solution to the SDP problem* (1). *We say that the strong second order sufficient condition holds at $\overline{X}$ if*

$$\text{(51)}\qquad \sup_{(y,S)\in\mathcal{M}(\overline{X})}\{-\Upsilon_{\overline{X}}(-S,H)\}>0\quad\forall\,0\neq H\in\Big\{\bigcap_{(y,S)\in\mathcal{M}(\overline{X})}\mathrm{app}(y,S)\Big\}.$$

The strong second order sufficient condition (51) may look very complicated. When $\mathcal{M}(\overline{X})$ is a singleton, the following result gives a very simple characterization.

LEMMA 14. *Let $\overline{X}\in\mathcal{S}^n_+$ be an optimal solution to the SDP problem* (1). *Assume that $\mathcal{M}(\overline{X})=\{(\bar{y},\overline{S})\}$. Let $\overline{X}$ and $\overline{S}$ have the spectral decompositions as in* (47). *Then the strong second order sufficient condition* (51) *holds at $\overline{X}$ if and only if, for any $H\in\mathcal{S}^n$, the following conditions hold*

$$\text{(52)}\qquad \mathcal{A}H=0,\ P_\beta^THP_\gamma=0,\ P_\gamma^THP_\gamma=0,\ \textit{and}\ P_\alpha^THP_\gamma=0\implies H=0.$$

Proof. For any $H\in\mathcal{S}^n$, write $\widetilde{H}=P^THP$. Since $\mathcal{M}(\overline{X})=\{(\bar{y},\overline{S})\}$, the strong second order sufficient condition (51) becomes

$$-\Upsilon_{\overline{X}}(-\overline{S},H)\ >\ 0\quad\forall\,H\in\mathrm{app}(\bar{y},\overline{S})\backslash\{0\},$$

which, by the definition of $\Upsilon_{\overline{X}}(-\overline{S},H)$ and (47), is equivalent to

$$2\sum_{i\in\alpha,j\in\gamma}\frac{-\lambda_j}{\lambda_i}(\widetilde{H}_{ij})^2>0\qquad\forall\,H\in\mathrm{app}(\bar{y},\overline{S})\backslash\{0\}.$$

For details, see [42]. Then, by (50), the strong second order sufficient condition (51) holds at $\overline{X}$ if and only if

$$\mathcal{A}H = 0,\ \widetilde{H}_{\beta\gamma} = 0,\ \widetilde{H}_{\gamma\gamma} = 0,\ \text{and}\ H \neq 0 \implies \widetilde{H}_{\alpha\gamma} \neq 0 \quad \forall\, H \in \mathcal{S}^n\,,$$

which is equivalent to (52). This completes the proof. □

Next, we shall establish a link between the strong second order sufficient condition and the dual constraint nondegeneracy.[2]

PROPOSITION 15. *Let $\overline{X} \in \mathcal{S}^n_+$ be an optimal solution to the SDP problem* (1). *Under the assumption $\mathcal{M}(\overline{X}) = \{(\bar{y}, \overline{S})\}$, the following are equivalent:*

(i) *The strong second order sufficient condition* (51) *holds at $\overline{X}$.*

(ii) *The dual constraint nondegenerate condition* (42) *holds at $(\bar{y}, \overline{S})$.*

Proof. Let $\overline{X}$ and $\overline{S}$ have the spectral decompositions as in (47). For any $H \in \mathcal{S}^n$, let $\widetilde{H} = P^T H P$. We prove "(i) $\Longrightarrow$ (ii)" first. By Lemma 14, (i) holds if and only if we have the following implication:

$$\mathcal{A}H = 0,\ \widetilde{H}_{\beta\gamma} = 0,\ \widetilde{H}_{\gamma\gamma} = 0,\ \text{and}\ \widetilde{H}_{\alpha\gamma} = 0 \Longrightarrow H = 0\ \forall\, H \in \mathcal{S}^n\,. \tag{53}$$

Suppose, for the sake of contradiction, that the dual constraint nondegenerate condition (42) does not hold at $(\bar{y}, \overline{S})$. Then, we have

$$[\,\mathcal{A}^*\Re^m\,]^\perp \cap \big[\,\mathrm{lin}\big(\mathcal{T}_{\mathcal{S}^n_+}(\overline{S})\big)\big]^\perp \neq \{0\}. \tag{54}$$

Take an arbitrary $0 \neq \overline{H} \in [\,\mathcal{A}^*\Re^m\,]^\perp \cap \big[\,\mathrm{lin}\big(\mathcal{T}_{\mathcal{S}^n_+}(\overline{S})\big)\,\big]^\perp$. We obtain from $\overline{H} \in [\,\mathcal{A}^*\Re^m\,]^\perp$ that

$$\langle \overline{H}, \mathcal{A}^* y\rangle = 0\ \ \forall\, y \in \Re^m \Longrightarrow \langle \mathcal{A}\overline{H}, y\rangle = 0\ \ \forall\, y \in \Re^m \Longrightarrow \mathcal{A}\overline{H} = 0 \tag{55}$$

and from $\overline{H} \in \big[\,\mathrm{lin}\big(\mathcal{T}_{\mathcal{S}^n_+}(\overline{S})\big)\,\big]^\perp$ that

$$\langle P^T\overline{H}P, P^T B P\rangle = \langle \overline{H}, B\rangle = 0\ \ \forall\, B \in \mathrm{lin}\big(\mathcal{T}_{\mathcal{S}^n_+}(\overline{S})\big)\,,$$

which, together with (49), implies

$$P_\alpha^T \overline{H} P_\gamma = 0, \quad P_\beta^T \overline{H} P_\gamma = 0, \quad \text{and}\ \ P_\gamma^T \overline{H} P_\gamma = 0\,. \tag{56}$$

By making use of (53), (55), and (56), we obtain $\overline{H} = 0$, which contradicts the choice of $\overline{H}$. This contradiction shows that (ii) holds.

Next, we show "(ii) $\Longrightarrow$ (i)." Since the dual constraint nondegenerate condition (42) holds at $(\bar{y}, \overline{S})$, for any $H \in \mathcal{S}^n$ such that $\mathcal{A}H = 0$, $\widetilde{H}_{\beta\gamma} = 0$, $\widetilde{H}_{\gamma\gamma} = 0$, and $\widetilde{H}_{\alpha\gamma} = 0$, there exist $y \in \Re^m$ and $S \in \mathrm{lin}\big(\mathcal{T}_{\mathcal{S}^n_+}(\overline{S})\big)$ such that

$$H = \mathcal{A}^* y + S\,,$$

which, together with (49), implies

$$\begin{aligned}
\langle H, H\rangle &= \langle H, \mathcal{A}^* y + S\rangle = \langle \mathcal{A}H, y\rangle + \langle H, S\rangle = 0 + \langle P^T H P, P^T S P\rangle \\
&= \left\langle \begin{bmatrix} \widetilde{H}_{\alpha\alpha} & \widetilde{H}_{\alpha\beta} & 0 \\ \widetilde{H}_{\alpha\beta}^T & \widetilde{H}_{\beta\beta} & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & P_\alpha^T S P_\gamma \\ 0 & 0 & P_\beta^T S P_\gamma \\ P_\gamma^T S P_\alpha & P_\gamma^T S P_\beta & P_\gamma^T S P_\gamma \end{bmatrix} \right\rangle = 0\,.
\end{aligned}$$

[2] A similar statement for the dual SDP problem (2) also holds. We omit it here for brevity.

Therefore, by Lemma 14, it follows that (i) holds. □

Let $(\overline{X}, \bar{y}, \overline{S}) \in \mathcal{S}^n \times \Re^m \times \mathcal{S}^n$ be a KKT point satisfying the KKT conditions (3), and let F be defined by (44). As we menioned in the introduction, by combining the two papers [14] and [18], we know that if the primal constraint nondegeneracy holds at $\overline{X}$ and the dual constraint nondegeneracy holds at $(\bar{y}, \overline{S})$, then every element in $\partial_B F(\overline{X}, \bar{y}, \overline{S})$ is nonsingular. Actually, Proposition 15 and [42, Proposition 3.2] allow us to prove even the nonsingularity of Clarke's generalized Jacobian $\partial F(\overline{X}, \bar{y}, \overline{S})$ under the same primal and dual constraint nondegenerate conditions.

PROPOSITION 16. *Let $(\overline{X}, \bar{y}, \overline{S}) \in \mathcal{S}^n \times \Re^m \times \mathcal{S}^n$ be a KKT point. Assume that the primal constraint nondegenerate condition* (40) *holds at $\overline{X}$ and the dual constraint nondegenerate condition* (42) *holds at $(\bar{y}, \overline{S})$, respectively. Then, every element in $\partial F(\overline{X}, \bar{y}, \overline{S})$ is nonsingular.*

Proof. Since the primal constraint nondegenerate condition (40) implies that $\mathcal{M}(\overline{X}) = \{(\bar{y}, \overline{S})\}$, we know from Proposition 15 that the strong second order sufficient condition (51) holds at $\overline{X}$. Consequently, by [42, Proposition 3.2], every element in $\partial F(\overline{X}, \bar{y}, \overline{S})$ is nonsingular. □

Proposition 16 says that the primal and dual constraint nondegenerate conditions are sufficient for the nonsingularity of all elements in $\partial F(\overline{X}, \bar{y}, \overline{S})$. Next, we shall show that the nonsingularity of only two elements in $\partial_B F(\overline{X}, \bar{y}, \overline{S})$ will imply both the primal and dual constraint nondegenerate conditions.

From Lemma 1, we know that $W \in \partial_B F(\overline{X}, \bar{y}, \overline{S})$ if and only if there exists a $V \in \partial_B \Pi_{\mathcal{S}^n_+}(\overline{A})$ such that

$$W(\Delta X, \Delta y, \Delta S) = \begin{bmatrix} -\mathcal{A}^*(\Delta y) - \Delta S \\ \mathcal{A}(\Delta X) \\ \Delta X - V(\Delta X - \Delta S) \end{bmatrix} \tag{57}$$

for all $(\Delta X, \Delta y, \Delta S) \in \mathcal{S}^n \times \Re^m \times \mathcal{S}^n$, where $\overline{A} \equiv \overline{X} - \overline{S}$. Let $\mathrm{ex}\,(\partial_B \Pi_{\mathcal{S}^n_+}(\overline{A}))$ be defined by (18). For $V^0, V^{\mathcal{I}} \in \mathrm{ex}\,(\partial_B \Pi_{\mathcal{S}^n_+}(\overline{A}))$, let W^0 and $W^{\mathcal{I}}$ be defined by (57), respectively. Denote

$$\mathrm{ex}\,(\partial_B F(\overline{X}, \bar{y}, \overline{S})) := \left\{ W^0, W^{\mathcal{I}} \right\} \subseteq \partial_B F(\overline{X}, \bar{y}, \overline{S}). \tag{58}$$

PROPOSITION 17. *Let $(\overline{X}, \bar{y}, \overline{S}) \in \mathcal{S}^n \times \Re^m \times \mathcal{S}^n$ be a KKT point. If both W^0 and $W^{\mathcal{I}}$ in $\mathrm{ex}\,(\partial_B F(\overline{X}, \bar{y}, \overline{S}))$ are nonsingular, then the primal constraint nondegenerate condition* (40) *holds at $\overline{X}$ and the dual constraint nondegenerate condition* (42) *holds at $(\bar{y}, \overline{S})$, respectively.*

Proof. First we show that the nonsingularity of W^0 implies the primal constraint nondegenerate condition (40). Assume on the contrary that (40) does not hold. Since, equivalently, (39) fails to hold, too, we have

$$\left\{ \begin{bmatrix} \mathcal{A} \\ I \end{bmatrix} \mathcal{S}^n \right\}^{\perp} \cap \begin{bmatrix} 0 \\ \mathrm{lin}\,(\mathcal{T}_{\mathcal{S}^n_+}(\overline{X})) \end{bmatrix}^{\perp} \neq \begin{bmatrix} 0 \\ 0 \end{bmatrix} \in \begin{bmatrix} \Re^m \\ \mathcal{S}^n \end{bmatrix},$$

which implies that there exists $0 \neq (\Delta y, \Delta S) \in \{[\begin{smallmatrix} \mathcal{A} \\ I \end{smallmatrix}] \mathcal{S}^n\}^{\perp} \cap \left[\begin{smallmatrix} 0 \\ \mathrm{lin}\,(\mathcal{T}_{\mathcal{S}^n_+}(\overline{X})) \end{smallmatrix}\right]^{\perp}$. We obtain from $(\Delta y, \Delta S) \in \{[\begin{smallmatrix} \mathcal{A} \\ I \end{smallmatrix}] \mathcal{S}^n\}^{\perp}$ that

$$\left\langle (\Delta y, \Delta S), (\mathcal{A}H, H) \right\rangle = 0 \quad \forall\, H \in \mathcal{S}^n \Longrightarrow \mathcal{A}^*(\Delta y) + \Delta S = 0, \tag{59}$$

and from $(\Delta y, \Delta S) \in \begin{bmatrix} 0 \\ \mathrm{lin}\,(\mathcal{T}_{\mathcal{S}^n_+}(\overline{X})) \end{bmatrix}^{\perp}$ we obtain that

$$\left\langle P^T(\Delta S)P, P^T H P \right\rangle = \left\langle \Delta S, H \right\rangle = 0 \quad \forall\, H \in \mathrm{lin}\left(\mathcal{T}_{\mathcal{S}^n_+}(\overline{X})\right),$$

which, together with (48), implies

$$P_\alpha^T(\Delta S)P_\alpha = 0, \quad P_\alpha^T(\Delta S)P_\beta = 0, \quad \text{and} \quad P_\alpha^T(\Delta S)P_\gamma = 0\,. \tag{60}$$

Let $U \in \mathcal{S}^n$ be defined by (8). Recall from Proposition 2 that for $V^0 \in \mathrm{ex}\,(\partial_B \Pi_{\mathcal{S}^n_+}(\overline{A}))$, it holds that

$$V^0(\Delta S) = P \begin{bmatrix} P_\alpha^T(\Delta S)P_\alpha & P_\alpha^T(\Delta S)P_\beta & U_{\alpha\gamma} \circ (P_\alpha^T(\Delta S)P_\gamma) \\ \left(P_\alpha^T(\Delta S)P_\beta\right)^T & 0 & 0 \\ \left(P_\alpha^T(\Delta S)P_\gamma\right)^T \circ U_{\alpha\gamma}^T & 0 & 0 \end{bmatrix} P^T,$$

which, together with (60), implies $V^0(\Delta S) = 0 \in \mathcal{S}^n$. Therefore, by (57) and (59), we have for $\Delta X \equiv 0$ that

$$W^0(\Delta X, \Delta y, \Delta S) = \begin{bmatrix} -\mathcal{A}^*(\Delta y) - \Delta S \\ \mathcal{A}\,(\Delta X) \\ \Delta X - V^0(\Delta X - \Delta S) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ V^0(\Delta S) \end{bmatrix} = 0\,,$$

which implies that W^0 is singular. This contradiction shows that the primal constraint nondegenerate condition (40) holds at $\overline{X}$.

Next, we show that the nonsingularity of $W^{\mathcal{I}}$ implies the dual constraint nondegenerate condition (42). Suppose not. Then,

$$[\mathcal{A}^*\Re^m]^{\perp} \cap \left[\,\mathrm{lin}\left(\mathcal{T}_{\mathcal{S}^n_+}(\overline{S})\right)\,\right]^{\perp} \neq \{0\}\,.$$

Let $0 \neq \Delta X \in [\mathcal{A}^*\Re^m]^{\perp} \cap \left[\,\mathrm{lin}\left(\mathcal{T}_{\mathcal{S}^n_+}(\overline{S})\right)\,\right]^{\perp}$. We obtain from $\Delta X \in [\mathcal{A}^*\Re^m]^{\perp}$ that

$$\langle \Delta X, \mathcal{A}^* y \rangle = 0 \quad \forall\, y \in \Re^m \Longrightarrow \mathcal{A}(\Delta X) = 0 \tag{61}$$

and from $\Delta X \in \left[\,\mathrm{lin}\left(\mathcal{T}_{\mathcal{S}^n_+}(\overline{S})\right)\,\right]^{\perp}$ that

$$\left\langle P^T(\Delta X)P, P^T S P \right\rangle = \left\langle \Delta X, S \right\rangle = 0 \quad \forall\, S \in \mathrm{lin}\left(\mathcal{T}_{\mathcal{S}^n_+}(\overline{S})\right),$$

which, together with (49), implies

$$P_\alpha^T(\Delta X)P_\gamma = 0, \quad P_\beta^T(\Delta X)P_\gamma = 0, \quad \text{and} \quad P_\gamma^T(\Delta X)P_\gamma = 0\,. \tag{62}$$

From Proposition 2, for $V^{\mathcal{I}} \in \mathrm{ex}\,(\partial_B \Pi_{\mathcal{S}^n_+}(\overline{A}))$, it holds that

$$V^{\mathcal{I}}(\Delta X) = P \begin{bmatrix} P_\alpha^T(\Delta X)P_\alpha & P_\alpha^T(\Delta X)P_\beta & U_{\alpha\gamma} \circ (P_\alpha^T(\Delta X)P_\gamma) \\ \left(P_\alpha^T(\Delta X)P_\beta\right)^T & P_\beta^T(\Delta X)P_\beta & 0 \\ \left(P_\alpha^T(\Delta X)P_\gamma\right)^T \circ U_{\alpha\gamma}^T & 0 & 0 \end{bmatrix} P^T,$$

which, together with (62), implies $V^{\mathcal{I}}(\Delta X) = \Delta X$. Therefore, by (57) and (61), we have for $(\Delta y, \Delta S) \equiv (0,0) \in \Re^m \times \mathcal{S}^n$ that

$$W^{\mathcal{I}}(\Delta X, \Delta y, \Delta S) = \begin{bmatrix} -\mathcal{A}^*(\Delta y) - \Delta S \\ \mathcal{A}(\Delta X) \\ \Delta X - V^{\mathcal{I}}(\Delta X - \Delta S) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \Delta X - V^{\mathcal{I}}(\Delta X) \end{bmatrix} = 0\,,$$

which implies that $W^{\mathcal{I}}$ is singular. This contradiction shows that the dual constraint nondegenerate condition (42) holds at $(\bar{y}, \overline{S})$. This completes the proof. ∎

Now, we are ready to state our main result of this paper.

THEOREM 18. *Let $(\overline{X}, \bar{y}, \overline{S}) \in \mathcal{S}^n \times \Re^m \times \mathcal{S}^n$ be a KKT point satisfying the KKT conditions* (3)*, and let F be defined by* (44)*. Then, the following are all equivalent:*

(i) *The KKT point $(\overline{X}, \bar{y}, \overline{S})$ is a strongly regular solution of the generalized equation* (45)*.*

(ii) *The function F is a locally Lipschitz homeomorphism near $(\overline{X}, \bar{y}, \overline{S})$.*

(iii) *The primal constraint nondegenerate condition* (40) *holds at $\overline{X}$, and the dual constraint nondegenerate condition* (42) *holds at $(\bar{y}, \overline{S})$.*

(iv) *Every element in $\partial F(\overline{X}, \bar{y}, \overline{S})$ is nonsingular.*

(v) *Every element in $\partial_B F(\overline{X}, \bar{y}, \overline{S})$ is nonsingular.*

(vi) *The two elements in $\mathrm{ex}\left(\partial_B F(\overline{X}, \bar{y}, \overline{S})\right)$ are nonsingular.*

Proof. We already know from Lemma 11 that (i) $\Longleftrightarrow$ (ii) and from Propositions 16 and 17 that (iii) $\Longleftrightarrow$ (iv) $\Longleftrightarrow$ (v) $\Longleftrightarrow$ (vi). Furthermore, Clarke's inverse function theorem for Lipschitz functions [11, 12] implies that (iv) $\Longrightarrow$ (ii). The proof of this theorem will be complete if one can show that (ii) $\Longrightarrow$ (v). However, the latter has been known to be true since 1991 [21] (Gowda [15] even obtained a stronger conclusion than this by employing the degree theory). ∎

Remark 19. Note that the relations (i) $\Longleftrightarrow$ (ii) $\Longleftrightarrow$ (iv) even hold for the general nonlinear semidefinite programming case [42, Theorem 4.1], whose proof further relies on a number of important results achieved by Bonnans and Shapiro in their excellent monograph [7] on sensitivity analysis in optimization and variational inequalities. Here, the structure displayed uniquely by the SDP problem (1) allows us to derive these relations directly by avoiding the detour employed in [42] for the nonlinear SDP problem. An SDP example satisfying (iii) but with the strict complementary condition failing to hold can be found in [2]. See also [20].

4. Quadratic convergence of smoothing Newton methods. In this section, we shall show how the theoretical results obtained in sections 2 and 3 can be used to provide a quadratic convergence analysis on smoothing Newton methods for solving the nonsmooth equation $F(X, y, S) = 0$, where F is defined by (44). Let $\Phi : \Re \times \mathcal{S}^n \to \mathcal{S}^n$ be defined by (19). We then introduce the following smoothing function for F:

$$G(\varepsilon, X, y, S) \equiv \begin{bmatrix} C - \mathcal{A}^* y - S \\ \mathcal{A}X - b \\ S - \Phi(\varepsilon, S - X) \end{bmatrix} = \begin{bmatrix} C - \mathcal{A}^* y - S \\ \mathcal{A}X - b \\ X - \Phi(\varepsilon, X - S) \end{bmatrix}, \tag{63}$$

where $(\varepsilon, X, y, S) \in \Re \times \mathcal{S}^n \times \Re^m \times \mathcal{S}^n$. The above function G is continuously differentiable around any $(\varepsilon, X, y, S) \in \Re \times \mathcal{S}^n \times \Re^m \times \mathcal{S}^n$ when $\varepsilon \neq 0$ and has been used by several authors [9, 10, 20, 46] to design smoothing Newton methods for solving SDP problems (1) and (2).

Define $E : \Re \times \mathcal{S}^n \times \Re^m \times \mathcal{S}^n \to \Re \times \mathcal{S}^n \times \Re^m \times \mathcal{S}^n$ by

$$E(\varepsilon, X, y, S) \equiv \begin{bmatrix} \varepsilon \\ G(\varepsilon, X, y, S) \end{bmatrix}, \quad (\varepsilon, X, y, S) \in \Re \times \mathcal{S}^n \times \Re^m \times \mathcal{S}^n. \tag{64}$$

Then we have

$$F(X, y, S) = 0 \Longleftrightarrow E(\varepsilon, X, y, S) = 0 \quad \forall\, (\varepsilon, X, y, S) \in \Re \times \mathcal{S}^n \times \Re^m \times \mathcal{S}^n.$$

Let $(\overline{X}, \bar{y}, \overline{S}) \in \mathcal{S}^n \times \Re^m \times \mathcal{S}^n$ be a KKT point satisfying the KKT conditions (3). Then

$$E(0, \overline{X}, \bar{y}, \overline{S}) = 0\,.$$

Write $\overline{A} \equiv \overline{X} - \overline{S}$. Let $\overline{A}$, $\overline{X}$, and $\overline{S}$ have the spectral decompositions as in (47). Let the linear-quadratic function $\Upsilon_{\overline{X}}(\cdot,\cdot)$ be defined as in Definition 12. Then, we have the following result, which will play a key role in our analysis of quadratic convergence of smoothing Newton methods.

PROPOSITION 20. *Let $V \in \partial\Phi(0, \overline{A})$. Then, for any ΔX and ΔS in $\mathcal{S}^n$ such that $\Delta X = V(0, \Delta X - \Delta S)$, it holds that*

$$\langle \Delta X, \Delta S\rangle \le \Upsilon_{\overline{X}}(-\overline{S}, \Delta X)\,. \tag{65}$$

Proof. Let ΔX and ΔS be in $\mathcal{S}^n$ such that $\Delta X = V(0, \Delta X - \Delta S)$. Write $\Delta\widetilde{X} \equiv P^T(\Delta X)P$ and $\Delta\widetilde{S} \equiv P^T(\Delta S)P$. Let $\Phi_{|\beta|}$ be defined by (24). Then, by Proposition 5, there exists $V_{|\beta|} \in \partial\Phi_{|\beta|}(0,0)$ such that

$$V(0, \Delta X - \Delta S) = P \begin{bmatrix} \Delta\widetilde{H}_{\alpha\alpha} & \Delta\widetilde{H}_{\alpha\beta} & U_{\alpha\gamma} \circ \Delta\widetilde{H}_{\alpha\gamma} \\ (\Delta\widetilde{H}_{\alpha\beta})^T & V_{|\beta|}(0, \Delta\widetilde{H}_{\beta\beta}) & 0 \\ (\Delta\widetilde{H}_{\alpha\gamma})^T \circ U_{\alpha\gamma}^T & 0 & 0 \end{bmatrix} P^T\,,$$

where $\Delta\widetilde{H} \equiv \Delta\widetilde{X} - \Delta\widetilde{S}$ and $U \in \mathcal{S}^n$ is defined by (8). Thus, by using $\Delta X = V(0, \Delta X - \Delta S)$, we obtain

$$\Delta\widetilde{S}_{\alpha\alpha} = 0, \quad \Delta\widetilde{S}_{\alpha\beta} = 0, \quad \Delta\widetilde{X}_{\beta\gamma} = 0, \quad \Delta\widetilde{X}_{\gamma\gamma} = 0, \tag{66}$$

$$\Delta\widetilde{X}_{\beta\beta} = V_{|\beta|}(0, \Delta\widetilde{X}_{\beta\beta} - \Delta\widetilde{S}_{\beta\beta})\,, \tag{67}$$

and

$$\Delta\widetilde{X}_{\alpha\gamma} - U_{\alpha\gamma} \circ \Delta\widetilde{X}_{\alpha\gamma} = -U_{\alpha\gamma} \circ \Delta\widetilde{S}_{\alpha\gamma}\,. \tag{68}$$

By applying Proposition 7 to $\Phi_{|\beta|}$ and using (67), we obtain

$$\begin{aligned} &\langle \Delta\widetilde{X}_{\beta\beta}, -\Delta\widetilde{S}_{\beta\beta}\rangle \\ &= \big\langle V_{|\beta|}(0, \Delta\widetilde{X}_{\beta\beta} - \Delta\widetilde{S}_{\beta\beta}), (\Delta\widetilde{X}_{\beta\beta} - \Delta\widetilde{S}_{\beta\beta}) - V_{|\beta|}(0, \Delta\widetilde{X}_{\beta\beta} - \Delta\widetilde{S}_{\beta\beta})\big\rangle \ge 0\,, \end{aligned} \tag{69}$$

Therefore, from (66), (68), and (69), we have

$$\begin{aligned} \langle \Delta X, \Delta S\rangle &= \langle \Delta\widetilde{X}, \Delta\widetilde{S}\rangle \\ &= \langle \Delta\widetilde{X}_{\beta\beta}, \Delta\widetilde{S}_{\beta\beta}\rangle + 2\langle \Delta\widetilde{X}_{\alpha\gamma}, \Delta\widetilde{S}_{\alpha\gamma}\rangle \\ &\le 2\langle \Delta\widetilde{X}_{\alpha\gamma}, \Delta\widetilde{S}_{\alpha\gamma}\rangle \\ &= 2\sum_{i\in\alpha, j\in\gamma} \frac{\lambda_j}{\lambda_i}((\Delta\widetilde{X})_{ij})^2\,, \end{aligned}$$

which, together with the fact that

$$\Upsilon_{\overline{X}}(-\overline{S}, \Delta X) = 2\sum_{i\in\alpha, j\in\gamma} \frac{\lambda_j}{\lambda_i}((\Delta\widetilde{X})_{ij})^2\,,$$

shows that (65) holds. □

The following result relates the nonsingularity of $\partial_B E(0, \overline{X}, \bar{y}, \overline{S})$ and $\partial E(0, \overline{X}, \bar{y}, \overline{S})$ to both the primal constraint nondegeneracy and the dual constraint nondegeneracy.

PROPOSITION 21. *Let $(\overline{X}, \bar{y}, \overline{S}) \in \mathcal{S}^n \times \Re^m \times \mathcal{S}^n$ be a KKT point satisfying the KKT conditions* (3)*, and let E be defined by* (64)*. Then the following are equivalent:*

(i) *The primal constraint nondegenerate condition* (40) *holds at $\overline{X}$, and the dual constraint nondegenerate condition* (42) *holds at $(\bar{y}, \overline{S})$.*

(ii) *Every element in $\partial_B E(0, \overline{X}, \bar{y}, \overline{S})$ is nonsingular.*

(iii) *Every element in $\partial E(0, \overline{X}, \bar{y}, \overline{S})$ is nonsingular.*

Proof. Since "(iii) $\Longrightarrow$ (ii)" holds trivially and "(ii) $\Longrightarrow$ (i)" follows from Proposition 6 and Theorem 18 directly, we need only to show "(i) $\Longrightarrow$ (iii)." So in the remaining part of our proof we always assume that part (i) holds.

Let W be an arbitrary element in $\partial E(0, \overline{X}, \bar{y}, \overline{S})$. We need to show that W is nonsingular. Let $(\Delta\varepsilon, \Delta X, \Delta y, \Delta S) \in \Re \times \mathcal{S}^n \times \Re^m \times \Re^n$ be such that

$$W(\Delta\varepsilon, \Delta X, \Delta y, \Delta S) = 0.$$

Then, by Lemma 1, there exists $V \in \partial\Phi(0, \overline{A})$ such that

$$W(\Delta\varepsilon, \Delta X, \Delta y, \Delta S) = \begin{bmatrix} \Delta\varepsilon \\ -\mathcal{A}^*(\Delta y) - \Delta S \\ \mathcal{A}(\Delta X) \\ \Delta X - V(\Delta\varepsilon, \Delta X - \Delta S) \end{bmatrix} = 0\,,$$

which implies that $\Delta\varepsilon = 0$. Thus, we have

$$W(0, \Delta X, \Delta y, \Delta S) = \begin{bmatrix} 0 \\ -\mathcal{A}^*(\Delta y) - \Delta S \\ \mathcal{A}(\Delta X) \\ \Delta X - V(0, \Delta X - \Delta S) \end{bmatrix} = 0\,. \tag{70}$$

Since the primal constraint nondegenerate condition (40) implies $\mathcal{M}(\overline{X}) = \{(\bar{y}, \overline{S})\}$, we know from Proposition 15 that the strong second order sufficient condition (51) holds at $\overline{X}$ and takes the form

$$-\Upsilon_{\overline{X}}(-\overline{S}, H) > 0 \quad \forall\, 0 \neq H \in \mathrm{app}(\bar{y}, \overline{S})\,, \tag{71}$$

where the set $\mathrm{app}(\bar{y}, \overline{S})$ is defined by (50). From Proposition 5, (70), and (50), we know that

$$\Delta X \in \mathrm{app}(\bar{y}, \overline{S})\,. \tag{72}$$

By the second and the third equations of (70), we obtain that

$$0 = \langle \Delta X, -\mathcal{A}^*(\Delta y) - \Delta S\rangle + \langle \Delta y, \mathcal{A}(\Delta X)\rangle = \langle \Delta X, -\Delta S\rangle\,,$$

which, together with Proposition 20 and the last equation of (70), implies that

$$\Upsilon_{\overline{X}}(-\overline{S}, \Delta X) \geq 0\,.$$

Hence, from (71) and (72), we can conclude that

$$\Delta X = 0\,.$$

Thus, from (70), we get

$$\begin{bmatrix} \mathcal{A}^*(\Delta y) + \Delta S \\ V(0, -\Delta S) \end{bmatrix} = 0\,, \tag{73}$$

which, by Proposition 5, gives rise to

$$P_\alpha^T(\Delta S)P_\alpha = 0, \quad P_\alpha^T(\Delta S)P_\beta = 0, \text{ and } P_\alpha^T(\Delta S)P_\gamma = 0\,. \tag{74}$$

From (39), which is equivalent to the primal constraint nondegenerate condition (40), we know that there exist $X \in \mathcal{S}^n$ and $S \in \operatorname{lin}\left(\mathcal{T}_{\mathcal{S}_+^n}(\overline{X})\right)$ such that

$$\mathcal{A}\,X = \Delta y \quad \text{and} \quad X + S = \Delta S\,,$$

which, together with (74), (49), and the first equation of (73), imply

$$\begin{aligned} \langle \Delta y, \Delta y\rangle + \langle \Delta S, \Delta S\rangle &= \langle \mathcal{A}\,X, \Delta y\rangle + \langle X + S, \Delta S\rangle \\ &= \langle \mathcal{A}\,X, \Delta y\rangle + \langle X, -\mathcal{A}^*(\Delta y)\rangle + \langle S, \Delta S\rangle \\ &= \langle S, \Delta S\rangle = \langle P^T S P, P^T(\Delta S)P\rangle = 0\,. \end{aligned}$$

Thus, $\Delta y = 0$ and $\Delta S = 0$, which, together with $\Delta\varepsilon = 0$ and $\Delta X = 0$, imply the following:

$$W(\Delta\varepsilon, \Delta X, \Delta y, \Delta S) = 0 \Longrightarrow (\Delta\varepsilon, \Delta X, \Delta y, \Delta S) = 0\,.$$

This shows that W is nonsingular. So, the proof is completed. □

The significance of Proposition 21 is that it allows us to offer a quadratic convergence analysis on several globally convergent smoothing Newton methods presented in [9, 10, 20, 46] for solving the SDP problem even when the strict complementarity condition is not satisfied, i.e., when the condition $\overline{X} + \overline{S} \succ 0$ fails to hold. Instead of working on these different smoothing Newton methods one by one (with some necessary modifications), for simplicity we use only the smoothing Newton method presented in [46] as an example of how this objective can be achieved.

For any $(\varepsilon, X, y, S) \in \Re \times \mathcal{S}^n \times \Re^m \times \mathcal{S}^n$, write $Z \equiv (X, y, S)$ and define $f(\varepsilon, Z) := \|E(\varepsilon, Z)\|^2$ and $\theta(\varepsilon, Z) := r \min\{1, f(\varepsilon, Z)\}$. Let $\bar{\varepsilon} \in (0, \infty)$ and $r \in (0, 1)$ be such that $r\bar{\varepsilon} < 1$. The smoothing Newton method presented in [46] can then be stated as follows.

Algorithm I (a squared smoothing Newton method).

Step 0. Select constants $\delta \in (0, 1)$ and $\sigma \in (0, 1/2)$. Let $\varepsilon_0 := \bar{\varepsilon}$, $Z^0 \in \mathcal{S}^n \times \Re^m \times \mathcal{S}^n$ be an arbitrary point, and $k := 0$.

Step 1. If $E(\varepsilon_k, Z^k) = 0$, then stop. Otherwise, let $\theta_k := \theta(\varepsilon_k, Z^k)$.

Step 2. Compute $(\Delta\varepsilon_k, \Delta Z^k)$ by

$$E(\varepsilon_k, Z^k) + E'(\varepsilon_k, Z^k)(\Delta\varepsilon_k, \Delta Z^k) = \theta_k \begin{bmatrix} \bar{\varepsilon} \\ 0 \end{bmatrix}. \tag{75}$$

Step 3. Let l_k be the smallest nonnegative integer l satisfying

$$f(\varepsilon_k + \delta^l \Delta\varepsilon_k, Z^k + \delta^l \Delta Z^k) \le [\,1 - 2\sigma(1 - r\bar{\varepsilon})\delta^l\,] f(\varepsilon_k, Z^k)\,. \tag{76}$$

Define $(\varepsilon_{k+1}, Z^{k+1}) := (\varepsilon_k + \delta^{l_k}\Delta\varepsilon_k, Z^k + \delta^{l_k}\Delta Z^k)$.

Step 4. Replace k by $k+1$ and go to Step 1.

The well posedness of Algorithm I hinges on the nonsingularity of $E'(\varepsilon, Z)$ for any $\varepsilon > 0$, which is equivalent to the surjectivity of the linear operator $\mathcal{A} : \mathcal{S}^n \to \Re^m$ [46]. The two conditions required for quadratic convergence of Algorithm I are (i) the strong semismoothness of the smoothing function E and (ii) the nonsingularity of all $W \in \partial_B E(0, Z^*)$ (or all $W \in \partial E(0, Z^*)$). However, (i) has been proven in [46] and (ii) can be derived from Proposition 21 under both the primal constraint nondegeneracy and the dual constraint nondegeneracy. Thus, by employing the standard convergence analysis detailed in [30] for the vector version of the squared smoothing Newton method, we have the following convergence theorem. For more explanation, see [46].

THEOREM 22. *Assume that $\mathcal{A} : \mathcal{S}^n \to \Re^m$ is onto. Then an infinite sequence $\{(\varepsilon_k, Z^k)\}$ is generated by Algorithm* I *and each accumulation point $(0, \overline{Z})$ of $\{(\varepsilon_k, Z^k)\}$ is a solution of $E(\varepsilon, Z) = 0$. Let $\overline{Z} = (\overline{X}, \bar{y}, \overline{S}) \in \mathcal{S}^n \times \Re^m \times \mathcal{S}^n$. If the primal constraint nondegenerate condition* (40) *holds at $\overline{X}$ and the dual constraint nondegenerate condition* (42) *holds at $(\bar{y}, \overline{S})$, then the whole sequence $\{(\varepsilon_k, Z^k)\}$ converges to $(0, \overline{Z})$,*

$$\| (\varepsilon_{k+1}, Z^{k+1}) - (0, \overline{Z}) \| = O(\| (\varepsilon_k, Z^k) - (0, \overline{Z}) \|^2), \tag{77}$$

and

$$\varepsilon_{k+1} = O(\varepsilon_k^2)\,. \tag{78}$$

Note that in Theorem 22, the quadratic convergence does not rely on the strict complementarity—one common condition that was assumed in all known smoothing Newton methods for solving the SDP problem (1) and its dual, as far as we know. The smoothing function G can certainly take other forms. For example, in order to improve the global convergence of Algorithm I, one may consider Tikhonov-type regularized smoothing functions such as

(79)

$$G(\varepsilon, X, y, S) := \begin{bmatrix} C - \mathcal{A}^* y - S + \varepsilon X \\ \mathcal{A}X - b + \varepsilon y \\ S - \Phi\big(\varepsilon, S - (X + \varepsilon S)\big) \end{bmatrix} = \begin{bmatrix} C - \mathcal{A}^* y - S + \varepsilon X \\ \mathcal{A}X - b + \varepsilon y \\ X - \Phi\big((X + \varepsilon S) - S\big) + \varepsilon S \end{bmatrix}.$$

The quadratic convergence of Algorithm I will not be affected because, by Lemma 1, the set $\partial_B E(0, X, S, Y)$ is still kept the same for any $(X, y, S) \in \mathcal{S}^n \times \Re^m \times \mathcal{S}^n$ if one replaces the smoothing function G in (64) by the one given in (79).

5. Conclusions. In this paper, we presented several equivalent links among the primal and dual constraint nondegenerate conditions, the strong regularity, and the nonsingularity of both the B-subdifferential and Clarke's generalized Jacobian of a nonsmooth system at a KKT point in the context of linear semidefinite programming. These links were further used to derive for the first time a quadratic convergence analysis of globally convergent smoothing Newton methods without assuming the strict complementarity. Variational analysis on the metric projector over the cone

of positive semidefinite matrices and its smoothed counterpart plays a fundamental role in achieving these. Given the fact that the metric projector over the more general symmetric cone behaves quite similarly to the metric projector over the cone of positive semidefinite matrices [45], one is tempted to wonder if the results obtained in this paper can be extended to linear symmetric cone programming. We leave this interesting question as our future research topic.

Acknowledgments. The authors are grateful to the referees and the associate editor for their helpful suggestions on improving the quality of this paper. Defeng Sun would also like to thank Jong-Shi Pang and Alexander Shapiro for discussions on the constraint nondegeneracy when they visited National University of Singapore during 2002–2005.

REFERENCES

[1] F. Alizadeh, *Interior point methods in semidefinite programming with applications to combinatorial optimization*, SIAM J. Optim., 5 (1995), pp. 13–51.
[2] F. Alizadeh, J.-P. A. Haeberly, and M. L. Overton, *Complementarity and nondegeneracy in semidefinite programming*, Math. Programming, 77 (1997), pp. 111–128.
[3] V. I. Arnold, *On matrices depending on parameters*, Russian Math. Surveys, 26 (1971), pp. 29–43.
[4] R. Bhatia, *Matrix Analysis*, Springer-Verlag, New York, 1997.
[5] J. F. Bonnans, R. Cominetti, and A. Shapiro, *Sensitivity analysis of optimization problems under second order regularity constraints*, Math. Oper. Res., 23 (1998), pp. 803–832.
[6] J. F. Bonnans, R. Cominetti, and A. Shapiro, *Second order optimality conditions based on parabolic second order tangent sets*, SIAM J. Optim., 9 (1999), pp. 466–492.
[7] J. F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer-Verlag, New York, 2000.
[8] S. Chandrasekaran and I. C. F. Ipsen, *Backward errors for eigenvalue and singular value decompositions*, Numer. Math., 68 (1994), pp. 215–223.
[9] X. Chen, H. -D. Qi, and P. Tseng, *Analysis of nonsmooth symmetric-matrix-valued functions with applications to semidefinite complementarity problems*, SIAM J. Optim., 13 (2003), pp. 960–985.
[10] X. Chen and P. Tseng, *Non-interior continuation methods for solving semidefinite complementarity problems*, Math. Program., 95 (2003), pp. 431–474.
[11] F. H. Clarke, *On the inverse function theorem*, Pacific J. Math., 64 (1976), pp. 97–102.
[12] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley and Sons, New York, 1983.
[13] B. C. Eaves, *On the basic theorem of complementarity*, Math. Programming, 1 (1971), pp. 68–75.
[14] M. L. Flegel and C. Kanzow, *Equivalence of two nondegeneracy conditions for semidefinite programs*, J. Optim. Theory Appl., 135 (2007), pp. 713–735.
[15] M. S. Gowda, *Inverse and implicit function theorems for H-differentiable and semismooth functions*, Optim. Methods Softw., 19 (2004), pp. 443-461.
[16] N. J. Higham, *Computing a nearest symmetric positive semidefinite matrix*, Linear Algebra Appl., 103 (1988), pp. 103–118.
[17] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.
[18] C. Kanzow and C. Nagel, *Semidefinite programs: New search directions, smoothing-type methods, and numerical results*, SIAM J. Optim., 13 (2002), pp. 1–23.
[19] C. Kanzow and C. Nagel, *Corrigendum: Semidefinite programs: New search directions, smoothing-type methods, and numerical results*, SIAM J. Optim., 14 (2004), pp. 936–937.
[20] C. Kanzow and C. Nagel, *Quadratic convergence of a nonsmooth Newton-type method for semidefinite programs without strict complementarity*, SIAM J. Optim., 15 (2005), pp. 654–672.
[21] B. Kummer, *Lipschitzian inverse functions, directional derivatives, and applications in $C^{1,1}$-optimization*, J. Optim. Theory Appl., 70 (1991), pp. 559–580.
[22] K. Löwner, *Über monotone matrixfunktionen*, Math. Z., 38 (1934), pp. 177–216.

[23] J. Malick and H. S. Sendov, *Clarke generalized Jacobian of the projection onto the cone of positive semidefinite matrices*, Set-Valued Anal., 14 (2006), pp. 273–293.
[24] T. Matsumoto, *An algebraic condition equivalent to strong stability of stationary solutions of nonlinear positive semidefinite programs*, SIAM J. Optim., 16 (2005), pp. 452–470.
[25] F. Meng, D. F. Sun, and G. Zhao, *Semismoothness of solutions to generalized equations and the Moreau–Yosida regularization*, Math. Program., 104 (2005), pp. 561–581.
[26] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optim., 15 (1977), pp. 959–972.
[27] Y. Nesterov and A. Nemirovskii, *Interiorpoint Polynomial Algorithms in Convex Programming*, SIAM Stud. Appl. Math. 13, SIAM, Philadelphia, 1994.
[28] J.-S. Pang, D. F. Sun, and J. Sun, *Semismooth homeomorphisms and strong stability of semidefinite and Lorentz complementarity problems*, Math. Oper. Res., 28 (2003), pp. 39–63.
[29] L. Qi, *Convergence analysis of some algorithms for solving nonsmooth equations*, Math. Oper. Res., 18 (1993), pp. 227–244.
[30] L. Qi, D. F. Sun, and G. Zhou, *A new look at smoothing Newton methods for nonlinear complementarity problems and box constrained variational inequalities*, Math. Program., 87 (2000), pp. 1–35.
[31] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Program., 58 (1993), pp. 353–367.
[32] S. M. Robinson, *Strongly regular generalized equations*, Math. Oper. Res., 5 (1980), pp. 43–62.
[33] S. M. Robinson, *Local structure of feasible sets in nonlinear programming* II: *Nondegeneracy*, Math. Program. Stud., 22 (1984), pp. 217–230.
[34] S. M. Robinson, *Local structure of feasible sets in nonlinear programming* III: *Stability and sensitivity*, Math. Program. Stud., 30 (1987), pp. 45–66.
[35] S. M. Robinson, *Constraint nondegeneracy in variational analysis*, Math. Oper. Res., 28 (2003), pp. 201–232.
[36] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[37] R. T. Rockafellar and R. J. -B. Wets, *Variational Analysis*, Springer-Verlag, New York, 1998.
[38] N. C. Schwertman and D. M. Allen, *Smoothing an indefinite variance-covariance matrix*, J. Statist. Comput. Simulation, 9 (1979), pp. 183–194.
[39] A. Shapiro, *First and second order analysis of nonlinear semidefinite programs*, Math. Program., 77 (1997), pp. 301–320.
[40] A. Shapiro, *Sensitivity analysis of generalized equations*, J. Math. Sci., 115 (2003), pp. 2554–2565.
[41] A. Shapiro and M. K. H. Fan, *On eigenvalue optimization*, SIAM J. Optim., 5 (1995), pp. 552–569.
[42] D. F. Sun, *The strong second-order sufficient condition and constraint nondegeneracy in nonlinear semidefinite programming and their implications*, Math. Oper. Res., 31 (2006), pp. 761–776.
[43] D. F. Sun and J. Sun, *Semismooth matrix valued functions*, Math. Oper. Res., 27 (2002), pp. 150–169.
[44] D. F. Sun and J. Sun, *Strong semismoothness of Fischer-Burmeister SDC and SOC complementarity functions*, Math. Program., 103 (2005), pp. 575–581.
[45] D. F. Sun and J. Sun, *Löwner's operator and spectral functions in Euclidean Jordan algebras*, Math. Oper. Res., 33 (2008), to appear.
[46] J. Sun, D. F. Sun, and L. Qi, *A squared smoothing Newton method for nonsmooth matrix equations and its applications in semidefinite optimization problems*, SIAM J. Optim., 14 (2004), pp. 783–806.
[47] M. J. Todd, *Semidefinite optimization*, Acta Numerica, 10 (2001), pp. 515–560.
[48] P. Tseng, *Merit functions for semi-definite complementarity problems*, Math. Programming, 83 (1998), pp. 159–185.
[49] E. H. Zarantonello, *Projections on convex sets in Hilbert space and spectral theory* I *and* II, in Contributions to Nonlinear Functional Analysis, E. H. Zarantonello, ed., Academic Press, New York, 1971, pp. 237–424.

SIAM J. OPTIM.
Vol. 19, No. 1, pp. 397–413

CONVERGENCE OF NEW INERTIAL PROXIMAL METHODS FOR DC PROGRAMMING*

PAUL-EMILE MAINGÉ† AND ABDELLATIF MOUDAFI†

Abstract. We present iterative methods for finding the critical points and/or the minima of extended real valued functions of the form $\phi = \psi + g - h$, where ψ is a differentiable function and g and h are convex, proper, and lower semicontinuous. The underlying idea relies upon the discretization of a first order dissipative dynamical system which allows us to preserve the local feature and to obtain some convergence results. The main theorems not only recover known convergence results in this field but also provide a theoretical basis for the development of new iterative methods.

Key words. DC minimization, proximal mappings, critical points, subdifferentials

AMS subject classifications. Primary, 49J53, 65K10; Secondary, 49M37, 90C25

DOI. 10.1137/060655183

1. Introduction.

In recent years there has been very active research in nonconvex programming. A great deal of the work involves global optimization, whose main tools and solution methods are developed according to the spirit of combinatorial optimization. Apart from this combinatorial approach to global continuous optimization, the convex analysis approach to nonconvex programming has been much less studied, and only on DC programming. In this paper these works are extended in a natural and logical way to find critical points of a nonconvex objective function $\phi : \mathcal{H} \to \overline{\mathbb{R}}$ of the form $\phi := \psi + g - h$, namely,

$$\mathcal{P}(\phi, \mathcal{H}) : \qquad \min_{x \in \mathcal{H}} (\psi(x) + g(x) - h(x)),$$

where ψ is differentiable (not necessarily convex) and g and h are convex, lower semicontinuous proper functions on a real Hilbert space $\mathcal{H}$ endowed with inner product and induced norm, respectively denoted by $\langle ., . \rangle$ and $|.|$.

Indeed, we wish to make an extension of DC programming—not too large to allow us the use of the arsenal of powerful tools in convex analysis and convex optimization but sufficiently wide to cover most real-world nonconvex optimization problems. The convexity of the two DC components and the differentiability of ψ of the objective function will be used to develop appropriate tools from both theoretical and algorithmic viewpoints. There is a great interest for this class of objective functions (see, for instance, [18, 28]). Let us mention that $\mathcal{P}(\phi, \mathcal{H})$, in the special case when $g = \delta_C$, the indicator function of a nonempty closed convex set C, becomes $\mathcal{P}(\psi - h, C) : \min_{x \in C}(\psi(x) - h(x))$, which includes constrained convex maximization problems. When $\psi \equiv 0$, ϕ is called a DC function. It is worth mentioning that the class of DC functions contains all lower-$\mathcal{C}^2$ functions and constitutes a minimal realistic extension of the class of convex functions. It has been successfully used in many nonconvex applications such as multicommodity network, image restoration processing, and semilinear elliptic problems arising in plasma physics and fluid mechanics

*Received by the editors March 25, 2006; accepted for publication (in revised form) September 20, 2007; published electronically April 16, 2008.
http://www.siam.org/journals/siopt/19-1/65518.html

†GRIMMAG, Département Scientifique Interfacultaire, Université des Antilles-Guyane, Campus de Schoelcher, 97230 Cedex, Martinique (F.W.I.) (Paul-Emile.Mainge@martinique.univ-ag.fr, Abdellatif.Moudafi@martinique.univ-ag.fr).

and seems particularly well suited to model several nonconvex industrial problems (computer vision, fuel mixture, molecular biology, etc.). Some interesting optimality conditions related to $\mathcal{P}(\phi, \mathcal{H})$ in the more general setting of Banach spaces are given by Penot [23] (see also [10, 11, 17] for the case when $\psi \equiv 0$).

(i) [23, Proposition 2.1]) A necessary condition for $\bar{x}$ to be a local minimizer of ϕ is $\partial^F h(\bar{x}) \subset \partial^F(\psi + g)(\bar{x})$, where ∂^F is the Fréchet subdifferential. Because of the regularity of the function ψ and the convexity of the functions g and h, we have

$$\partial^F(\psi + g)(\bar{x}) = \partial^F g(\bar{x}) + \nabla\psi(\bar{x}),\ \partial^F g = \partial g, \quad \text{and } \partial^F h = \partial h,$$

where ∂ denotes the standard Fenchel subdifferential of convex analysis, so that this necessary condition may be rewritten as $\partial h(\bar{x}) \subset \partial g(\bar{x}) + \nabla\psi(\bar{x})$.

(ii) [23, Corollary 2.4]) Conversely, when $\mathcal{H}$ is finite dimensional, $\bar{x}$ is a local strict minimizer of ϕ provided that $\partial h(\bar{x}) \subset \text{int}(\partial g(\bar{x})) + \nabla\psi(\bar{x})$.

Being aware that such optimality conditions are not easily reached from the numerical viewpoint, we will focus on the problem of finding critical points of ϕ, namely,

$$\text{find } \bar{x} \in S := \{x \in \mathcal{H};\quad (\nabla\psi(x) + \partial g(x)) \cap \partial h(x) \neq \emptyset\}. \tag{1.1}$$

In contrast with the combinatorial approach, from which many algorithms have been studied, there have been very few algorithms for solving DC programs from the convex analysis approach (see, for instance, [24, 27]). DC algorithms, based on local optimality conditions and duality in DC programming, have been introduced by Pham Dinh Tao [24] as an extension of the subgradient algorithms to DC programming. Due to its local character it cannot guarantee the globality of computed solutions for general DC programs. It is worth mentioning that x^* is a global solution to a DC problem (i.e., when $\psi \equiv 0$) if and only if

$$\partial_\varepsilon h(x^*) \subset \partial_\varepsilon g(x^*) \quad \forall \varepsilon \geq 0.$$

Unfortunately, as we can foresee, the conditions are rather difficult to use for devising solution methods to DC programs.

Another important feature of the DC structure, which must be taken into account while studying solution algorithms, is the regularization techniques in DC programming. To the best of our knowledge, the corresponding scheme, which has been studied in [27], is the most recent algorithm in the special case of DC functions. Each of its iterates consists in combining an ascent subgradient step on h with a proximal step on g (see [27]). An approximate version of this algorithm using the ϵ-subdifferential was also discussed in Moudafi and Maingé [20]. They proved that if the sequence generated by their method is bounded, then every cluster-point is a critical point of ϕ.

It is worth noting, for instance, that by using suitable DC decompositions of convex functions we can obtain almost standard algorithms for convex and nonconvex programming. The choice of the DC decomposition in the objective function strongly depends on the very specific structure of the problem being considered. In practice, for solving problem (1.1), we try to choose g and h such that the sequences x_n and q_n generated by (1.4) can be easily calculated: i.e., either they are in explicit form or their computations are inexpensive. Let us remember that the convergence of most of the numerical methods in convex minimization are obtained under the condition of the existence of a local minimizer (which then is a global minimizer) for the objective function. In the present setting of nonconvex minimization, the existence of local minima does not imply the existence of global minima, even if the objective function

is bounded from below. Therefore, one may expect not to succeed in proving the boundedness of the iterates of the proposed numerical algorithms, except for some restrictive assumptions.

Let us now come to the original aspect of our approach. In order to solve (1.1), we suggest and analyze a new and promising iterative method obtained by coupling the ϵ-subdifferential of h and the approximate proximal mapping of g. These new schemes generalize the algorithms proposed in [27] and most of the existing methods in convex minimization, e.g., the proximal point algorithm (see [9, 16, 25]), the standard projected gradient algorithm (see [21, 15]), and the inertial proximal algorithm and its approximate variants (see [1, 2, 12, 19]). Furthermore, our proposed algorithm is based upon an implicit discretization of the following first order dissipative dynamical system discussed in [3] (see also [5]):

$$\left[\begin{array}{l} x^{(1)}(t) + \beta\nabla\phi(x(t)) + ax(t) + by(t) = 0, \\ y^{(1)}(t) + ax(t) + by(t) = 0, \\ x(0) = x_0, \quad y(0) = y_0, \end{array}\right. \tag{1.2}$$

where ϕ is assumed to be a convex differentiable function, where the initial data x_0 and y_0 belong to $\mathcal{H}$ and where $\beta > 0$, $b > 0$, and $a + b > 0$.

Let us also specify the motivation for this dynamical system in view of building iterative schemes. Most of the existing numerical methods for minimizing a function, whether in the convex setting or not, are based upon a discretization of some continuous equations with appropriate convergence properties. The algorithm of Sun, Sampaio, and Candido [27], the standard proximal point algorithm, and the gradient method come from the first order steepest descent equation $x^{(1)}(t) + \nabla\phi(x(t)) = 0$. Conversely, the inertial proximal method is inspired by the heavy ball with friction dynamical system $x^{(2)}(t)+\alpha x^{(1)}(t)+\nabla\phi(x(t)) = 0$. The latter system was introduced by Attouch, Goudon, and Redont [6] for overcoming some of the drawbacks of the steepest descent method. It turns out that when ϕ is convex, the two trajectories of the last two equations weakly converge to minimizers of ϕ. Nevertheless, there is a tremendous difference between them. By contrast with the steepest descent method, the heavy ball with friction dynamical system is no longer a descent method. In the latter case, it is not the function $\phi(x(t))$ which decreases along trajectories in general but the energy of the system $E(t) := \frac{1}{2}|x^{(1)}(t)|^2 + \phi(x(t))$. Consequently, the latter benefits from interesting properties for the exploration of local minima of ϕ (see [6] for more details). It also appears that its trajectories may exhibit oscillations which are not desirable. In view of numerical optimization purposes, Alvarez and Pérez [4] have studied the "continuous Newton" method $\nabla^2\phi(x(t))x^{(1)}(t) + \nabla\phi(x(t)) = 0$. See also [8], where an ordinary differential equation model within the framework of a preconditioned gradient method is considered, the objective being the developement of an implicit scheme to approximate the preconditioned search direction at every iteration, without a priori knowledge of the Hessian of the objective function. If one combines this last system with the heavy ball system with friction, the system thus obtained,

$$x^{(2)}(t) + \alpha x^{(1)}(t) + \beta\nabla^2\phi(x(t))x^{(1)}(t) + \nabla\phi(x(t)) = 0, \tag{1.3}$$

inherits most of the advantages of the two preceding systems and corrects both of the above-mentioned drawbacks: the term $\nabla^2\phi(x(t))x^{(1)}(t)$ is a clever geometric damping term, while the acceleration term $x^{(2)}(t)$ makes the Newton dynamical system well-posed, even if $\nabla^2\phi(x(t))$ is degenerate (see Attouch and Redont [7] for a first study of

this question). It is worth mentioning that (1.3) is a second order system both in time and in space. Furthermore, it was proved that (1.3) is equivalent in some sense to the system (1.2) which is first order in time and with no occurrence of the Hessian (see [3] and [5] for more details). This matter opens new interesting perspectives such as considering (1.3) for nonsmooth functions which are only lower semicontinuous or involving constraints, with clear applications to mechanics and PDEs (wave equations, shocks). Moreover, the following theorem related to (1.2) is proved.

THEOREM 1.1 (see [3]). *Let ϕ satisfy the following hypotheses:*

- *ϕ is defined and continuously differentiable on $\mathcal{H}$,*
- *ϕ is bounded from below, and*
- *the gradient $\nabla\phi$ is Lipschitz continuous on the bounded subsets.*

Assume further that $\beta > 0$, $b > 0$, $b + a > 0$ in (1.2). Then the following properties hold.

(i) *For each $(x_0, y_0) \in \mathcal{H} \times \mathcal{H}$, there exists a unique solution (x, y) of (1.2) defined on the whole interval $[0, +\infty)$, which satisfies the initial conditions $x(0) = x_0$, $y(0) = y_0$; (x,y) belongs to $C^1(0, \infty; \mathcal{H}) \times C^2(0, \infty; \mathcal{H})$ and satisfies the initial conditions $x(0) = x_0$ and $y(0) = y_0$.*

(ii) *For every trajectory $(x(t), y(t))$ of (1.2) and for $\lambda \in [\beta(\sqrt{a+b} - \sqrt{b})^2, \beta(\sqrt{a+b} + \sqrt{b})^2]$, the energy $F_\lambda : (x, y) \in \mathcal{H} \times \mathcal{H} \to \lambda\phi(x) + \frac{1}{2}|ax + by|^2$ is a Lyapounov function of (1.2); the energy $F_\lambda(x(t), y(t))$ is decreasing on $[0, +\infty)$ and bounded from below and hence converges to some real value as $t \to +\infty$. Moreover, we have*

- *$x^{(1)}$ and $y^{(1)}$ in $L^2(0, +\infty; \mathcal{H})$,*
- *$\lim_{t\to+\infty} \phi(x(t))$ exists,*
- *$\lim_{t\to+\infty} y^{(1)}(t) = 0$, and*
- *$\nabla\phi(x) \in L^2(0, +\infty; \mathcal{H})$ and $\lim_{t\to+\infty} x^{(1)}(t) + \beta\nabla\phi(x(t)) = 0$.*

(iii) *Assuming, in addition, that x is in $L^\infty(0, +\infty; \mathcal{H})$, we have*

- *$\nabla\phi(x)$, y, $x^{(1)}$ are bounded on $[0, +\infty)$,*
- *$\lim_{t\to+\infty} x^{(1)}(t) = 0$, and*
- *$\lim_{t\to+\infty} \nabla\phi(x(t)) = 0$.*

It then seems natural to investigate implicit and/or nonimplicit discretization of (1.2) for numerical and theoretical optimization purposes, because an accurate discrete version of an equation is supposed to preserve the essential properties of the continuous one. Moreover, this approach permits us to avoid the use of the Hessian of the function and thus, from a numerical point of view, its aproximation and the related computational cost (see, for instance, [8]). Now, in order to introduce and analyze convergence properties of a proximal iteration called DPM (dissipative proximal method) for finding critical points of $\mathcal{P}(\phi, \mathcal{H})$ which is motivated by the above arguments and inspired by the system (1.2), let us recall the following definitions.

Given $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$, a convex, proper, and lower semicontinuous function,

- the ϵ-subdifferential of f at any point x in the domain of f is defined by

$$\partial_\epsilon f(x) := \{p_x \in \mathcal{H}; \quad f(y) - f(x) \geq \langle y - x, p_x\rangle - \epsilon \quad \forall y \in \mathcal{H}\}.$$

Clearly, $\partial_\epsilon f(x)$ is an enlargement of $\partial f(x)$ in the sense that $\partial_0 f(x) = \partial f(x)$ and $0 \leq \epsilon_1 \leq \epsilon_2 \Rightarrow \partial_{\epsilon_1} f(x) \subset \partial_{\epsilon_2} f(x)$. The use of elements in $\partial_\epsilon f(x)$ instead of $\partial f(x)$ allows an extra degree of freedom, which is of very practical interest in various applications.

- The approximate proximal mapping of f (denoted $J_c^{\partial_\epsilon f}$ or $prox_{cf}^\epsilon$) is defined for any $c > 0$ by $J_c^{\partial_\epsilon f} := (I + c\partial_\epsilon f)^{-1}$.

- The conjugate function of f is defined by $f^*(y) = \sup_{x\in\mathcal{H}} (\langle y, x\rangle - f(y))$. If $k : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is also a convex, proper, and lower semicontinuous function, by Toland's duality, we have $\inf_{x\in\mathcal{H}}(f(x) - k(x)) = \inf_{x\in\mathcal{H}}(k^*(x) - f^*(x))$.

Finally, we will use the usual notation

$$\Gamma_0(\mathcal{H}) := \{v : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}; \quad v \text{ is convex, proper, and lower semicontinuous}\}$$

and adopt in what follows the convention $+\infty - (+\infty) = +\infty$.

Now, choose parameters λ, μ, (ϵ_n), α, γ, ν, τ such that

- $\lambda > 0$, $\mu > 0$, $(\epsilon_n) \subset [0, +\infty)$;
- $\alpha + \gamma > 0$ and $\gamma > 0$; and
- $\nu > 0$ and $\tau > -(2 + \alpha)/(2\gamma)$.

Consider then Algorithm DPM (for finding critical points of $\psi(x) + g(x) - h(x)$):

$$\left[\begin{array}{l} \text{Initialization: } (x_0, y_0) \text{ in } \mathcal{H} \times \mathcal{H}. \\ \\ \text{Step 1: Compute} \quad q_n \in \partial_{\epsilon_n} h(x_n). \\ \\ \text{Step 2: Compute} \\ \quad x_{n+1} \in J_\lambda^{\partial_{\epsilon_n} g}[x_n - \lambda\,(\nabla\psi(x_n) - q_n) - \mu(\alpha x_n + \gamma y_n)]. \\ \\ \text{Step 3: Compute} \\ \quad y_{n+1} = y_n - \dfrac{1}{k}[\alpha x_n + \gamma y_n + \nu\alpha(x_{n+1} - x_n)], \\ \quad \text{where } k := \left(1 + \tau\gamma + \frac{(\alpha+\gamma)}{2}\right). \end{array}\right. \tag{1.4}$$

Let us observe that under the conditions $\gamma > 0$ and $\tau > -(2 + \alpha)/(2\gamma)$, we have $k \geq \gamma/2 > 0$, so that Step 3 in (1.4) makes sense and hence (1.4) is well defined.

The goal of this paper is to provide a broad framework for the design and the analysis of promising algorithms based on a discretization of the new system (1.2). This framework will not only lead to a unified convergence analysis of some existing algorithms in convex optimization and DC programming but will also propose original and theoretically interesting results and will serve as a basis for the development of a new class of algorithms. More specifically, we would like to emphasize that the method proposed by [27] is nothing but the special case of DPM when $\alpha = \gamma = 0$ and g, h are convex functions. The classical proximal algorithm is recovered by taking $\psi = h \equiv 0$ and $\mu = 0$, and the inertial proximal method initiated in [1] for convex minimization (see also [2]) corresponds to the particular case of DPM when $\psi = h \equiv 0$, $\gamma > 0$, $\alpha > 0$, and the parameters τ and ν satisfy

$$\tau = -\tfrac{1}{2\gamma}(\alpha + 2) + \tfrac{1}{2} \text{ and } \nu = 1 + \tfrac{1}{\alpha}. \tag{1.5}$$

Indeed, it is easy to verify that in this case $k = \gamma$ and $\alpha(1 - \nu) = -1$; then Step 2 becomes $y_{n+1} = -\frac{\alpha}{\gamma}\,(x_n + \nu(x_{n+1} - x_n))$, which (in Step 1) entails

$$\alpha x_n + \gamma y_n = \alpha x_n - \alpha\,(x_{n-1} + \nu(x_n - x_{n-1})) = \alpha(1-\nu)(x_n - x_{n-1}) = -(x_n - x_{n-1}),$$

so that DPM reduces to

$$x_{n+1} \in J_\lambda^{\partial_{\epsilon_n} g}\,(x_n + \mu\,(x_n - x_{n-1}))\,.$$

Under suitable conditions on the parameters and a summability condition on the errors, we prove a discrete version of the main theorem, Theorem 1.1, in a more general context. More precisely, we prove that the proposed algorithm generates an asymptotically regular sequence (x_n) and that its weak-cluster points are in S. Moreover, in the convex case we obtain the weak convergence of the whole sequence to a minimizer of the function under consideration.

2. Preliminaries. To begin, let us state a remark which will be useful in what follows.

Remark 2.1. Step 3 in (1.4) is equivalent to each of the following equalities:

$$\begin{aligned}
&\text{(i)} - k(y_{n+1}-y_n) = \alpha x_n + \gamma y_n + \nu\alpha(x_{n+1}-x_n),\\
&\quad \text{where} \quad k := \left(1+\tau\gamma+\tfrac{(\alpha+\gamma)}{2}\right);\\
&\text{(ii)} \quad y_n - y_{n+1} = a[\nu x_{n+1} + (1-\nu)x_n] + b[\tau y_{n+1} + (1-\tau)y_n],\\
&\quad \text{where } a = \tfrac{2\alpha}{2+\alpha+\gamma} \text{ and } b = \tfrac{2\gamma}{2+\alpha+\gamma}.
\end{aligned} \tag{2.1}$$

The fact that Step 3 in (1.4) is equivalent to equality (2.1)(i) is obvious. To complete the verification, we will see that (2.1)(ii) is in turn equivalent to (2.1)(i). Indeed, by an elementary computation (2.1)(ii) is equivalent to

$$\frac{2+\alpha+\gamma}{2}(y_n - y_{n+1}) = \alpha\nu(x_{n+1}-x_n) + \alpha x_n + \gamma\tau(y_{n+1}-y_n) + \gamma y_n;$$

in other words,

$$-\left(\frac{2+\alpha+\gamma}{2}+\gamma\tau\right)(y_n - y_{n+1}) = \alpha\nu(x_{n+1}-x_n) + \alpha x_n + \gamma y_n,$$

which is (2.1)(i).

Now, let us clarify the link between (1.2) and (1.4).

Remark 2.2. To see the connection between (1.2) and (1.4), one may observe that Step 2 in (1.4) is equivalent to the monotone inclusion

$$x_{n+1} - x_n + \lambda\left(\partial_{\epsilon_n} g(x_{n+1}) + \nabla\psi(x_n) - q_n\right) + \mu[\alpha x_n + \gamma y_n] \ni 0, \tag{2.2}$$

while by (2.1)(i) we have

$$\alpha x_n + \gamma y_n = -k(y_{n+1}-y_n) - \nu\alpha(x_{n+1}-x_n),$$

where $k := (1+\tau\gamma+\frac{(\alpha+\gamma)}{2})$, so that we deduce

$$(1-\mu\nu\alpha)(x_{n+1}-x_n) + \lambda(\partial_{\epsilon_n} g(x_{n+1}) + \nabla\psi(x_n) - q_n) - \mu k(y_{n+1}-y_n) \ni 0,$$

namely,

$$\frac{(1-\mu\nu\alpha)}{\mu k}(x_{n+1}-x_n) + \frac{\lambda}{\mu k}[\partial_{\epsilon_n} g(x_{n+1}) + \nabla\psi(x_n) - q_n] - (y_{n+1}-y_n) \ni 0. \tag{2.3}$$

Setting $\beta := \lambda/(\mu k\Delta t)$ (for some positive Δt), we obtain

$$\frac{(1-\mu\nu\alpha)}{\mu k}(x_{n+1}-x_n) + \Delta t\beta[\partial_{\epsilon_n} g(x_{n+1}) + \nabla\psi(x_n) - q_n] - (y_{n+1}-y_n) \ni 0.$$

Hence, for $\mu = \frac{1}{k+\nu\alpha}$, one has $1 - \mu\nu\alpha = \mu k$, and we get

$$\frac{1}{\Delta t}(x_{n+1} - x_n) + \beta[\partial_{\epsilon_n} g(x_{n+1}) + \nabla\psi(x_n) - q_n] - \frac{1}{\Delta t}(y_{n+1} - y_n) \ni 0, \tag{2.4}$$

which in light of (2.1)(ii) can be viewed as a discrete variant of (1.2).

The following lemma will be needed in the next sections.

LEMMA 2.1. *Let w be a differentiable function on $\mathcal{H}$ with an L-Lipschitz continuous gradient for some $L \in]0, +\infty[$ and $u, v \in \Gamma_0(\mathcal{H})$. Then for any points s and t in $\mathcal{H}$, we have*

$$\forall p \in \partial_\epsilon u(t), \quad \forall q \in \partial_\epsilon v(s),$$
$$(w + u - v)(t) - (w + u - v)(s) \le \langle \nabla w(s) + p - q, t - s\rangle + L|t - s|^2 + 2\epsilon.$$

Proof. Taking $u \in \Gamma_0(\mathcal{H})$ and $p \in \partial_\epsilon u(s)$, we have $u(t) - u(s) \le \langle p, t - s\rangle + \epsilon$. Moreover, taking $v \in \Gamma_0(\mathcal{H})$ and $q \in \partial_\epsilon v(s)$, we also have $v(s) - v(t) \le \langle q, s - t\rangle + \epsilon$. Then, by combining the two previous inequalities, we get

$$(u - v)(t) - (u - v)(s) \le \langle p - q, t - s\rangle + 2\epsilon. \tag{2.5}$$

Now, by a classical result there exists $c_\theta := \theta s + (1 - \theta)t$ for some $\theta \in]0, 1[$ such that $w(t) = w(s) + \langle \nabla w(c_\theta), t - s\rangle$, that is, $w(t) - w(s) = \langle \nabla w(c_\theta) - \nabla w(s), t - s\rangle + \langle \nabla w(s), t - s\rangle$. Since ∇w is L-Lipschitz continuous on $\mathcal{H}$, we then obtain

$$w(t) - w(s) \le L|s - t|^2 + \langle \nabla w(s), t - s\rangle,$$

which in light of (2.5) yields the desired result. □

In what follows, $E_n(\delta)$, the discrete energy of the method (1.4), is defined for $n \ge 1$ and $\delta > 0$ by

$$E_n(\delta) = \delta\phi(x_n) + \frac{1}{2}|ax_n + by_n|^2 - 2\delta \sum_{j=0}^{n-1} \epsilon_j, \tag{2.6}$$

where $a = \frac{2\alpha}{2+\alpha+\gamma}$ and $b = \frac{2\gamma}{2+\alpha+\gamma}$.

It is worth mentioning that (2.6) has the same form as the one used in [3] and will permit us to make a connection with the framework developed by Attouch, Bolte, and Redont [5]. Indeed, applying the continuous steepest descent method to the functional

$$E(\delta) : (x, y) \to \frac{1}{b^2}\phi(x) + \frac{1}{2}|ax + by|^2 \tag{2.7}$$

(i.e., $E_n(\delta)$ with $\delta = \frac{1}{b^2}, \epsilon_j = 0$ for $j = 1 \cdots n - 1$) provides the following first order system which is close to (1.2):

$$\left[\begin{array}{l} x^{(1)}(t) + \frac{1}{b^2}\partial\phi(x(t)) + a(ax(t) + by(t)) \ni 0, \\ y^{(1)}(t) + b(ax(t) + by(t)) = 0. \end{array}\right.$$

The latter was introduced by Attouch, Bolte, and Redont [5] in view of minimizing ϕ, which in light of [5, Proposition 6.1.1], is equivalent to minimizing the functional $E(\frac{1}{b^2})$.

3. Asymptotic convergence of DPM. The following lemma contains a useful property of the discrete energy.

LEMMA 3.1. *Let ψ be a differentiable function on $\mathcal{H}$ with an L-Lipschitz continuous gradient for some $L \in]0,+\infty[$ and $g, h \in \Gamma_0(\mathcal{H})$. Then, the two sequences (x_n) and (y_n) generated by scheme (1.4) satisfy the following inequality: $\forall \delta > 0, \forall n \geq 1$,*

$$\begin{aligned} E_{n+1}(\delta) - E_n(\delta) \leq & \left(-\delta \tfrac{1-\mu\nu\alpha}{\lambda} + \delta L - aC_1\right)|x_{n+1}-x_n|^2 - bC_2|y_{n+1}-y_n|^2 \\ & + \left(\frac{\delta\mu k}{\lambda} - (aC_2+bC_1)\right)\langle x_{n+1}-x_n, y_{n+1}-y_n\rangle, \end{aligned} \tag{3.1}$$

where $k := (1+\tau\gamma+\frac{(\alpha+\gamma)}{2})$, $C_2 := [1+b(\tau-1/2)]$, and $C_1 := a(\nu-1/2)$. $(E_j(\delta))_{j\geq 0}$ is the discrete energy defined in (2.6), and a, b stand for the two parameters defined in (2.1)(ii).

Proof. From (2.3), there exists $p_{n+1} \in \partial_{\epsilon_n} g(x_{n+1})$ such that

$$\frac{1-\mu\nu\alpha}{\mu k}|x_{n+1}-x_n|^2 + \frac{\lambda}{\mu k}\langle p_{n+1}+\nabla\psi(x_n)-q_n, x_{n+1}-x_n\rangle + \langle y_n - y_{n+1}, x_{n+1}-x_n\rangle = 0,$$

that is,

$$\begin{aligned} & \langle p_{n+1}+\nabla\psi(x_n)-q_n, x_{n+1}-x_n\rangle \\ & \qquad = \frac{\mu k}{\lambda}\left(-\frac{1-\mu\nu\alpha}{\mu k}|x_{n+1}-x_n|^2 + \langle y_{n+1}-y_n, x_{n+1}-x_n\rangle\right). \end{aligned} \tag{3.2}$$

Thanks to definition (2.6) of the energy function, we have

$$\begin{aligned} E_{n+1}(\delta) - E_n(\delta) = & \delta(\phi(x_{n+1})-\phi(x_n)) - 2\delta\epsilon_n \\ & + \left\langle a(x_{n+1}-x_n)+b(y_{n+1}-y_n), a\frac{x_{n+1}+x_n}{2} + b\frac{y_{n+1}+y_n}{2}\right\rangle. \end{aligned} \tag{3.3}$$

By virtue of (2.1)(ii), we also have

$$\begin{aligned} & y_n - y_{n+1} - a\frac{x_{n+1}+x_n}{2} - b\frac{y_{n+1}+y_n}{2} \\ & \qquad = a[\nu x_{n+1} + (1-\nu)x_n] - a\frac{x_{n+1}+x_n}{2} \\ & \qquad\quad + b[\tau y_{n+1} + (1-\tau)y_n] - b\frac{y_{n+1}+y_n}{2} \\ & \qquad = a\left(\nu-\frac{1}{2}\right)(x_{n+1}-x_n) + b\left(\tau-\frac{1}{2}\right)(y_{n+1}-y_n), \end{aligned}$$

that is,

$$\begin{aligned} & a\frac{x_{n+1}+x_n}{2} + b\frac{y_{n+1}+y_n}{2} \\ & \qquad = -[1+b(\tau-1/2)](y_{n+1}-y_n) - a(\nu-1/2)(x_{n+1}-x_n) \\ & \qquad = -C_2(y_{n+1}-y_n) - C_1(x_{n+1}-x_n), \end{aligned}$$

where $C_2 := [1 + b\,(\tau - 1/2)]$ and $C_1 := a\,(\nu - 1/2)$, which by (3.3) entails

$$\begin{aligned} E_{n+1}(\delta) - E_n(\delta) - \delta(\phi(x_{n+1}) - \phi(x_n)) + 2\delta\epsilon_n \\ &= \left\langle a(x_{n+1} - x_n) + b(y_{n+1} - y_n), a\frac{x_{n+1} + x_n}{2} + b\frac{y_{n+1} + y_n}{2} \right\rangle \\ &= -C_2\langle a(x_{n+1} - x_n) + b(y_{n+1} - y_n), y_{n+1} - y_n\rangle \\ &\quad - C_1\langle a(x_{n+1} - x_n) + b(y_{n+1} - y_n), x_{n+1} - x_n\rangle \\ &= -bC_2|y_{n+1} - y_n|^2 - aC_1|x_{n+1} - x_n|^2 \\ &\quad - (aC_2 + bC_1)\langle x_{n+1} - x_n, y_{n+1} - y_n\rangle. \end{aligned} \tag{3.4}$$

As a result we equivalently get

$$\begin{aligned} E_{n+1}(\delta) - E_n(\delta) = \delta(\phi(x_{n+1}) - \phi(x_n)) - 2\delta\epsilon_n \\ - (aC_2 + bC_1)\langle x_{n+1} - x_n, y_{n+1} - y_n\rangle \\ - bC_2|y_{n+1} - y_n|^2 - aC_1|x_{n+1} - x_n|^2. \end{aligned} \tag{3.5}$$

As $(q_n, p_{n+1}) \in \partial_{\epsilon_n} h(x_n) \times \partial_{\epsilon_n} g(x_{n+1})$ and remembering that $\phi = \psi + g - h$, from Lemma 2.1 we obtain

$$\begin{aligned} \phi(x_{n+1}) - \phi(x_n) \\ \le \langle p_{n+1} + \nabla\psi(x_n) - q_n, x_{n+1} - x_n\rangle + L|x_{n+1} - x_n|^2 + 2\epsilon_n. \end{aligned} \tag{3.6}$$

Combining (3.5) and (3.6), we consequently obtain

$$\begin{aligned} E_{n+1}(\delta) - E_n(\delta) \le \delta\langle p_{n+1} + \nabla\psi(x_n) - q_n, x_{n+1} - x_n\rangle \\ - (aC_2 + bC_1)\langle x_{n+1} - x_n, y_{n+1} - y_n\rangle \\ - bC_2|y_{n+1} - y_n|^2 + (\delta L - aC_1)|x_{n+1} - x_n|^2, \end{aligned} \tag{3.7}$$

which along with (3.2) yields

$$\begin{aligned} E_{n+1}(\delta) - E_n(\delta) \le \delta\frac{\mu k}{\lambda}\left(-\frac{1 - \mu\nu\alpha}{\mu k}|x_{n+1} - x_n|^2 + \langle y_{n+1} - y_n, x_{n+1} - x_n\rangle\right) \\ - (aC_2 + bC_1)\langle x_{n+1} - x_n, y_{n+1} - y_n\rangle \\ - bC_2|y_{n+1} - y_n|^2 + (\delta L - aC_1)|x_{n+1} - x_n|^2, \end{aligned} \tag{3.8}$$

which is the desired result. □

Now, let us establish the main result of this section.

THEOREM 3.2. *Assume that the following conditions are satisfied:*

$$\gamma > 0, \quad \alpha + \gamma > 0, \quad \tau > -\frac{2 + \alpha}{2\gamma}, \quad \nu \ge 1/2, \tag{3.9}$$

$$\sum_{n\ge 0} \epsilon_n < \infty. \tag{3.10}$$

Suppose also that $f, g \in \Gamma_0(\mathcal{H})$ and ψ is a differentiable function on $\mathcal{H}$ with an L-Lipschitz continuous gradient for some $L \in]0, +\infty[$, and set $\phi := \psi + g - h$ and $k := (1 + \tau\gamma + \frac{(\alpha+\gamma)}{2})$. If in addition ϕ is bounded from below and $\mu, \lambda > 0$ verify

$$\lambda L + \mu(\nu\alpha + k) \le 1, \tag{3.11}$$

then the sequences (x_n) *and* (y_n) *generated by scheme* (1.4) *satisfy the following properties:*

(i) *For* $\delta \in [\beta(\sqrt{b_2} - \sqrt{a_2+b_2})^2, \beta(\sqrt{b_2} + \sqrt{a_2+b_2})^2]$, *where* $\beta = \lambda/(k\mu)$, $a_2 := a[1+2b(\nu-1/2)]$, $b_2 := b[1+b(\tau-1/2)]$ ($a := \frac{2\alpha}{2+\alpha+\gamma}$ *and* $b := \frac{2\gamma}{2+\alpha+\gamma}$), *the energy* $(E_n(\delta))$ *is a decreasing and converging sequence.*

(ii) $\lim_{n\to+\infty} \phi(x_n)$ *exists.*

(iii) $\lim_{n\to+\infty} |x_{n+1} - x_n| = \lim_{n\to+\infty} |y_{n+1} - y_n| = 0$.

(iv) $\lim_{n\to+\infty} |\alpha x_n + \gamma y_n| = 0$.

(v) *There exists a sequence* $p_{n+1} \in \partial_{\epsilon_n} g(x_{n+1})$ *such that*

$$\lim_{n\to+\infty} |p_{n+1} + \nabla\psi(x_n) - q_n| = 0.$$

(vi) *If* $\mathcal{H}$ *is a finite dimensional space and if the sequences* (x_n) *and* (q_n) *are bounded, then every cluster-point* x_∞ *of the sequences* (x_n) *is a critical point of the function* $\psi + g - h$.

Proof. To begin, regarding the parameters involving in this theorem, we wish to emphasize that the conditions given in (3.9) ensure that

$$a > 0, \quad a+b > 0, \quad k > 0, \quad \beta > 0.$$

Furthermore, setting $\beta = \lambda/(k\mu)$ (hence $\lambda = k\mu\beta$), we have

$$-\frac{1-\mu\nu\alpha}{\lambda} + L = -\frac{1}{\beta}\left(\frac{1}{k\mu}(1-L\lambda) - \frac{\nu\alpha}{k}\right),$$

which by (3.1) yields

$$\begin{aligned} E_{n+1}(\delta) - E_n(\delta) \le{}& \left(-\frac{\delta}{\beta}\left(\frac{1}{k\mu}(1-L\lambda) - \frac{\nu\alpha}{k}\right) - aC_1\right)|x_{n+1}-x_n|^2 \\ &- bC_2|y_{n+1}-y_n|^2 \\ &+ \left(\frac{\delta}{\beta} - (aC_2+bC_1)\right)\langle x_{n+1}-x_n, y_{n+1}-y_n\rangle, \end{aligned} \tag{3.12}$$

where $C_2 := [1 + b\,(\tau - 1/2)]$ and $C_1 := a\,(\nu - 1/2)$. Clearly, we have

$$aC_1 = a^2\,(\nu - 1/2) \ge 0 \quad \text{for} \quad \nu \ge 1/2,$$

while it is immediate that $\frac{1}{k\mu}(1-L\lambda) - \frac{\nu\alpha}{k} \ge 1$ when (3.11) is satisfied.

As a consequence, by (3.12) we have

$$\begin{aligned} E_{n+1}(\delta) - E_n(\delta) \le{}& -\frac{\delta}{\beta}|x_{n+1}-x_n|^2 - bC_2|y_{n+1}-y_n|^2 \\ &+ \left(\frac{\delta}{\beta} - (aC_2+bC_1)\right)\langle x_{n+1}-x_n, y_{n+1}-y_n\rangle. \end{aligned} \tag{3.13}$$

Let us denote $a_2 := (aC_2 + bC_1)$ and $b_2 := bC_2$, so that the previous inequality becomes

$$E_{n+1}(\delta) - E_n(\delta) \le q_n(\delta), \tag{3.14}$$

where $q_n(\delta)$ is defined by

$$q_n(\delta) := -\frac{\delta}{\beta}|x_{n+1}-x_n|^2 - b_2|y_{n+1}-y_n|^2 + \left(\frac{\delta}{\beta}-a_2\right)\langle x_{n+1}-x_n, y_{n+1}-y_n\rangle. \tag{3.15}$$

Observe that $q_n(\delta)$ is an affine function with respect to δ, since the sequences (x_n) and (y_n) are independent of δ.

Moreover, for $\tau > 1/2 - 1/b$ (hence $1 + b\,(\tau - 1/2) > 0$), we obviously have

$$b_2 = bC_2 = b[1 + b\,(\tau - 1/2)] > 0 \quad (\text{since } b > 0)$$

and

$$a_2 + b_2 = (a+b)C_2 + bC_1 > 0, \quad \text{because } a+b>0,\ C_2>0,\ b>0, \text{ and } C_1 \geq 0.$$

It is also easily checked that for

$$\delta_1 := \beta(\sqrt{b_2} - \sqrt{a_2+b_2})^2, \quad \delta_2 := \beta(\sqrt{b_2} + \sqrt{a_2+b_2})^2, \tag{3.16}$$

we obtain

$$\begin{aligned} q_n(\delta_1) &= -|(\sqrt{b_2} - \sqrt{a_2+b_2})(x_{n+1}-x_n) + \sqrt{b_2}(y_{n+1}-y_n)|^2, \\ q_n(\delta_2) &= -|(\sqrt{b_2} + \sqrt{a_2+b_2})(x_{n+1}-x_n) - \sqrt{b_2}(y_{n+1}-y_n)|^2. \end{aligned} \tag{3.17}$$

Therefore, for $\delta \in [\delta_1, \delta_2]$, we deduce that

$$q_n(\delta) \leq \max\{q_n(\delta_1), \quad q_n(\delta_2)\} \leq 0, \tag{3.18}$$

which by (3.14) yields

$$E_{n+1}(\delta) - E_n(\delta) \leq 0,$$

so that $(E_n(\delta))$ is a decreasing sequence. As a consequence, $(E_n(\delta))$ is a converging and bounded sequence, provided that $\sum_{j\geq 0}\epsilon_j < \infty$ and ϕ is bounded from below. On the other hand, taking into account the identity

$$\begin{aligned} \langle x_{n+1}-x_n, y_{n+1}-y_n\rangle = &-\frac{1}{2}|(x_{n+1}-x_n)-(y_{n+1}-y_n)|^2 \\ &+\frac{1}{2}|x_{n+1}-x_n|^2 + \frac{1}{2}|y_{n+1}-y_n|^2, \end{aligned} \tag{3.19}$$

inequality (3.14) can be rewritten as

$$\begin{aligned} &E_{n+1}(\delta) - E_n(\delta) \\ &+\frac{1}{2}\left(a_2 + \frac{\delta}{\beta}\right)|x_{n+1}-x_n|^2 + \left(b_2 + \frac{a_2}{2} - \frac{\delta}{2\beta}\right)|y_{n+1}-y_n|^2 \\ &+\frac{1}{2}\left(\frac{\delta}{\beta} - a_2\right)|(x_{n+1}-x_n)-(y_{n+1}-y_n)|^2 \leq 0. \end{aligned}$$

In the particular case when $\delta = \delta_3 := \beta(a_2 + 2b_2)$, this inequality becomes

$$E_{n+1}(\delta_3) - E_n(\delta_3) + (a_2+b_2)\,|x_{n+1}-x_n|^2 + b_2|(x_{n+1}-x_n)-(y_{n+1}-y_n)|^2 \leq 0;$$

hence

$$E_{n+1}(\delta_3) - E_0(\delta_3) + (a_2+b_2)\sum_{j=1}^{n}|x_{j+1}-x_j|^2 + b_2\sum_{j=1}^{n}|(x_{j+1}-x_j)-(y_{j+1}-y_j)|^2 \le 0. \tag{3.20}$$

Consequently, by the boundedness of $E_n(\delta_3)$ and remembering that $b_2>0$, $a_2+b_2>0$, we deduce that $\sum_{j\ge1}|(x_{j+1}-x_j)-(y_{j+1}-y_j)|^2<\infty$ and $\sum_{j\ge1}|x_{j+1}-x_j|^2<\infty$, which leads to

$$\lim_{n\to+\infty}|x_{n+1}-x_n| = \lim_{n\to+\infty}|y_{n+1}-y_n| = 0. \tag{3.21}$$

This combined with (2.1)(i) yields $\lim_{n\to+\infty}(ax_n+by_n)=0$, and (2.3) ensures then the existence of $p_{n+1}\in\partial_{\epsilon_n}g(x_{n+1})$ satisfying $\lim_{n\to+\infty}|p_{n+1}+\nabla\psi(x_n)-q_n|=0$.

Taking μ_1, μ_2 in $[\beta(\sqrt{b_2}-\sqrt{a_2+b_2})^2, \beta(\sqrt{b_2}+\sqrt{a_2+b_2})^2]$ such that $\mu_1\neq\mu_2$, we have

$$\phi(x_n) = \frac{1}{\mu_1-\mu_2}(E_n(\mu_1)-E_n(\mu_2)),$$

so that $\lim_{n\to+\infty}\phi(x_n)$ exists, because both $(E_n(\mu_1))$ and $(E_n(\mu_2))$ are converging sequences.

Now let us consider two subsequences (x_{n_ν}) and (q_{n_ν}) of (x_n) and (q_n) (we will use the same notation for the index even if this requires extracting other subsequences) converging, respectively, to x_∞ and q_∞. By passing to the limit in (2.5) and in the relation $q_{k_\nu}\in\partial_{\varepsilon_{k_\nu}}h(x_{k_\nu})$ and taking into account the fact that the multivalued maps $\partial_{(\cdot)}f(\cdot)$ and $\partial_{(\cdot)}h(\cdot)$ are closed on $\mathbb{R}_+\times\mathbb{R}^n$ and that $\nabla\psi$ is Lipschitz continuous, we obtain

$$q_\infty\in\partial g(x_\infty)+\nabla\psi(x_\infty) \quad\text{and}\quad q_\infty\in\partial h(x_\infty),$$

from which we infer that $(\partial g(x_\infty)+\nabla\psi(x_\infty))\cap\partial h(x_\infty)\neq\emptyset$; in other words, x_∞ is a critical point of $\psi+g-h$.

Remark 3.1. By virtue of [5, Proposition 6.1.1], minimizing ϕ is equivalent to minimizing the functional $E(\frac{1}{b^2})$ defined by (2.7); thus our result could also be considered as a discrete version of Theorem 6.1.1. by Attouch, Bolte, and Redont [5].

4. Convex minimization case. In this section, we will focus on the case where the objective function is convex. Our interest is in solving the convex minimization problem $\min_{x\in\mathcal{H}}g(x)$ or, equivalently,

$$\text{find } \bar{x}\in\mathcal{H} \text{ such that } 0\in\partial g(\bar{x}). \tag{4.1}$$

In this context, (1.4) is nothing else than the following algorithm:

$$\left[\begin{array}{l}\text{Initialization: } (x_0,y_0) \text{ in } \mathcal{H}\times\mathcal{H}.\\[2ex] \text{Step 1: Compute}\\ \quad x_{n+1}\in J_\lambda^{\partial_{\epsilon_n}g}[x_n-\mu(\alpha x_n+\gamma y_n)].\\[2ex] \text{Step 2: Compute}\\ \quad y_{n+1}=y_n-\dfrac{1}{k}[\alpha x_n+\gamma y_n+\nu\alpha(x_{n+1}-x_n)],\\ \quad \text{where } k:=\left(1+\tau\gamma+\frac{(\alpha+\gamma)}{2}\right).\end{array}\right. \tag{4.2}$$

We are going to prove, under adequate conditions, the weak convergence of the sequence (x_n) to a minimizer of the function g. This result is the discrete version of the main result in [3]. However, the proof is much more technical than in the continuous case and will be obtained by verifying that conditions of Opial's lemma are fulfilled.

LEMMA 4.1. *Let $\mathcal{H}$ be a Hilbert space and (x_n) a sequence in $\mathcal{H}$ such that there exists a nonempty set $C \subset \mathcal{H}$ satisfying the following:*

(i) *For every $\tilde{x} \in C$, $\lim_n |x_n - \tilde{x}|$ exists.*

(ii) *Any weak-cluster point of the sequence (x_n) belongs in C.*

Then, there exists $\bar{x} \in C$ such that (x_n) weakly converges to $\bar{x}$.

Now, we are in a position to present the main convergence result of this section.

THEOREM 4.2. *Assume that the following conditions are satisfied:*

$$\gamma > 0,\ \alpha + \gamma > 0, \tag{4.3}$$

$$\tau > -\tfrac{2+\alpha}{2\gamma},\ \nu \geq 1/2, \tag{4.4}$$

$$\textstyle\sum_{n\geq 0} \epsilon_n < \infty. \tag{4.5}$$

Suppose also that a function $g \in \Gamma_0(\mathcal{H})$ is bounded from below and such that Argmin *g, the set of minimizers of g on $\mathcal{H}$, is nonempty. Then for any $\lambda > 0$ and $\mu > 0$ verifying*

$$\begin{aligned} &\mu(\nu\alpha + k) \leq 1, \\ &\text{where } k := \Big(1 + \tau\gamma + \tfrac{(\alpha+\gamma)}{2}\Big), \end{aligned} \tag{4.6}$$

the sequences (x_n) and (y_n) generated by scheme (4.2) satisfy the following properties:

(i) $\sum |x_{n+1} - x_n|^2 < \infty$, $\sum |y_{n+1} - y_n|^2 < \infty$,

(ii) $\lim_{n\to+\infty} |\alpha x_n + \gamma y_n| = 0$, *and*

(iii) *(x_n) weakly converges to a minimizer of the function g on $\mathcal{H}$.*

Proof. Let $\tilde{x}$ be a minimizer of g, set $c_\mu := \frac{(1-\mu\nu\alpha)}{\mu k}$, and introduce the energy-like sequence U_n defined by

$$U_n = \langle \tilde{x} - x_n, ax_n + by_n \rangle + \frac{1}{2}(a + bc_\mu)|\tilde{x} - x_n|^2 + b\beta g(x_n), \tag{4.7}$$

where $\beta = \lambda/(k\mu)$, $a := \frac{2\alpha}{2+\alpha+\gamma}$, and $b := \frac{2\gamma}{2+\alpha+\gamma}$.

It should be noticed that (4.7) can be viewed as a discrete version of a slight modification of the Lyapunov functional used in [3]. Moreover, under the assumptions of the present theorem, Theorem 3.2 clearly holds true. Furthermore, it is easily observed that

$$\begin{aligned} U_{n+1} - U_n - \beta b(g(x_{n+1}) - g(x_n)) &= \langle \tilde{x} - x_{n+1}, ax_{n+1} + by_{n+1} \rangle - \langle \tilde{x} - x_n, ax_n + by_n \rangle \\ &\quad + \frac{1}{2}(a + bc_\mu)\left(|\tilde{x} - x_{n+1}|^2 - |\tilde{x} - x_n|^2\right) \\ &= \langle \tilde{x} - x_{n+1}, a(x_{n+1} - x_n) + b(y_{n+1} - y_n) \rangle + \langle x_n - x_{n+1}, ax_n + by_n \rangle \\ &\quad + (a + bc_\mu)\left\langle x_n - x_{n+1}, \tilde{x} - \frac{x_{n+1} + x_n}{2} \right\rangle, \end{aligned}$$

that is,

$$\begin{aligned}U_{n+1}-U_n-\beta b(g(x_{n+1})-g(x_n))&\\ =a\langle\tilde{x}-x_{n+1},x_{n+1}-x_n\rangle+b\langle\tilde{x}-x_{n+1},y_{n+1}-y_n\rangle&\\ +\langle x_n-x_{n+1},ax_n+by_n\rangle&\\ +(a+bc_\mu)\langle x_n-x_{n+1},\tilde{x}-x_{n+1}\rangle+(a+bc_\mu)\left\langle x_n-x_{n+1},\frac{x_{n+1}-x_n}{2}\right\rangle;&\end{aligned}$$

hence

$$\begin{aligned}U_{n+1}-U_n&=\beta b(g(x_{n+1})-g(x_n))\\ &\quad+b\langle\tilde{x}-x_{n+1},y_{n+1}-y_n-c_\mu(x_{n+1}-x_n)\rangle\\ &\quad+\langle x_n-x_{n+1},ax_n+by_n\rangle-\frac{a+bc_\mu}{2}|x_{n+1}-x_n|^2.\end{aligned}\tag{4.8}$$

Let us estimate the first term and the third term in the right-hand side of the previous equality. By (2.1)(ii), we also have

$$\begin{aligned}y_n-y_{n+1}&=a[\nu x_{n+1}+(1-\nu)x_n]+b[\tau y_{n+1}+(1-\tau)y_n]\\ &=a\nu(x_{n+1}-x_n)+b\tau(y_{n+1}-y_n)+(ax_n+by_n),\end{aligned}$$

that is,

$$ax_n+by_n=-(1+b\tau)(y_{n+1}-y_n)-a\nu(x_{n+1}-x_n).\tag{4.9}$$

Referring to (3.6), for any $p_{n+1}\in\partial_{\epsilon_n}g(x_{n+1})$ we additionally get

$$g(x_{n+1})-g(x_n)\le\langle p_{n+1},x_{n+1}-x_n\rangle+\epsilon_n.\tag{4.10}$$

Combining (4.8), (4.9), and (4.10), we deduce

$$\begin{aligned}U_{n+1}-U_n&\le\beta b(\langle p_{n+1},x_{n+1}-x_n\rangle+\epsilon_n)\\ &\quad+b\langle\tilde{x}-x_{n+1},y_{n+1}-y_n-c_\mu(x_{n+1}-x_n)\rangle\\ &\quad+\langle x_n-x_{n+1},-(1+b\tau)(y_{n+1}-y_n)-a\nu(x_{n+1}-x_n)\rangle\\ &\quad-\frac{a+bc_\mu}{2}|x_{n+1}-x_n|^2,\end{aligned}\tag{4.11}$$

namely,

$$\begin{aligned}U_{n+1}-U_n&\le b\left(\beta\langle p_{n+1},x_{n+1}-x_n\rangle+\langle\tilde{x}-x_{n+1},y_{n+1}-y_n-c_\mu(x_{n+1}-x_n)\rangle\right)\\ &\quad+b\beta\epsilon_n+(1+\tau b)\langle x_{n+1}-x_n,y_{n+1}-y_n\rangle\\ &\quad-[bc_\mu/2-a(\nu-1/2)]|x_{n+1}-x_n|^2.\end{aligned}\tag{4.12}$$

In addition, by (2.3) with $\psi=h\equiv 0$ and taking into account the fact that $\beta=\frac{\lambda}{\mu k}$, we have

$$c_\mu(x_{n+1}-x_n)+\beta[\partial_{\epsilon_n}g(x_{n+1})]-(y_{n+1}-y_n)\ni 0,\tag{4.13}$$

so that there exists $p_{n+1} \in \partial_{\epsilon_n} g(x_{n+1})$ such that

$$\beta p_{n+1} = (y_{n+1} - y_n) - c_\mu(x_{n+1} - x_n); \tag{4.14}$$

hence

$$\begin{aligned} &\beta\langle p_{n+1}, x_{n+1} - \tilde{x}\rangle + \beta\langle p_{n+1}, x_{n+1} - x_n\rangle \\ &\quad = \langle (y_{n+1} - y_n) - c_\mu(x_{n+1} - x_n), (x_{n+1} - \tilde{x}) + (x_{n+1} - x_n)\rangle. \end{aligned} \tag{4.15}$$

As $p_{n+1} \in \partial_{\epsilon_n} g(x_{n+1})$ and $0 \in \partial g(\tilde{x})$, from the definitions of ∂g and $\partial_{\epsilon_n} g$, we additionally have

$$\langle p_{n+1}, x_{n+1} - \tilde{x}\rangle \geq -\epsilon_n,$$

so that

$$\begin{aligned} &\beta\langle p_{n+1}, x_{n+1} - x_n\rangle \\ &\quad \leq \langle (y_{n+1} - y_n) - c_\mu(x_{n+1} - x_n), (x_{n+1} - \tilde{x}) + (x_{n+1} - x_n)\rangle + \beta\epsilon_n. \end{aligned} \tag{4.16}$$

This can be equivalently rewritten as

$$\begin{aligned} &\beta\langle p_{n+1}, x_{n+1} - x_n\rangle + \langle \tilde{x} - x_{n+1}, y_{n+1} - y_n - c_\mu(x_{n+1} - x_n)\rangle \\ &\quad \leq \langle (y_{n+1} - y_n) - c_\mu(x_{n+1} - x_n), x_{n+1} - x_n\rangle + \beta\epsilon_n, \end{aligned} \tag{4.17}$$

which along with (4.12) leads to

$$\begin{aligned} U_{n+1} - U_n \leq\ & b\langle (y_{n+1} - y_n) - c_\mu(x_{n+1} - x_n), x_{n+1} - x_n\rangle \\ & + 2b\beta\epsilon_n + (1 + \tau b)\langle x_{n+1} - x_n, y_{n+1} - y_n\rangle \\ & - [bc_\mu/2 - a(\nu - 1/2)]|x_{n+1} - x_n|^2, \end{aligned}$$

that is,

$$\begin{aligned} U_{n+1} - U_n \leq\ & [1 + (\tau + 1)b]\langle y_{n+1} - y_n, x_{n+1} - x_n\rangle \quad + 2b\beta\epsilon_n \\ & - \left(\frac{3}{2}bc_\mu - a\left(\nu - \frac{1}{2}\right)\right)|x_{n+1} - x_n|^2. \end{aligned} \tag{4.18}$$

This last inequality can be obviously written as

$$U_{n+1} - U_n \leq S_n, \tag{4.19}$$

where

$$\begin{aligned} S_n := \ & 2b\beta\epsilon_n + [1 + (\tau + 1)b]\langle x_{n+1} - x_n, y_{n+1} - y_n\rangle \\ & -[\tfrac{3}{2}bc_\mu - a(\nu - \tfrac{1}{2})]|x_{n+1} - x_n|^2. \end{aligned} \tag{4.20}$$

It is then immediate that $U_n - \sum_{k=0}^{n-1} S_k$ is a nonincreasing sequence. Furthermore, we obviously observe that $\sum_{n\geq 0} S_n$ is bounded, because $\sum_{n\geq 0} \epsilon_n < \infty$ and, by Theorem 3.2, the two sums $\sum_{n\geq 0} |x_{n+1} - x_n|^2$ and $\sum_{n\geq 0} |x_{n+1} - x_n|^2$ are finite. It turns out that U_n is bounded from above, and by (4.7) we have

$$-|\tilde{x} - x_n| \times |ax_n + by_n| + \frac{1}{2}(bc_\mu + a)|\tilde{x} - x_n|^2 + b\beta g(x_n) \leq U_n; \tag{4.21}$$

from (4.9) and Theorem 3.2, we also note that the quantity $|ax_n + by_n|$ is bounded since it converges to zero. By an easy computation, we also have

$$\frac{(2+\alpha+\gamma)(a+bc_\mu)}{2} = \alpha + \gamma\frac{1-\mu\nu\alpha}{k\mu} = (\alpha+\gamma) + \gamma\frac{1-\mu(\nu\alpha+k)}{k\mu},$$

so that $a + bc_\mu > 0$, since $\alpha + \gamma > 0$, $\gamma > 0$, $\mu > 0$, $k > 0$, and $1 - \mu(\nu\alpha + k) \geq 0$.

Consequently, we deduce that $|\tilde{x} - x_n|$ is bounded; hence U_n is bounded from below and is convergent. On the other hand, by Theorem 3.2, we know that $g(x_n)$ also converges. As a consequence, observing that

$$|\tilde{x} - x_n|^2 = \frac{2}{a+bc_\mu}\left(U_n - \langle \tilde{x} - x_n, ax_n + by_n\rangle - b\beta g(x_n)\right)$$

and since $\langle \tilde{x}-x_n, ax_n+by_n\rangle \to 0$, we deduce the convergence of the sequence $(|\tilde{x}-x_n|)$. Taking into account (4.13), we can easily check that any weak-cluster point of (x_n) is in Argmin g, because $|x_{n+1} - x_n| \to 0$, $|y_{n+1} - y_n| \to 0$, and the graph of ∂g is weakly-strongly closed. Applying Opial's lemma with $C = \text{Argmin } g$, we conclude that the sequence (x_n) weakly converges to a minimizer of g. This completes the proof. □

Finally, we would like to emphasize again that in the case where $\psi = h \equiv 0$, $\alpha > 0, \gamma > 0$, and the parameters τ and ν satisfy (1.5), DPM is nothing but the inertial proximal method. Moreover, $k = \gamma$, and thus condition (4.6) can be rewritten as $\mu < \frac{1}{\alpha+\gamma+1}$. This ensures that $\mu \in (0,1)$, and we recover Theorem 3.1 by Alvarez [1].

Conclusion. The main purpose of this article is to establish the asymptotic convergence of some new implicit iterative methods for solving a nonconvex minimization problem which is a natural extension of differentiable, convex, and DC programming. These algorithms, which generalize Sun, Sampaio, and Candido's scheme [27], the classical proximal algorithm [25], the inertial proximal method [1], and the gradient algorithm [15], are obtained by a discretization of a first order dissipative dynamical system. Particular attention to the convex case is also given, and the results obtained are nothing but discrete versions of those proposed in the continuous case by Alvarez, Attouch, Bolte, and Redont [3] and [5]. We think that the results obtained in this paper may inspire and pave the way for future research in this field, especially in developing new hybrid algorithms which admit less stringent requirements on solving proximal subproblems in the spirit of Solodov and Svaiter [26], who showed that the tolerance requirements for solving subproblems can be significantly relaxed if the solving of each subproblem is followed by a projection onto a certain hyperplane which separates the current iterate from the solution set of the problem.

Acknowledgments. The authors would like to thank the anonymous referees and Professor Adrian Lewis for their careful reading of the paper and for their comments and suggestions which permitted us to improve the presentation.

REFERENCES

[1] F. Alvarez, *On the minimizing property of a second order dissipative dynamical system in Hilbert spaces*, SIAM J. Control Optim., 38 (2000), pp. 1102–1119.

[2] F. Alvarez and H. Attouch, *An inertial proximal method for monotone operators via discretization of a nonlinear oscillator with damping*, Set-Valued Anal., 9 (2001), pp. 3–31.

[3] F. Alvarez, H. Attouch, J. Bolte, and P. Redont, *A second-order gradient-like dissipative dynamical system with Hessian driven damping. Application to optimization and mechanics*, J. Math. Pures Appl. (9), 81 (2002), pp. 747–779.

[4] F. Alvarez and J. M. Pérez, *A dynamical system associated with Newton's method for parametric approximations of convex minimization problems*, Appl. Math. Optim., 38 (1998), pp. 193–217.

[5] H. Attouch, J. Bolte, and P. Redont, *A second-order gradient-like dissipative dynamical system with Hessian driven damping. Link with the proximal algorithm*, Control Cybernet., 31 (2002), pp. 643–657.

[6] H. Attouch, X. Goudon, and P. Redont, *The heavy ball system with friction method*, I. *The continuous dynamical system*, Commun. Contemp. Math., 2 (2000), pp. 1–34.

[7] H. Attouch and P. Redont, *The second-order in time continuous Newton method*, in Approximation Optimization and Mathematical Economics, M. Lassonde ed., Physica, Heidelberg, 2001, pp. 25–36.

[8] J.-P. Chehab and M. Raydan, *Implicit and adaptive inverse preconditioned gradient methods for nonlinear problems*, Appl. Numer. Math., 55 (2005), pp. 32–47.

[9] O. Güler, *On the convergence of the proximal point algorithm for convex minimization*, SIAM J. Control Optim., 29 (1991), pp. 403–419.

[10] J. B. Hiriart-Urruty, *From convex optimization to nonconvex optimization*, in Nonsmooth Optimization and Related Topics, F. H. Clarke, V. F. Demyanov, and F. Giannessi, eds., Plenum Press, New York, 1989, pp. 219–239.

[11] J. B. Hiriart-Urruty, *Generalized differentiability, duality and optimization for problems dealing with difference of convex functions*, in Convexity and Duality in Optimization, Lecture Notes in Econom. Math. Systems 256, Springer-Verlag, New York, 1986, pp. 37–70.

[12] F. Jules and P. E. Maingé, *Numerical approaches to stationary solution of a second order dissipative dynamical system*, Optimization, 51 (2002), pp. 235–255.

[13] K. C. Kiwiel, *Proximity control in bundle methods for convex nondifferentiable minimization*, Math. Programming, 46 (1990), pp. 105–122.

[14] K. Goebel and W. A. Kirk, *Topics in Metric Fixed Points Theory*, Cambridge Studies in Advanced Mathematics 28, Cambridge University Press, Cambridge, UK, 1990.

[15] B. Lemaire, *On the convergence of some iterative methods for convex minimization*, in Recent Developments in Optimization, Lecture Notes in Econom. and Math. Systems 429, Springer-Verlag, Berlin, 1995, pp. 252–268.

[16] B. Martinet, *Algorithmes pour la résolution des problèmes d'optimisation et de minmax*, Thèse d'état Université de Grenoble, Grenoble, France, 1972.

[17] J. E. Martinez-Legaz and A. Seeger, *A formula on the approximate subdifferential of the difference of convex functions*, Bull. Austral. Math. Soc., 45 (1992), pp. 37–41.

[18] Ph. Michel, *Problème d'optimisation défini par des fonctions qui sont somme de fonctions convexes et de fonctions dérivables*, J. Math. Pures Appl. (9), 53 (1974), pp. 321–329.

[19] A. Moudafi and E. Elisabeth, *An approximate inertial proximal method using enlargement of a maximal monotone operator*, Int. J. Pure Appl. Math., 5 (2003), pp. 283–299.

[20] A. Moudafi and P.-E. Maingé, *On the convergence of an approximate proximal method for DC functions*, J. Comput. Math., 24 (2006), pp. 475–480.

[21] A. Moudafi and M. Théra, *Finding the zero for the sum of two maximal monotone operators*, J. Optim. Theory Appl., 94 (1997), pp. 425–448.

[22] Z. Opial, *Weak convergence of the sequence of successive approximations for nonexpansive mappings*, Bull. Amer. Math. Soc., 73 (1967), pp. 591–597.

[23] J. P. Penot, *On the minimization of difference functions*, J. Global Optim., 12 (1998), pp. 373–382.

[24] D. T. Pham and E. B. Souad, *Algorithms for solving a class of nonconvex optimization problems: Methods of subgradient*, in Fermat Days 85: Mathematics for Optimization, North-Holland Math. Stud. 129, Elsevier, Amsterdam, 1986, pp. 249–270.

[25] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.

[26] M. V. Solodov and B. F. Svaiter, *A hybrid projection-proximal point algorithm*, J. Convex Anal., 6 (1999), pp. 59–70.

[27] W. Y. Sun, R. J. B. Sampaio, and M. A. B. Candido, *Proximal point algorithm for minimization of DC functions*, J. Comput. Math., 21 (2003), pp. 451–462.

[28] A. Szulkin, *Minimax principles for lower semi-continuous functions and applications to nonlinear boundary value problems*, Annales Inst. H. Poincaré Anal. Non Linéaire, 3 (1986), pp. 77–109.

[29] J. F. Toland, *Duality in nonconvex optimization*, J. Math. Anal. Appl., 66 (1978), pp. 399–415.

SIAM J. OPTIM.
Vol. 19, No. 1, pp. 414–444

RECURSIVE TRUST-REGION METHODS FOR MULTISCALE NONLINEAR OPTIMIZATION*

SERGE GRATTON†, ANNICK SARTENAER‡, AND PHILIPPE L. TOINT‡

Abstract. A class of trust-region methods is presented for solving unconstrained nonlinear and possibly nonconvex discretized optimization problems, like those arising in systems governed by partial differential equations. The algorithms in this class make use of the discretization level as a means of speeding up the computation of the step. This use is recursive, leading to true multilevel/multiscale optimization methods reminiscent of multigrid methods in linear algebra and the solution of partial differential equations. A simple algorithm of the class is then described and its numerical performance is shown to be numerically promising. This observation then motivates a proof of global convergence to first-order stationary points on the fine grid that is valid for all algorithms in the class.

Key words. nonlinear optimization, multiscale problems, simplified models, recursive algorithms, convergence theory

AMS subject classifications. 90C30, 65K05, 90C26, 90C06

DOI. 10.1137/050623012

1. Introduction.

Large-scale finite-dimensional optimization problems often arise from the discretization of infinite-dimensional problems, a primary example being optimal control problems defined in terms of either ordinary or partial differential equations [8]. While the direct solution of such problems for a discretization level is often possible using existing packages for large-scale numerical optimization, this technique typically makes very little use of the fact that the underlying infinite-dimensional problem may be described at several discretization levels; the approach thus rapidly becomes cumbersome. Motivated by this observation, we explore here a class of algorithms which makes explicit use of this fact.

Using the different levels of discretization for an infinite-dimensional problem is not a new idea. A simple first approach is to use coarser grids in order to compute approximate solutions which can then be used as starting points for the optimization problem on a finer grid (see [5, 6, 7, 22], for instance). Other efficient techniques are inspired from the multigrid paradigm in the solution of partial differential equations and associated systems of linear algebraic equations (see, for example, [10, 11, 12, 23, 41, 43] for descriptions and references).

The purpose of our paper is threefold. We first introduce a new extension of the full approximation scheme (FAS) (see, for instance, Chapter 3 of [12] or [25]), an existing multigrid-type method, to a class of trust-region based optimization algorithms. We then indicate that this class contains numerically efficient members, thereby motivating further analysis. We finally provide a global convergence proof for all members of the class, which gives a robustness guarantee typical in optimization but, to the authors' knowledge, uncommon in multigrid approaches. Significantly, this guarantee holds even for nonconvex (nonelliptic) problems.

*Received by the editors January 20, 2005; accepted for publication (in revised form) September 17, 2007; published electronically April 16, 2008.
http://www.siam.org/journals/siopt/19-1/62301.html
†CNES and CERFACS, Toulouse, France (gratton@cerfacs.fr).
‡Department of Mathematics, University of Namur, B-5000 Namur, Belgium (annick.sartenaer@fundp.ac.be, philippe.toint@fundp.ac.be).

The work presented here was in particular motivated by the paper by Gelman and Mandel [16], the "generalized truncated Newton algorithm" presented in Fisher [15], a talk by Moré [28], and the contributions by Nash and coauthors [26, 27, 30]. These latter three papers present the description of MG/OPT, a linesearch-based recursive algorithm, an outline of its convergence properties, and impressive numerical results. The generalized truncated Newton algorithm and MG/OPT are very similar and, like many linesearch methods, naturally suited to convex problems, although their generalization to the nonconvex case is possible. An older contribution for convex problems is the damped nonlinear multilevel method by Hackbusch and Reusken [24], where convergence is analyzed for a variant of the FAS under the condition that a Lipschitz constant for the problem Hessian is explicitly known or can be numerically estimated. In the same spirit, the very recent contribution by Yavneh and Dardyk [45] considers a linesearch to improve the radius of local convergence of a nonlinear equations solver. Further motivation to consider the more general nonconvex problem is also provided by the computational success of the low/high-fidelity model management techniques of Alexandrov, Lewis, and coauthors [2, 3] and a paper by Borzi and Kunisch [9] on multigrid globalization.

The class of algorithms discussed in this note can be viewed as an alternative where one uses the trust-region technology whose efficiency and reliability in the solution of nonconvex problems is well known (we refer the reader to [13] for more complete coverage of this subject). Our developments are organized as follows. We first describe our class of multiscale trust-region algorithms in section 2 and show in section 3 that it can be specialized to a multigrid method that performs well on examples. This observation then motivates the proof of global convergence to first-order critical points presented in section 4. The main results of this section are Theorem 4.10, which establishes a level-independent complexity bound for general trust-region algorithms, and Theorem 4.13, which is the desired convergence property. Some conclusions and perspectives are presented in section 5.

2. Recursive multiscale trust-region algorithms. We start by considering the solution of the unconstrained optimization problem

$$\min_{x\in\Re^n} f(x), \tag{2.1}$$

where f is a twice-continuously differentiable objective function which maps $\Re^n$ into $\Re$ and is bounded below. The trust-region methods which we investigate are iterative: given an initial point x_0, they produce a sequence $\{x_k\}$ of iterates (hopefully) converging to a first-order critical point for the problem, i.e., to a point where $g(x) \stackrel{\text{def}}{=} \nabla f(x) = 0$. At each iterate x_k, trust-region methods build a model $m_k(x)$ of $f(x)$ around x_k. This model is then assumed to be adequate in a "trust region," defined as a sphere of radius $\Delta_k > 0$ centered at x_k, and a step s_k is then computed such that the trial point $x_k + s_k$ sufficiently reduces this model in the region. The objective function is computed at $x_k + s_k$ and the trial point is accepted as the next iterate if the ratio of achieved to predicted reduction is larger than a small positive constant. The value of the radius is finally updated to ensure that it is decreased when the trial point cannot be accepted as the next iterate and is increased or unchanged otherwise. In many practical trust-region algorithms, the model m_k is quadratic, and obtaining sufficient decrease then amounts to (approximately) solving

$$\min_{\|s\|\le\Delta_k} m_k(x_k+s) = \min_{\|s\|\le\Delta_k} f(x_k) + \langle g_k, s\rangle + \tfrac{1}{2}\langle s, H_k s\rangle \tag{2.2}$$

for s, where $g_k \stackrel{\rm def}{=} \nabla f(x_k)$, H_k is a symmetric $n \times n$ approximation of $\nabla^2 f(x_k)$, $\langle \cdot, \cdot \rangle$ is the Euclidean inner product, and $\| \cdot \|$ is the Euclidean norm.

Such methods are efficient and reliable and provably converge to first-order critical points whenever the sequence $\{\|H_k\|\}$ is uniformly bounded. Besides computing the value $f(x_k + s_k)$, their work per iteration is dominated by the numerical solution of the subproblem (2.2), which crucially depends on the dimension n of the problem. When (2.1) results from the discretization of some infinite-dimensional problem on a relatively fine grid, the solution cost is therefore often significant.

In what follows, we investigate what can be done to reduce this cost by exploiting the knowledge of alternative simplified expressions of the objective function, when available. More specifically, we assume that we know a collection of functions $\{f_i\}_{i=0}^r$ such that each f_i is a twice-continuously differentiable function from $\Re^{n_i}$ to $\Re$ (with $n_i \geq n_{i-1}$), the connection with our original problem being that $n_r = n$ and $f_r(x) = f(x)$ for all $x \in \Re^n$. We will also assume that, for each $i = 1, \ldots, r$, f_i is "more costly" to minimize than f_{i-1}. This may be because f_i has more variables than f_{i-1} (as would typically be the case if the f_i represent increasingly finer discretizations of the same infinite-dimensional objective), or because the structure (in terms of partial separability, sparsity, or eigenstructure) of f_i is more complex than that of f_{i-1}, or for any other reason. To fix terminology, we will refer to a particular i as a *level*.

Of course, for f_{i-1} to be useful at all in minimizing f_i, there should be some relation between the variables of these two functions. We henceforth assume that, for each $i = 1, \ldots, r$, there exist a full-rank linear operator R_i from $\Re^{n_i}$ into $\Re^{n_{i-1}}$ (the restriction) and another full-rank operator P_i from $\Re^{n_{i-1}}$ into $\Re^{n_i}$ (the prolongation) such that

$$\sigma_i P_i = R_i^T \tag{2.3}$$

for some known constant $\sigma_i > 0$. In the context of multigrid algorithms, P_i and R_i are interpreted as restriction and prolongation between a fine and a coarse grid (see [12], for instance). This assumption is also used in Nash [30].

The main idea is then to use f_{r-1} to construct an alternative model h_{r-1} for $f_r = f$ in the neighborhood of the current iterate that is cheaper than the quadratic model at level r, and to use this alternative model, whenever suitable, to define the step in the trust-region algorithm. If more than two levels are available ($r > 1$), this can be done recursively, the approximation process stopping at level 0, where the quadratic model is always used. In what follows, we use a simple notation where quantities of interest have a double subscript i, k. The first, i ($0 \leq i \leq r$), is the level index (meaning, in particular, if applied to a vector, that this vector belongs to $\Re^{n_i}$), and the second, k, is the index of the current iteration *within level* i and is *reset to 0 each time level* i *is entered*.[1]

Consider now some iteration k at level i (with current iterate $x_{i,k}$) and suppose that one decides to use the lower level model h_{i-1} based on f_{i-1} to compute a step. The first task is to restrict $x_{i,k}$ to create the starting iterate $x_{i-1,0}$ at level $i-1$, that is, $x_{i-1,0} = R_i x_{i,k}$. We then define the lower level model by

$$h_{i-1}(x_{i-1,0} + s_{i-1}) \stackrel{\rm def}{=} f_{i-1}(x_{i-1,0} + s_{i-1}) + \langle v_{i-1}, s_{i-1} \rangle, \tag{2.4}$$

[1] We are well aware that this creates some ambiguities, since a sequence of indices i, k can occur more than once if level i ($i < r$) is used more than once, implying the existence of more than one starting iterate at this level. This ambiguity is resolved by the context.

where $v_{i-1} = R_i g_{i,k} - \nabla f_{i-1}(x_{i-1,0})$ with $g_{i,k} \stackrel{\text{def}}{=} \nabla h_i(x_{i,k})$. By convention, we set $v_r = 0$ such that, for all s_r,

$$h_r(x_{r,0} + s_r) = f_r(x_{r,0} + s_r) = f(x_{r,0} + s_r) \text{ and } g_{r,k} = \nabla h_r(x_{r,k}) = \nabla f(x_{r,k}).$$

The function h_i therefore corresponds to a modification of f_i by a linear term that enforces the relation

$$g_{i-1,0} = \nabla h_{i-1}(x_{i-1,0}) = R_i g_{i,k}. \tag{2.5}$$

The first-order modification (2.4) is not unusual in multigrid applications in the context of the FAS and is also used by Fisher [15] and Nash [30]. It crucially ensures that the first-order behaviors of h_i and h_{i-1} are coherent in a neighborhood of $x_{i,k}$ and $x_{i-1,0}$, respectively: indeed, one verifies that, if s_i and s_{i-1} satisfy $s_i = P_i s_{i-1}$, then

$$\langle g_{i,k}, s_i\rangle = \langle g_{i,k}, P_i s_{i-1}\rangle = \frac{1}{\sigma_i}\langle R_i g_{i,k}, s_{i-1}\rangle = \frac{1}{\sigma_i}\langle g_{i-1,0}, s_{i-1}\rangle, \tag{2.6}$$

where we have also used (2.3) and (2.5). This coherence was independently imposed in [26] and, in a slightly different context, in [2] and other papers on first-order model management.

Our task, when entering level $i = 0, \ldots, r$, is then to (locally) minimize h_i starting from $x_{i,0}$. At iteration k of this minimization, we first choose, at iterate $x_{i,k}$, either the model $h_{i-1}(x_{i-1,0} + s_{i-1})$ (given by (2.4)) or

$$m_{i,k}(x_{i,k} + s_i) = h_i(x_{i,k}) + \langle g_{i,k}, s_i\rangle + \tfrac{1}{2}\langle s_i, H_{i,k} s_i\rangle, \tag{2.7}$$

where the latter is the usual truncated Taylor series in which $H_{i,k}$ is a symmetric $n_i \times n_i$ approximation to the second derivatives of h_i (which are also the second derivatives of f_i) at $x_{i,k}$. Once the model is chosen (we will return to the conditions of this choice below), we then compute a step $s_{i,k}$ that generates a decrease on this model within a trust region $\{s_i \mid \|s_i\|_i \le \Delta_{i,k}\}$ for some trust-region radius $\Delta_{i,k} > 0$. The norm $\|\cdot\|_i$ in this last expression is level-dependent and defined, for some symmetric positive-definite matrix M_i, by

$$\|s_i\|_i \stackrel{\text{def}}{=} \sqrt{\langle s_i, M_i s_i\rangle} \stackrel{\text{def}}{=} \|s_i\|_{M_i}. \tag{2.8}$$

If the model (2.7) is chosen,[2] this is nothing but a usual ellipsoidal trust-region subproblem solution yielding a step $s_{i,k}$. The decrease of the model $m_{i,k}$ is then understood in its usual meaning for trust-region methods, which is to say that $s_{i,k}$ is such that

$$m_{i,k}(x_{i,k}) - m_{i,k}(x_{i,k} + s_{i,k}) \ge \kappa_{\text{red}} \|g_{i,k}\| \min\left[\frac{\|g_{i,k}\|}{1 + \|H_{i,k}\|}, \Delta_{i,k}\right] \tag{2.9}$$

for some constant $\kappa_{\text{red}} \in (0, 1)$. This condition is known as the "sufficient decrease" or "Cauchy point" condition. Chapter 7 of [13] reviews several techniques that enforce it, including the exact minimization of $m_{i,k}$ within the trust region or an approximate minimization using (possibly preconditioned) Krylov space methods. On the other

[2] Observe that this is the only possible choice for $i = 0$.

hand, if the model h_{i-1} is chosen, minimization of this latter model (hopefully) produces a new point $x_{i-1,*}$ such that $h_{i-1}(x_{i-1,*}) < h_{i-1}(x_{i-1,0})$ and a corresponding step $x_{i-1,*} - x_{i-1,0}$ which must then be brought back to level i by the prolongation P_i. Since

$$\|s_i\|_i = \|s_i\|_{M_i} = \|P_i s_{i-1}\|_{M_i} = \|s_{i-1}\|_{P_i^T M_i P_i} \stackrel{\text{def}}{=} \|s_{i-1}\|_{M_{i-1}} = \|s_{i-1}\|_{i-1} \tag{2.10}$$

(which is well defined since P_i is full-rank), the trust-region constraint at level $i-1$ then becomes

$$\|x_{i-1,*} - x_{i-1,0}\|_{i-1} \leq \Delta_{i,k}. \tag{2.11}$$

The lower level subproblem consists in (possibly approximately) solving

$$\min_{\|s_{i-1}\|_{i-1} \leq \Delta_{i,k}} h_{i-1}(x_{i-1,0} + s_{i-1}). \tag{2.12}$$

The relation (2.10) also implies that, for $i = 0 \ldots, r-1$,

$$M_i = Q_i^T Q_i, \ \text{ where } \ Q_i = P_r \ldots P_{i+2} P_{i+1}, \tag{2.13}$$

while we define $M_r = I$ for consistency. Preconditioning can also be accommodated by choosing M_r more elaborately.

Is the cheaper model h_{i-1} always useful? Obviously not, as it may happen, for instance, that $g_{i,k}$ lies in the nullspace of R_i and thus that $R_i g_{i,k}$ is zero while $g_{i,k}$ is not. In this case, the current iterate appears to be first-order critical for h_{i-1} in $\Re^{n_{i-1}}$ while it is not for h_i in $\Re^{n_i}$. Using the model h_{i-1} is hence potentially useful only if $\|g_{i-1,0}\| = \|R_i g_{i,k}\|$ is large enough compared to $\|g_{i,k}\|$. We therefore restrict the use of the model h_{i-1} to iterations where

$$\|R_i g_{i,k}\| \geq \kappa_{\text{g}} \|g_{i,k}\| \ \text{ and } \ \|R_i g_{i,k}\| > \epsilon_{i-1}^{\text{g}} \tag{2.14}$$

for some constant $\kappa_{\text{g}} \in (0, \min[1, \min_i \|R_i\|])$ and where $\epsilon_{i-1}^{\text{g}} \in (0,1)$ is a measure of the first-order criticality for h_{i-1} that is judged sufficient at level $i-1$. Note that, given $g_{i,k}$ and R_i, this condition is easy to check before even attempting to compute a step at level $i-1$.

We are now in a position to describe our recursive multiscale trust-region (RMTR) algorithm more formally as Algorithm 2.1.

In this description, we use the constants η_1, η_2, γ_1, and γ_2 satisfying the conditions $0 < \eta_1 \leq \eta_2 < 1$ and $0 < \gamma_1 \leq \gamma_2 < 1$. It is assumed that the prolongations/restrictions P_i and R_i are known, as are the description of the levels $i = 0, \ldots, r$. An initial trust-region radius for each level $\Delta_i^{\text{s}} > 0$ is also defined, as well as level-dependent gradient norm tolerances $\epsilon_i^{\text{g}} \in (0,1)$ and trust-region tolerances $\epsilon_i^{\Delta} \in (0,1)$ for $i = 0, \ldots, r$. The algorithm's initial data consists of the level index i ($0 \leq i \leq r$), a starting point $x_{i,0}$, the gradient $g_{i,0}$ at this point, the radius Δ_{i+1} of the level-$(i+1)$ trust region, and the tolerances ϵ_i^{g} and ϵ_i^{Δ}. The original task of minimizing $f(x) = f_r(x_r) = h_r(x_r)$ (up to the gradient norm tolerance $\epsilon_r^{\text{g}} < \|\nabla f_r(x_{r,0})\|$) is achieved by calling RMTR(r, $x_{r,0}$, $\nabla f_r(x_{r,0})$, $\Delta_{r+1,0}$, ϵ_r^{g}, ϵ_r^{Δ}, Δ_r^{s}) for some starting point $x_{r,0}$ and initial trust-region radius Δ_r^{s}, and where we define $\Delta_{r+1,0} = \infty$. For coherence of notation, we thus view this call as being made with an infinite radius from some (virtual) iteration 0 at level $r+1$. The motivation for (2.17) in Step 6 of the algorithm and the termination test $\|x_{i,k+1} - x_{i,0}\|_i > (1 - \epsilon_i^{\Delta})\Delta_{i+1}$ in Step 5 are

Algorithm 2.1: RMTR(i, $x_{i,0}$, $g_{i,0}$, Δ_{i+1}, ϵ_i^{g}, ϵ_i^{Δ}, Δ_i^{s})

Step 0: Initialization. Compute $v_i = g_{i,0} - \nabla f_i(x_{i,0})$ and $h_i(x_{i,0})$. Set $\Delta_{i,0} = \min[\Delta_i^{\mathrm{s}}, \Delta_{i+1}]$ and $k = 0$.

Step 1: Model choice. If $i = 0$ or if (2.14) fails, go to Step 3. Otherwise, choose to go to Step 2 (recursive step) or to Step 3 (Taylor step).

Step 2: Recursive step computation.
Call Algorithm RMTR($i-1$, $R_i x_{i,k}$, $R_i g_{i,k}$, $\Delta_{i,k}$, $\epsilon_{i-1}^{\mathrm{g}}$, ϵ_{i-1}^{Δ}, $\Delta_{i-1}^{\mathrm{s}}$), yielding an approximate solution $x_{i-1,*}$ of (2.12). Then define $s_{i,k} = P_i(x_{i-1,*} - R_i x_{i,k})$, set $\delta_{i,k} = h_{i-1}(R_i x_{i,k}) - h_{i-1}(x_{i-1,*})$, and go to Step 4.

Step 3: Taylor step computation. Choose $H_{i,k}$ and compute a step $s_{i,k} \in \Re^{n_i}$ that sufficiently reduces the model $m_{i,k}$ (given by (2.7)) in the sense of (2.9) and such that $\|s_{i,k}\|_i \le \Delta_{i,k}$. Set $\delta_{i,k} = m_{i,k}(x_{i,k}) - m_{i,k}(x_{i,k}+s_{i,k})$.

Step 4: Acceptance of the trial point. Compute $h_i(x_{i,k} + s_{i,k})$ and define

$$\rho_{i,k} = (h_i(x_{i,k}) - h_i(x_{i,k} + s_{i,k}))/\delta_{i,k}. \tag{2.15}$$

If $\rho_{i,k} \ge \eta_1$, then define $x_{i,k+1} = x_{i,k}+s_{i,k}$; otherwise define $x_{i,k+1} = x_{i,k}$.

Step 5: Termination. Compute $g_{i,k+1}$. If $\|g_{i,k+1}\|_\infty \le \epsilon_i^{\mathrm{g}}$ or $\|x_{i,k+1} - x_{i,0}\|_i > (1 - \epsilon_i^{\Delta})\Delta_{i+1}$, then return with the approximate solution $x_{i,*} = x_{i,k+1}$.

Step 6: Trust-region radius update. Set

$$\Delta_{i,k}^{+} \in \begin{cases} [\Delta_{i,k}, +\infty) & \text{if } \rho_{i,k} \ge \eta_2, \\ [\gamma_2 \Delta_{i,k}, \Delta_{i,k}] & \text{if } \rho_{i,k} \in [\eta_1, \eta_2), \\ [\gamma_1 \Delta_{i,k}, \gamma_2 \Delta_{i,k}] & \text{if } \rho_{i,k} < \eta_1, \end{cases} \tag{2.16}$$

and

$$\Delta_{i,k+1} = \min\left[\Delta_{i,k}^{+}, \Delta_{i+1} - \|x_{i,k+1} - x_{i,0}\|_i\right]. \tag{2.17}$$

Increment k by one and go to Step 1.

to guarantee that iterates at a lower level in a recursion remain in the trust region defined at the calling level, as verified below in Lemma 4.1.

Iteration k at level i, associated with the computation of the step $s_{i,k}$, will be referred to as iteration (i, k). It will be called a *Taylor iteration* if Step 3 is used (that is, if Taylor's model $m_{i,k}(x_{i,k} + s_i)$ is chosen at Step 1). If Step 2 is used instead, iteration (i, k) will then be called a *recursive iteration*. We emphasize that we expect the most efficient algorithms in our class to make use of a combination of both iteration types, which means, in particular, that recursive iterations should not be automatic if (2.14) holds. As is usual for trust-region methods, iteration (i, k) is said to be *successful* if $\rho_{i,k} \ge \eta_1$, that is, if the trial point $x_{i,k} + s_{i,k}$ is accepted as the next iterate $x_{i,k+1}$. It is said to be *very successful* if $\rho_{i,k} \ge \eta_2$, implying that $\Delta_{i,k}^{+} \ge \Delta_{i,k}$.

In the case where $r = 0$, that is, if there is only one level in the problem, the algorithm reduces to the well-known usual trust-region method (see p. 116 of [13]) and enjoys all the desirable properties of this method. If $r > 0$, the recursive nature

of Algorithm RMTR is clear from Step 2. It is, in that sense, reminiscent of multigrid methods for linear systems [23] and is close in spirit to the MG/OPT method [30]. However, this latter method differs from ours in two main respects: Algorithm RMTR is of trust-region type, and its global convergence properties considered in this paper do not rely on performing Taylor iterations before or after a recursive one. Algorithm RMTR can also be viewed as an extension of the low-/high-fidelity model management method of [2] and [3]. The main differences are that our framework explicitly uses prolongation and restriction operators between possibly different variable spaces, allows more than two nested levels of fidelity, and, maybe less importantly, does not require coherence of low-fidelity model values with the high-fidelity objective function (zeroth-order model management). On the other hand, Algorithm RMTR does not fit in the framework of [16] because this latter formalism considers only "memoryless" iterations and therefore does not cover adaptive algorithmic features such as the trust-region radius. Moreover, the convergence results analyzed in this reference require nonlocal properties on the involved functions and the limit points are proved only to belong to a set containing the problem's critical points and the iteration fixed points. Finally, the proposal by Borzi and Kunisch [9] differs from ours in that it emphasizes convergence to minimizers on the coarsest grid but does not directly consider globalization on finer ones.

3. A practical algorithm and some numerical motivation. Clearly, our algorithmic description so far leaves a number of practical choices unspecified and is best viewed at this stage as a theoretical shell which potentially contains both efficient and inefficient algorithms. Can efficient algorithms be found in this shell? It is the purpose of this section to show that this is indeed the case. Instead of considering the RMTR class in its full generality, we will therefore focus on a simple implementation of our framework, and then show that the resulting method is, in our view, numerically promising.

3.1. Algorithm definition.

Smoothing and Taylor iterations. The most important of the open algorithmic questions is how one enforces sufficient decrease at Taylor iterations. A first answer is provided by existing algorithms for large-scale optimization, such as truncated conjugate-gradient (TCG) [37, 38] or generalized Lanczos trust-region (GLTR) [19] methods, in which the problem of minimizing (2.7) is solved in successive embedded Krylov subspaces (see also section 7.5 in [13]). This method is known to ensure (2.9). While it can be viewed as a Ritz procedure where solutions of subproblems of increasing sizes approach the desired high-dimensional one, the definition of these embedded subspaces does not exploit the explicit knowledge of discretization grids. We are thus interested in alternatives that exploit this knowledge.

If the model (2.7) is strictly convex and the trust-region radius Δ_k sufficiently large, minimizing (2.7) amounts to an (approximate) solution of the classical Newton equations $H_{i,k}s_i = -g_{i,k}$. If the problem additionally results from discretizing a convex operator on successively finer grids, then multigrid solvers constitute a most interesting alternative. Our intention is not to review this vast class of numerical algorithms here (we refer the reader to [12] for an excellent introduction to the field), but we briefly outline their main characteristics. Multigrid algorithms are based on three complementary observations. The first is that some algorithms, called *smoothers*, are very efficient at selectively reducing the high-frequency components of the error on a grid, that is (in most cases), components whose "wavelength" is comparable to the grid's mesh-size. The second is that a low-frequency error component on a fine grid

appears more oscillatory on a coarse grid and may thus be viewed as a high-frequency component on this grid. The third is that computations on coarse grids are typically much cheaper than on finer grids. These observations may be exploited by a two-grid procedure, as follows. A few iterations of a smoother are first applied on the fine grid, reducing the error's high frequencies. The residual is then projected on the coarse grid where the low frequencies are more oscillatory and thus efficiently and cheaply reduced by the smoother applied on the coarse grid. The remaining error on the coarse grid is then prolongated back to the fine grid, which reintroduces a small amount of high-frequency error. A few more steps of the fine-grid smoother are finally applied to eliminate it. The multigrid algorithm is obtained by recursively replacing the error smoothing on the coarse grid by another two-grid procedure. Multigrid methods for positive-definite systems of equations typically result in remarkably efficient linearly convergent processes. Our intention here is to exploit the same features in minimizing (2.7), although it is expected only to reduce to a positive-definite system of linear equations asymptotically, when a minimizer of the problem is approached.

At the coarsest level, where further recursion is impossible, the cost of exactly minimizing (2.7) within the trust region remains small, because of the low dimensionality of the subproblem. Our strategy is thus to solve it using the method by Moré and Sorensen [29] (see also section 7.3 in [13]), whose very acceptable cost is then dominated by that of a small number of small-scale Cholesky factorizations. At finer levels, we have the choice of using the TCG or GLTR algorithms mentioned above, or an adaptation of the multigrid smoothing techniques that guarantees sufficient descent inside the trust region and also handles the possible nonconvexity of the model. The remainder of this section is devoted to describing this last option.

A very well-known multigrid smoother for systems of equations is the Gauss–Seidel method, in which every individual equation of the Newton system is solved in succession.[3] This procedure can be extended to optimization without major difficulty as follows: instead of successively solving equations, we may perform cyclic successive one-dimensional minimizations along the coordinate axes of the model (2.7), provided the curvature of this model along each axis is positive. Thus, if j is an index such that the jth diagonal entry of $H_{i,k}$ is strictly positive, the updates

$$\alpha_j = -[g]_j/[H_{i,k}]_{jj}, \quad [s]_j \leftarrow [s]_j + \alpha_j, \quad \text{and} \quad g \leftarrow g + \alpha_j H_{i,k} e_{i,j}$$

are performed for the minimization along the jth axis (starting each cycle from s such that $\nabla m_{i,k}(x_{i,k} + s) = g$), where we denote by $[v]_j$ the jth component of the vector v and by $[M]_{ij}$ the (i,j)th entry of the matrix M, and where $e_{i,j}$ is the jth vector of the canonical basis of $\Re^{n_i}$. This is nothing but the well-known (and widely ill-considered) sequential coordinate minimization (see, for instance, [33, section 14.6]), which we abbreviate as SCM. In order to enforce convergence on nonconvex problems to first-order points, we still have to ensure sufficient model decrease (2.9) while keeping the step in the trust region. This can be achieved in various ways, but we choose here to start the SCM cycle by initiating the cycle with the axis corresponding to the largest component of the gradient $g_{i,k}$ in absolute value. Indeed, if this component is the ℓth one and if $d_\ell = -\text{sign}([g_{i,k}]_\ell)e_{i,\ell}$, then minimization of the model $m_{i,k}$ along d_ℓ within the trust region is guaranteed to yield a *Cauchy step* $\alpha_\ell d_\ell$ such that the inequality

$$m_{i,k}(x_{i,k}) - m_{i,k}(x_{i,k} + \alpha_\ell d_\ell) \geq \tfrac{1}{2}|[g_{i,k}]_\ell| \min\left[\frac{|[g_{i,k}]_\ell|}{1 + |[H_{i,k}]_{\ell\ell}|}, \Delta_{i,k}\right] \tag{3.1}$$

[3] See [12, p. 10], or [18, p. 510], or [33, p. 214] amongst many others.

holds. But

$$|[g_{i,k}]_\ell| = \max_j |\,[g_{i,k}]_j| \geq \frac{1}{\sqrt{n}}\|g_{i,k}\|, \quad \text{and} \quad |[H_{i,k}]_{\ell\ell}| \leq \|H_{i,k}\|,$$

and (2.9) then follows from these inequalities and (3.1) since the remaining SCM operations only reduce the value of the model $m_{i,k}$ further. If, after completing one SCM cycle, one then notices that the overall step s lies outside of the trust region, we then apply a variant of the *dogleg* strategy (see [35], or [13, section 7.5.3]) to the step, by minimizing $m_{i,k}$ along the segment $[\alpha_\ell d_\ell, s]$ *restricted to the trust region*. The final step is then given by $\alpha_\ell d_\ell + \alpha_s(s - \alpha_\ell d_\ell)$, where α_s is the multiple of $s - \alpha_\ell d_\ell$ where the minimizer is achieved.

Our description of the smoothing method is complete if we finally specify what is done when negative curvature is encountered along one of the coordinate axes, the jth one, say, during the SCM cycles. In this case, the model minimizer along $e_{i,j}$ lies on the boundary of the trust region, and it is very easy to compute the associated model reduction. The largest of these reductions is remembered (along with the corresponding step) if negative curvature is met along more than one axis. It is then compared to the reduction obtained by minimizing along the axes with positive curvature, and the step is finally chosen as that giving the maximum reduction.

The V-cycles. One of the flexible features of our RMTR framework is that the minimization at lower levels ($i = 1, \ldots, r-1$) can be stopped after the first successful iteration without affecting convergence properties (as will become clear in section 4). This therefore opens the possibility of considering *fixed form* recursion patterns and *free form* ones. A free form pattern is obtained when Algorithm RMTR is run without using the premature termination option, in which case minimization is carried out at each level until the gradient becomes small enough or the relevant trust-region boundary is approached sufficiently (see Step 5 of Algorithm RMTR). The actual recursion pattern is then uniquely determined by the progress of minimization at each level and may be difficult to forecast. By contrast, the fixed form recursion patterns are obtained by specifying a maximum number of successful iterations at each level, a technique directly inspired from the definitions of V- and W-cycles in multigrid algorithms (see [12, p. 40], for instance).

In this section, we consider only V-cycle iterations, where minimization at lower levels (above the coarsest) consists in, at most, one successful smoothing iteration followed by either a successful TCG Taylor iteration (if (2.14) fails) or a recursive iteration (if (2.14) holds), itself followed by a second successful smoothing iteration. The lower iteration is however terminated if the boundary of the upper-level trust region is met, which typically occurs only far from a solution, or if the gradient becomes sufficiently small.

Second-order and Galerkin models. The definition of the gradient correction v_{i-1} in (2.4) is engineered to ensure (2.6), which is to say that h_i and h_{i-1} coincide at first order (up to the constant σ_i) in the range of the prolongation operator. But coherence of the models can also be achieved at second order: if we choose

$$h_{i-1}(x_{i-1,0} + s_{i-1}) = f_{i-1}(x_{i-1,0} + s_{i-1}) + \langle v_{i-1}, s_{i-1}\rangle + \tfrac{1}{2}\langle s_{i-1}, W_{i-1}s_{i-1}\rangle, \tag{3.2}$$

where $W_{i-1} = R_i H_{i,k} P_i - \nabla^2 f_{i-1}(x_{i-1,0})$, then we also have that

$$\langle P_i s_{i-1}, H_{i,k} P_i s_{i-1,}\rangle = \frac{1}{\sigma_i}\langle s_{i-1}, \nabla^2 h_{i-1}(x_{i-1,0}) s_{i-1}\rangle,$$

as desired. An even more radical strategy is to choose $f_{i-1}(x_{i-1,0} + s_{i-1}) = 0$ for all s_{i-1} in (3.2), which amounts to choosing the lower-level objective function as the "restricted" version of the quadratic model at the upper level, also known as the "Galerkin approximation." This technique is known to improve performance for difficult cases involving an underlying infinite-dimensional problem with discontinuous coefficients (see, in particular, the recent analysis in [45]). This is also the option considered in this section. In the case where this model is strictly convex and the trust-region radius is large enough, an iteration of the algorithm reduces to the solution of a positive-definite linear system; multigrid algorithms for solving this system, such as the multigrid V-Cycle scheme of [12, p. 44], can then be viewed as instances of Algorithm RMTR.

Computing the starting point at successively finer levels. It is clear that, if the multilevel recursion idea has any power within an iteration from the finest level down and back, it must also be advantageous to use the lower-level problems for computing the starting point $x_{r,0}$. In our motivating application, we have chosen to compute $x_{r,0}$ by successively minimizing at levels 0 up to $r-1$ starting from the lowest one, where an initial starting point is assumed to be supplied by the user. (Note that, in general, the starting point can be supplied at any discretization level and transferred to other levels by using the prolongations or restrictions.) At level $i < r$, the accuracy on the gradient infinity norm that is required for termination is given by

$$\epsilon_i^{\mathrm{g}} = \min(0.01, \epsilon_{i+1}^{\mathrm{g}}/\nu_i^{\psi}), \tag{3.3}$$

where ψ is the dimension of the underlying continuous problem, ν_i is the discretization mesh-size along one of these dimensions, and ϵ_r^{g} is the user-supplied gradient accuracy requirement for the topmost level. Once computed at level i, the solution is prolongated to level $i+1$ using cubic interpolation.

3.2. Two test examples.

A simple quadratic example. We consider here the two-dimensional model problem for multigrid solvers in the unit square domain S_2,

$$-\Delta u(x,y) = f \quad \text{in } S_2, \qquad\qquad u(x,y) = 0 \quad \text{on } \partial S_2,$$

where f is such that the analytical solution to this problem is

$$u(x,y) = \sin[2\pi x(1-x)]\sin[2\pi y(1-y)].$$

This problem is discretized using a five-point finite-difference scheme, giving a linear system $A_i x = b_i$ at level i, where A_i is a symmetric positive-definite matrix. Algorithm RMTR is used on the variational minimization problem

$$\min_{x \in \Re^{n_r}} \tfrac{1}{2} x^T A_r x - x^T b_r,$$

which is equivalent to the linear system $A_r x = b_r$. The starting point for the values of u not on the boundary is chosen as a random perturbation (of amplitude 10^{-5}) of the vector of all ones. This example illustrates that RMTR exhibits performance similar to traditional linear multigrid solvers on a model problem.

A nonconvex example. We introduce the nonlinear least-squares problem

$$\min_{u,\gamma} \quad \tfrac{1}{1000} \int_{S_2} \gamma(x,y)^2 + \int_{S_2} [u(x,y) - u_0(x,y)]^2 + \int_{S_2} [\Delta u(x,y) - \gamma(x,y) u(x,y)]^2,$$

where the unknown functions $u(x,y)$ and $\gamma(x,y)$ are defined on the unit square S_2 and the function $u_0(x,y)$ is defined on S_2 by $u_0(x,y) = \sin(6\pi x)\sin(2\pi y)$. This problem is again discretized using five-point finite differences, but the square in the last term makes the Hessian denser than for the pure Laplacian. The starting values for u and γ are random perturbations (of amplitude 100) of u_0 and zero, respectively. The nonconvexity of the resulting discretized problem on the fine grid has been assessed by a direct eigenvalue computation on the Hessian of the problem.

Prolongations and restrictions. In both examples, we have defined the prolongation to be the linear interpolation operator and the restriction to be its transpose normalized to ensure that $\|R_i\| = 1$. These operators are never assembled but are applied locally for improved efficiency.

3.3. Numerical results. The algorithm described above has been coded in MATLAB (Release 7.0.0) and the experiments below were run on a Dell Precision M70 laptop computer with 2MBytes of RAM. The test problems are solved with $\epsilon_r^g = 0.5\times10^{-9}$. Smoothing iterations use a single SCM cycle, and we choose $\eta_1 = 0.01$, $\eta_2 = 0.95$, $\gamma_1 = 0.05$, $\gamma_2 = 0.25$, $\kappa_g = 0.5$, and $\epsilon_i^\Delta = 0.001$ for all i. The choice of Δ_r^s, the initial trust-region radius at level r, is slightly more difficult (see, for instance, [34, 36] for suggested strategies), but here we choose to use $\Delta_r^s = 1$. The gradient thresholds ϵ_i^g are chosen according to the rule (3.3).

We consider the simple quadratic example first. In this example, recursive iterations were always accepted by the test (2.14). As a result, the work consisted only in exactly minimizing (2.7) in the trust region at the coarsest level and SCM smoothing at higher levels. Table 3.1 gives the problem dimension (n) for each level and the number of smoothing SCM cycles (# fine SCM) at the finest level required to solve the complete problem from scratch. This is, by far, the dominant linear algebra cost. For completeness, we also report the solution time in seconds (as reported by MATLAB) in the line "CPU(s)" of the same table.

For comparison, we also tested an efficient classical trust-region method using mesh-refinement with cubic interpolation and a TCG solver, where the conjugate-gradient minimization at iteration (i,k) is terminated as soon as the model gradient falls under the threshold

$$\max\left[\min\left(0.1, \sqrt{\|g_{i,k}\|}\right)\|g_{i,k}\|, 0.95\,\epsilon_r^g\right]$$

(see section 7.5.1 of [13], for instance). This algorithm solved the level-7 problem ($n = 261,121$) with 657 conjugate-gradient iterations at the finest level in 190.54 seconds, and solved the level-8 problem ($n = 1,046,529$) with 1,307 conjugate-gradient iterations at the finest level in 2,463.33 seconds. (Note that this TCG solver can also be obtained as a special case of our framework by replacing smoothing iterations by TCG ones and disabling the recursive calls to RMTR.) As expected for a typical multigrid algorithm for linear equations, we observe that the number of smoothing cycles is fairly independent of the mesh size and dimension, which indicates that the trust-region machinery does not alter this property.

TABLE 3.1
Performance on the simple quadratic example.

Level	0	1	2	3	4	5	6	7	8
n	9	49	225	961	3,969	16,129	65,025	261,121	1,046,529
# fine SCM	-	11	11	11	9	8	6	5	3
CPU(s)	-	0.05	0.14	0.37	0.97	2.84	9.4	38.4	150.88

TABLE 3.2
Performance on the nonconvex example.

Level	0	1	2	3	4	5	6	7
n	18	98	450	1,922	7,938	32,258	130,050	522,242
# fine SCM	-	21	19	21	28	32	14	9
CPU(s)	-	0.43	1.05	3.60	14.90	73.63	151.53	560.26

We now consider our nonconvex test problem, for which the same statistics are given in Table 3.2. As for the quadratic example, the test (2.14) was always satisfied and the algorithm thus never had to use TCG iterations for levels above the coarsest.

On this example, the mesh-refinement algorithm using the TCG solver solved the level-6 problem ($n = 130,050$) with 33,033 conjugate-gradient iterations at level 6 in 3,262.06 seconds, and solved the level-7 problem ($n = 522,242$) with 3,926 conjugate-gradient iterations at level 7 in 6,154.96 seconds.

Even if these results were obtained by a very simple implementation of our framework, they are nevertheless highly encouraging, as they suggest that speed-ups of one order of magnitude or more could be obtained over (good) contending methods. Moreover, the statistics presented here also suggest that, at least for not too nonlinear problems, performance in CPU time can be essentially proportional to problem size, a very desirable property. The authors are of course aware that only continued experience with more advanced implementations will vindicate those preliminary tests (this work is currently under way) but consider that the potential numerical benefits justify a sound convergence analysis of the algorithm, which is best carried out considering the general RMTR class. This is the purpose of the next section.

4. Global convergence. Our exposition of the global convergence properties of our general class of recursive multiscale algorithms starts with the analysis of properties that are specific to our class. The main concepts and developments of section 6.4 in [13] are subsequently revisited to conclude the case of the multiscale algorithm. Interestingly, the techniques of proof are different and lead to a new complexity result (Theorem 4.10) that is also valid in the classical single-level case.

We first complete our assumptions by supposing that the Hessians of each h_i and their approximations are bounded above by the constant $\kappa_{\text{H}} \geq 1$, i.e., more formally, that, for $i = 0, \ldots, r$,

$$1 + \|\nabla^2 h_i(x_i)\| \leq \kappa_{\text{H}} \tag{4.1}$$

for all $x_i \in \Re^{n_i}$, and

$$1 + \|H_{i,k}\| \leq \kappa_{\text{H}} \tag{4.2}$$

for all k. In order to keep our notation simple, we also assume, without loss of generality, that

$$\sigma_i = 1 \tag{4.3}$$

in (2.3) for $i = 0, \ldots, r$ (this can be directly obtained from the original form by scaling P_i and/or R_i). We also define the constants

$$\kappa_{\text{PR}} \stackrel{\text{def}}{=} \max\left[1, \max_{i=1,\ldots,r} \|P_i\|\right] = \max\left[1, \max_{i=1,\ldots,r} \|R_i\|\right] \tag{4.4}$$

(where we used (2.3) and (4.3) to deduce the second equality) and

$$\kappa_\sigma \stackrel{\rm def}{=} \min\left[1, \min_{i=0,\ldots,r} \sigma_{\min}(M_i)\right] > 0, \tag{4.5}$$

where $\sigma_{\min}(A)$ denotes the smallest singular value of the matrix A. We finally define

$$\Delta^{\rm s}_{\min} = \min_{i=0,\ldots,r} \Delta^{\rm s}_i, \qquad \epsilon^{\rm g}_{\min} = \min_{i=0,\ldots,r} \epsilon^{\rm g}_i, \ \text{ and } \ \epsilon^{\Delta}_{\min} = \min_{i=0,\ldots,r} \epsilon^{\Delta}_i. \tag{4.6}$$

We also introduce some additional concepts and notation.

1. If iteration (i,k) is recursive, we say that this iteration initiates a *minimization sequence* at level $i-1$, which consists of all successive iterations *at this level* (starting from the point $x_{i-1,0} = R_i x_{i,k}$) until a return is made to level i within iteration (i,k). In this case, we also say that iteration (i,k) is the *predecessor* of the minimization sequence at level $i-1$. If $(i-1,\ell)$ belongs to this minimization sequence, this is written as $(i,k) = \pi(i-1,\ell)$.
2. To a given iteration (i,k), we associate the set

$$\mathcal{R}(i,k) \stackrel{\rm def}{=} \{(j,\ell) \mid \text{iteration } (j,\ell) \text{ occurs within iteration } (i,k)\}. \tag{4.7}$$

   The set $\mathcal{R}(i,k)$ always contains the pair (i,k) and contains only that pair if Step 3 is used at iteration (i,k). If Step 2 is used instead of Step 3, then it additionally contains the pairs of level and iteration numbers of all iterations that occur in the potential recursion started in Step 2 and terminating on return within iteration (i,k). Because $\mathcal{R}(i,k)$ is defined in terms of iterations, it does *not* contain the pairs of indices corresponding to the terminating iterates $(j,*)$ of its (internal) minimization sequences. One easily verifies that $j \le i$ for every j such that $(j,\ell) \in \mathcal{R}(i,k)$ for some nonnegative k and ℓ. The mechanism of the algorithm also ensures that

$$\Delta_{j,\ell} \le \Delta_{i,k} \ \text{ whenever } \ (j,\ell) \in \mathcal{R}(i,k), \tag{4.8}$$

   because of the choice of $\Delta_{j,0}$ in Step 0 and (2.17). Note that $\mathcal{R}(i,k)$ contains at most one minimization sequence at level $i-1$ but may contain more than one at level $i-2$, since each iteration at level $i-1$ may generate its own.
3. For any iteration $(j,\ell) \in \mathcal{R}(i,k)$, there exists a unique *path* from (j,ℓ) to (i,k) defined by taking the predecessor of iteration (j,ℓ), say, $(j+1,q) = \pi(j,\ell)$, and then the predecessor of $(j+1,q)$ and so on until iteration (i,k). We also define

$$d(i,k) = \min_{(j,\ell)\in\mathcal{R}(i,k)} j, \tag{4.9}$$

   which is the index of the deepest level reached by the potential recursion of iteration (i,k). The path from $(d(i,k),\ell)$ to (i,k) is the longest in $\mathcal{R}(i,k)$.
4. We use the symbol

$$\mathcal{T}(i,k) \stackrel{\rm def}{=} \{(j,\ell) \in \mathcal{R}(i,k) \mid \text{iteration } (j,\ell) \text{ uses Step 3}\}$$

   to denote the subset of Taylor iterations in $\mathcal{R}(i,k)$, that is, iterations at which Taylor's model $m_{j,\ell}(x_{j,\ell} + s_j)$ is chosen.

We start the analysis of Algorithm RMTR by proving that it has a central property of trust-region methods, namely, that the steps remain in the trust region.

LEMMA 4.1. *For each iteration (i,k), we have that*

$$\|s_{i,k}\|_i \leq \Delta_{i,k}. \tag{4.10}$$

Moreover, if $\Delta_{j+1,q}$ is the trust-region radius of iteration $(j+1,q) = \pi(j,\ell)$, we have that, for each $(j,\ell) \in \mathcal{R}(i,k)$,

$$\|x_{j,\ell} - x_{j,0}\|_j \leq \Delta_{j+1,q} \quad \textit{and} \quad \|x_{j,*} - x_{j,0}\|_j \leq \Delta_{j+1,q}. \tag{4.11}$$

Proof. The constraint (4.10) is explicit for Taylor iterations. We therefore have to verify that it holds only if Step 2 is chosen at iteration (i,k). If this is the case, consider $j = d(i,k)$, and consider the first time it occurs in $\mathcal{R}(i,k)$. Assume, furthermore, that $x_{j,*} = x_{j,p}$. Because no recursion occurs to a level lower than j, one must have (from Step 3) that

$$\|s_{j,\ell}\|_j \leq \Delta_{j,\ell} \quad (\ell = 0, \ldots, p-1). \tag{4.12}$$

Then we obtain, for $\ell = 1, \ldots, p$, that, if iteration $(j,\ell-1)$ is successful,

$$\|x_{j,\ell} - x_{j,0}\|_j = \|x_{j,\ell-1} - x_{j,0} + s_{j,\ell-1}\|_j \leq \|x_{j,\ell-1} - x_{j,0}\|_j + \|s_{j,\ell-1}\|_j$$

because of the triangle inequality, while

$$\|x_{j,\ell} - x_{j,0}\|_j = \|x_{j,\ell-1} - x_{j,0}\|_j \leq \|x_{j,\ell-1} - x_{j,0}\|_j + \|s_{j,\ell-1}\|_j$$

if it is unsuccessful. Combining these two bounds and (4.12), we have that

$$\begin{aligned} \|x_{j,\ell} - x_{j,0}\|_j &\leq \|x_{j,\ell-1} - x_{j,0}\|_j + \Delta_{j,\ell-1} \\ &\leq \|x_{j,\ell-1} - x_{j,0}\|_j + \Delta_{j+1,q} - \|x_{j,\ell-1} - x_{j,0}\|_j \\ &= \Delta_{j+1,q} \end{aligned} \tag{4.13}$$

for $\ell = 2, \ldots, p$, where the last inequality results from (2.17). The same result also holds for $\ell = 1$, since $\|x_{j,1} - x_{j,0}\|_j \leq \Delta_{j,0} \leq \Delta_{j+1,q}$ because of Step 0 in the algorithm. We then verify, using (2.10), that

$$\|s_{j+1,q}\|_{j+1} = \|P_{j+1}\,(x_{j,*} - x_{j,0})\,\|_{j+1} = \|x_{j,*} - x_{j,0}\|_j = \|x_{j,p} - x_{j,0}\|_j \leq \Delta_{j+1,q},$$

which is nothing but inequality (4.12) at iteration $(j+1,q)$. The same reasoning may then be applied to each iteration at level $j+1$ that uses Step 2. Since inequality (4.12) is guaranteed for all other iterations of that level by Step 3, we obtain that (4.12) also holds with j replaced by $j+1$. The same must therefore be true for (4.13). The induction can then be continued up to level i, yielding both (4.10) and (4.11) (for which the case $\ell = 0$ is obvious). ∎

In the same vein, the algorithm also ensures the following two properties.

LEMMA 4.2. *The mechanism of Algorithm RMTR guarantees that, for each iterate of index (j,ℓ) such that $(j,\ell) \neq (j,*)$ (i.e., for all iterates at level j but the last one),*

$$\|g_{j,\ell}\| > \epsilon_j^g \tag{4.14}$$

and

$$\|x_{j,\ell} - x_{j,0}\|_j \leq (1 - \epsilon_j^{\Delta})\Delta_{j+1,q}, \tag{4.15}$$

where $\Delta_{j+1,q}$ *is the trust-region radius of iteration* $(j+1,q) = \pi(j,\ell)$.

Proof. These bounds directly follow from the stopping criteria for minimization at level j, in Step 5 of the algorithm. ∎

We now prove some useful bounds on the gradient norms for all iterates that belong to a recursion process initiated within a sufficiently small trust region.

LEMMA 4.3. *Assume that, for some iteration* (i,k),

$$\Delta_{i,k} \leq \frac{\sqrt{\kappa_\sigma}\kappa_{\rm g}^r}{2r\kappa_{\rm H}}\|g_{i,k}\| \stackrel{\rm def}{=} \kappa_1\|g_{i,k}\|, \tag{4.16}$$

where $\kappa_1 \in (0,1)$. *Then one has that, for all* $(j,\ell) \in \mathcal{R}(i,k)$,

$$\tfrac{1}{2}\kappa_{\rm g}^r\|g_{i,k}\| \leq \|g_{j,\ell}\| \leq \kappa_{\rm PR}^r(1 + \tfrac{1}{2}\kappa_{\rm g}^r)\|g_{i,k}\|. \tag{4.17}$$

Proof. The result is obvious for $(j,\ell) = (i,k)$ since, by definition, $\kappa_{\rm g} < 1$ and $\kappa_{\rm PR} \geq 1$. Let us now consider some iteration $(j,\ell) \in \mathcal{R}(i,k)$ with $j < i$. From the mean-value theorem, we know that, for any iteration (j,ℓ),

$$g_{j,\ell} = g_{j,0} + G_{j,\ell}(x_{j,\ell} - x_{j,0}), \tag{4.18}$$

where

$$G_{j,\ell} = \int_0^1 \nabla^2 h_j(x_{j,0} + t(x_{j,\ell} - x_{j,0}))\, dt. \tag{4.19}$$

But

$$\|G_{j,\ell}\| \leq \max_{t\in[0,1]} \|\nabla^2 h_j(x_{j,0} + t(x_{j,\ell} - x_{j,0}))\| \leq \kappa_{\rm H}, \tag{4.20}$$

and hence, by definition of the norms and (4.5),

$$\|g_{j,\ell}\| \geq \|g_{j,0}\| - \kappa_{\rm H}\|x_{j,\ell} - x_{j,0}\| \geq \|g_{j,0}\| - \frac{\kappa_{\rm H}}{\sqrt{\kappa_\sigma}}\|x_{j,\ell} - x_{j,0}\|_j \tag{4.21}$$

for all (j,ℓ). On the other hand, if $(j+1,q) = \pi(j,\ell)$, we have also that, for all $(j,\ell) \in \mathcal{R}(i,k)$,

$$\|x_{j,\ell} - x_{j,0}\|_j \leq \Delta_{j+1,q} \leq \Delta_{i,k} \tag{4.22}$$

because of (4.11) and (4.8) (as $(j+1,q) \in \mathcal{R}(i,k)$). Combining (4.21) and (4.22), we obtain that, for all $(j,\ell) \in \mathcal{R}(i,k)$,

$$\|g_{j,\ell}\| \geq \|g_{j,0}\| - \frac{\kappa_{\rm H}}{\sqrt{\kappa_\sigma}}\Delta_{i,k}. \tag{4.23}$$

Consider now the path from (j,ℓ) to (i,k) in $\mathcal{R}(i,k)$. Let this path consist of the iterations (j,ℓ), $(j+u,t_{j+u})$ for $u = 1,\ldots,i-j-1$, and (i,k). We then have that

$$\begin{aligned}
\|g_{j,\ell}\| &\geq \|g_{j,0}\| - \tfrac{\kappa_{\rm H}}{\sqrt{\kappa_\sigma}}\Delta_{i,k} \geq \kappa_{\rm g}\|g_{j+1,t_{j+1}}\| - \tfrac{\kappa_{\rm H}}{\sqrt{\kappa_\sigma}}\Delta_{i,k} \\
&\geq \kappa_{\rm g}\|g_{j+1,0}\| - 2\tfrac{\kappa_{\rm H}}{\sqrt{\kappa_\sigma}}\Delta_{i,k} \geq \kappa_{\rm g}^2\|g_{j+2,t_{j+2}}\| - 2\tfrac{\kappa_{\rm H}}{\sqrt{\kappa_\sigma}}\Delta_{i,k} \\
&\geq \kappa_{\rm g}^r\|g_{i,k}\| - r\tfrac{\kappa_{\rm H}}{\sqrt{\kappa_\sigma}}\Delta_{i,k},
\end{aligned}$$

where we successively used (4.23), (2.5), the first part of (2.14), and the inequality $\kappa_{\rm g} < 1$. We then deduce the first inequality of (4.17) from (4.16).

To prove the second, we reuse (4.18)–(4.20) to obtain that

$$\|g_{j,\ell}\| \le \|g_{j,0}\| + \kappa_{\rm H}\|x_{j,\ell} - x_{j,0}\| \le \|g_{j,0}\| + \frac{\kappa_{\rm H}}{\sqrt{\kappa_\sigma}}\|x_{j,\ell} - x_{j,0}\|_j. \tag{4.24}$$

Combining this with (4.22), we conclude that

$$\|g_{j,\ell}\| \le \|g_{j,0}\| + \frac{\kappa_{\rm H}}{\sqrt{\kappa_\sigma}}\Delta_{i,k}. \tag{4.25}$$

We now retrack the iteration path from (j,ℓ) back to (i,k) as above and successively deduce from (4.25), (2.5), and (4.4) that

$$\begin{aligned}
\|g_{j,\ell}\| &\le \|g_{j,0}\| + \tfrac{\kappa_{\rm H}}{\sqrt{\kappa_\sigma}}\Delta_{i,k} \le \kappa_{\rm PR}\|g_{j+1,t_{j+1}}\| + \tfrac{\kappa_{\rm H}}{\sqrt{\kappa_\sigma}}\Delta_{i,k} \\
&\le \kappa_{\rm PR}\|g_{j+1,0}\| + (\kappa_{\rm PR}+1)\tfrac{\kappa_{\rm H}}{\sqrt{\kappa_\sigma}}\Delta_{i,k} \le \kappa_{\rm PR}^2\|g_{j+2,t_{j+2}}\| + 2\tfrac{\kappa_{\rm PR}\kappa_{\rm H}}{\sqrt{\kappa_\sigma}}\Delta_{i,k} \\
&\le \kappa_{\rm PR}^r\|g_{i,k}\| + r\tfrac{\kappa_{\rm PR}^{r-1}\kappa_{\rm H}}{\sqrt{\kappa_\sigma}}\Delta_{i,k} \le \kappa_{\rm PR}^r\left[\|g_{i,k}\| + r\tfrac{\kappa_{\rm H}}{\sqrt{\kappa_\sigma}}\Delta_{i,k}\right],
\end{aligned}$$

using $\kappa_{\rm PR} \ge 1$. We may now use the bound (4.16) to conclude that the second inequality of (4.17) must hold. ∎

We now investigate what happens at noncritical points if the trust-region radius $\Delta_{i,k}$ is small enough. This investigation is conducted by considering the subset $\mathcal{V}(i,k)$ of $\mathcal{R}(i,k)$ defined by

$$\mathcal{V}(i,k) = \left\{(j,\ell) \in \mathcal{R}(i,k) \mid \delta_{j,\ell} \ge \tfrac{1}{2}\kappa_{\rm red}\kappa_{\rm g}^r\kappa_\epsilon^{j-d(i,k)}\|g_{i,k}\|\Delta_{j,\ell}\right\}, \tag{4.26}$$

where

$$\kappa_\epsilon \stackrel{\rm def}{=} \eta_2\epsilon_{\rm min}^{\Delta} < 1. \tag{4.27}$$

$\mathcal{V}(i,k)$ is the subset of iterations within the recursion at iteration (i,k) for which the model decrease is bounded below by a (level-dependent) factor times the product of the gradient norm $\|g_{i,k}\|$ and the trust-region radius $\Delta_{j,\ell}$. Note that, if iteration (j,ℓ) belongs to $\mathcal{V}(i,k)$, this implies that $\delta_{j,\ell}$ can be computed in a finite number of iterations, and thus that $\mathcal{R}(j,\ell)$ is finite. The idea of the next two results is to show that $\mathcal{V}(i,k)$ and $\mathcal{R}(i,k)$ coincide for a sufficiently small radius $\Delta_{i,k}$.

THEOREM 4.4. *Consider an iteration (i,k) for which $\|g_{i,k}\| > 0$ and*

$$\begin{aligned}
\Delta_{i,k} &\le \min\left[\Delta_{\rm min}^s, \min\left(\kappa_1, \frac{\kappa_{\rm red}\kappa_\sigma\kappa_{\rm g}^r\kappa_\epsilon^r(1-\eta_2)}{2\kappa_{\rm H}}\right)\|g_{i,k}\|\right] \\
&\stackrel{\rm def}{=} \min[\Delta_{\rm min}^s, \kappa_2\|g_{i,k}\|],
\end{aligned} \tag{4.28}$$

where $\kappa_2 \in (0,1)$. Then the following conclusions hold:

1. *every iteration using Taylor's model belongs to* (4.26)*, that is,*

$$\mathcal{T}(i,k) \subseteq \mathcal{V}(i,k), \text{ and} \tag{4.29}$$

2. *iteration (j,ℓ) is very successful for every $(j,\ell) \in \mathcal{V}(i,k)$.*

Moreover, if all iterations (j,ℓ) of a minimization sequence at level $j < i$ belong to $\mathcal{V}(i,k)$ and if $\pi(j,\ell) = (j+1,q)$, then

3. *the decrease in the objective function at level j satisfies, for each $\ell > 0$,*

$$h_j(x_{j,0}) - h_j(x_{j,\ell}) \geq \tfrac{1}{2}\kappa_{\rm red}\kappa_{\rm g}^r\kappa_\epsilon^{j-d(i,k)+1}\ell\,\|g_{i,k}\|\Delta_{j+1,q}, \tag{4.30}$$

4. *there are at most*

$$p_* \stackrel{\rm def}{=} \left\lceil \frac{\kappa_{\rm PR}^r\sqrt{\kappa_\sigma}(2+\kappa_{\rm g}^r)+\kappa_2\kappa_{\rm H}}{\kappa_{\rm red}\kappa_\sigma\kappa_{\rm g}^r\kappa_\epsilon^r} \right\rceil \tag{4.31}$$

iterations in the minimization sequence at level j, and

5. *we have that*

$$(j+1,q) \in \mathcal{V}(i,k). \tag{4.32}$$

Proof. 1. We start by proving (4.29). Note that, for $(j,\ell) \in \mathcal{R}(i,k)$, (4.8), the fact that the positive constants $\kappa_{\rm red}$, κ_σ, κ_ϵ, and η_2 are all bounded above by one, (4.28), the left inequality in (4.17), and (4.2) allow us to conclude that

$$\Delta_{j,\ell} \leq \Delta_{i,k} \leq \frac{\kappa_{\rm g}^r}{2\kappa_{\rm H}}\|g_{i,k}\| \leq \frac{\|g_{j,\ell}\|}{1+\|H_{j,\ell}\|}. \tag{4.33}$$

If we now assume that $(j,\ell) \in \mathcal{T}(i,k)$, the decrease condition (2.9) must hold at this iteration, which, together with the left part of (4.17) and (4.33), gives that

$$\delta_{j,\ell} = m_{j,\ell}(x_{j,\ell}) - m_{j,\ell}(x_{j,\ell}+s_{j,\ell}) \geq \kappa_{\rm red}\|g_{j,\ell}\|\Delta_{j,\ell} \geq \tfrac{1}{2}\kappa_{\rm red}\kappa_{\rm g}^r\|g_{i,k}\|\Delta_{j,\ell}, \tag{4.34}$$

which then implies (4.29) since $\kappa_\epsilon < 1$.

2. We prove item 2 separately for $(j,\ell) \in \mathcal{T}(i,k)$ and for $(j,\ell) \in \mathcal{V}(i,k) \setminus \mathcal{T}(i,k)$. Consider the case where $(j,\ell) \in \mathcal{T}(i,k)$ first. We deduce from Taylor's theorem that, for $(j,\ell) \in \mathcal{T}(i,k)$,

$$|h_j(x_{j,\ell}+s_{j,\ell}) - m_{j,\ell}(x_{j,\ell}+s_{j,\ell})| \leq \kappa_{\rm H}\left(\frac{\|s_{j,\ell}\|}{\|s_{j,\ell}\|_j}\right)^2\Delta_{j,\ell}^2 \tag{4.35}$$

(see, for instance, Theorem 6.4.1 on p. 133 of [13]). But, by definition of the norms and (4.5), we know that $\|s_{j,\ell}\|_j \geq \sqrt{\kappa_\sigma}\|s_{j,\ell}\|$. Hence, (4.35) becomes

$$|h_j(x_{j,\ell}+s_{j,\ell}) - m_{j,\ell}(x_{j,\ell}+s_{j,\ell})| \leq \frac{\kappa_{\rm H}}{\kappa_\sigma}\Delta_{j,\ell}^2.$$

Combining this last bound with (4.34), we obtain from (2.15) that

$$|\rho_{j,\ell}-1| \leq \left|\frac{h_j(x_{j,\ell}+s_{j,\ell}) - m_{j,\ell}(x_{j,\ell}+s_{j,\ell})}{m_{j,\ell}(x_{j,\ell}) - m_{j,\ell}(x_{j,\ell}+s_{j,\ell})}\right| \leq \frac{2\kappa_{\rm H}}{\kappa_{\rm red}\kappa_\sigma\kappa_{\rm g}^r\|g_{i,k}\|}\Delta_{j,\ell} \leq 1-\eta_2,$$

where the last inequality is deduced from (4.8) and the fact that (4.28) implies the bound

$$\Delta_{i,k} \leq \kappa_{\rm red}\kappa_\sigma\kappa_{\rm g}^r\|g_{i,k}\|(1-\eta_2)/2\kappa_{\rm H}$$

since $\kappa_\epsilon < 1$. Hence $\rho_{j,\ell} \geq \eta_2$ and iteration $(j,\ell) \in \mathcal{T}(i,k)$ is very successful, as requested in item 2.

We next prove item 2 for $(j,\ell) \in \mathcal{V}(i,k) \setminus \mathcal{T}(i,k)$, which implies, in particular, that $\mathcal{R}(j,\ell)$ is finite and $x_{j-1,*}$ is well defined. If we consider iteration (j,ℓ), we may still deduce from the mean-value theorem that

$$h_j(x_{j,\ell}) - h_j(x_{j,\ell} + s_{j,\ell}) = -\langle g_{j,\ell}, s_{j,\ell}\rangle - \tfrac{1}{2}\langle s_{j,\ell}, \nabla^2 h_j(\xi_j)\, s_{j,\ell}\rangle$$

for some $\xi_j \in [x_{j,\ell}, x_{j,\ell} + s_{j,\ell}]$ and also that

$$h_{j-1}(x_{j-1,0}) - h_{j-1}(x_{j-1,*}) = -\langle g_{j-1,0}, z_{j-1}\rangle - \tfrac{1}{2}\langle z_{j-1}, \nabla^2 h_{j-1}(\xi_{j-1})\, z_{j-1}\rangle$$

for some $\xi_{j-1} \in [x_{j-1,0}, x_{j-1,0} + z_{j-1}]$, where $z_{j-1} = x_{j-1,*} - x_{j-1,0} = x_{j-1,*} - R_j x_{j,\ell}$. Now, because $s_{j,\ell} = P_j z_{j-1}$, we deduce from (2.6) and (4.3) that $\langle g_{j,\ell}, s_{j,\ell}\rangle = \langle g_{j-1,0}, z_{j-1}\rangle$ and therefore that

$$\begin{aligned} h_j(x_{j,\ell}) - h_j(x_{j,\ell} + s_{j,\ell}) &= h_{j-1}(x_{j-1,0}) - h_{j-1}(x_{j-1,*}) \\ &\quad - \tfrac{1}{2}\langle s_{j,\ell}, \nabla^2 h_j(\xi_j)\, s_{j,\ell}\rangle \\ &\quad + \tfrac{1}{2}\langle z_{j-1}, \nabla^2 h_{j-1}(\xi_{j-1})\, z_{j-1}\rangle. \end{aligned} \tag{4.36}$$

But Lemma 4.1 implies that $\|s_{j,\ell}\|_j \le \Delta_{j,\ell}$ and $\|z_{j-1}\|_{j-1} \le \Delta_{j,\ell}$, which, in turn with the Cauchy–Schwarz inequality, gives that

$$|\langle s_{j,\ell}, \nabla^2 h_j(\xi_j)\, s_{j,\ell}\rangle| \le \kappa_{\mathrm{H}} \|s_{j,\ell}\|^2 \le \kappa_{\mathrm{H}} \left(\frac{\|s_{j,\ell}\|}{\|s_{j,\ell}\|_j}\right)^2 \Delta_{j,\ell}^2 \le \frac{\kappa_{\mathrm{H}}}{\kappa_\sigma} \Delta_{j,\ell}^2. \tag{4.37}$$

Similarly,

$$|\langle z_{j-1}, \nabla^2 h_{j-1}(\xi_{j-1})\, z_{j-1}\rangle| \le \frac{\kappa_{\mathrm{H}}}{\kappa_\sigma} \Delta_{j,\ell}^2. \tag{4.38}$$

Combining (4.36), (4.37), (4.38), and the definition of $\delta_{j,\ell}$, we obtain that

$$h_j(x_{j,\ell}) - h_j(x_{j,\ell} + s_{j,\ell}) \ge \delta_{j,\ell} - \frac{\kappa_{\mathrm{H}}}{\kappa_\sigma} \Delta_{j,\ell}^2. \tag{4.39}$$

But since $(j,\ell) \in \mathcal{V}(i,k)$ and $\kappa_\epsilon < 1$, we have that

$$\delta_{j,\ell} \ge \tfrac{1}{2}\kappa_{\mathrm{red}} \kappa_{\mathrm{g}}^r \kappa_\epsilon^{j-d(i,k)} \|g_{i,k}\| \Delta_{j,\ell} \ge \tfrac{1}{2}\kappa_{\mathrm{red}} \kappa_{\mathrm{g}}^r \kappa_\epsilon^r \|g_{i,k}\| \Delta_{j,\ell} > 0,$$

and we conclude from (4.39), the definition of $\rho_{j,\ell}$, and this last bound that

$$\rho_{j,\ell} = \frac{h_j(x_{j,\ell}) - h_j(x_{j,\ell} + s_{j,\ell})}{\delta_{j,\ell}} \ge 1 - \frac{\kappa_{\mathrm{H}} \Delta_{j,\ell}^2}{\kappa_\sigma \delta_{j,\ell}} \ge 1 - \frac{2\kappa_{\mathrm{H}} \Delta_{j,\ell}}{\kappa_{\mathrm{red}} \kappa_\sigma \kappa_{\mathrm{g}}^r \kappa_\epsilon^r \|g_{i,k}\|}.$$

Noting now that (4.28) implies the inequality

$$\Delta_{i,k} \le \tfrac{1}{2}\kappa_{\mathrm{red}} \kappa_\sigma \kappa_{\mathrm{g}}^r \kappa_\epsilon^r \|g_{i,k}\| (1 - \eta_2)$$

and using the bound (4.8), we obtain that $\rho_{j,\ell} \ge \eta_2$. Iteration (j,ℓ) is thus very successful, which completes the proof of item 2.

3. We now assume that all iterations (j,ℓ) of a minimization sequence at level $j < i$ belong to $\mathcal{V}(i,k)$ with $(j+1,q) = \pi(j,\ell)$. We first notice that $(j+1,q) \in \mathcal{R}(i,k)$, (4.8), (4.28), and (4.6) imply that $\Delta_{j+1,q} \le \Delta_{i,k} \le \Delta_{\min}^{\mathrm{s}} \le \Delta_j^{\mathrm{s}}$. Hence Step 0 gives

that $\Delta_{j,0} = \Delta_{j+1,q}$, and since all iterations at level j are very successful because of item 2, we have from Step 6 that, for all (j,ℓ) with $\ell > 0$,

$$\begin{aligned}
\Delta_{j,\ell} &= \min\left[\Delta^+_{j,\ell-1},\, \Delta_{j+1,q} - \|x_{j,\ell} - x_{j,0}\|_j\right] \\
&\geq \min\left[\Delta_{j,\ell-1},\, \Delta_{j+1,q} - \|x_{j,\ell} - x_{j,0}\|_j\right] \\
&= \min\left[\min[\Delta^+_{j,\ell-2}, \Delta_{j+1,q} - \|x_{j,\ell-1} - x_{j,0}\|_j],\, \Delta_{j+1,q} - \|x_{j,\ell} - x_{j,0}\|_j\right] \\
&\geq \min\left[\Delta_{j,\ell-2},\, \Delta_{j+1,q} - \max_{p=\ell-1,\ell} \|x_{j,p} - x_{j,0}\|_j\right] \\
&\geq \min\left[\Delta_{j,0},\, \Delta_{j+1,q} - \max_{p=1,\ldots,\ell} \|x_{j,p} - x_{j,0}\|_j\right] \\
&= \Delta_{j+1,q} - \max_{p=1,\ldots,\ell} \|x_{j,p} - x_{j,0}\|_j \\
&\geq \epsilon^{\Delta}_j \Delta_{j+1,q},
\end{aligned}$$

where we used (4.15) to deduce the last inequality. Note that $\Delta_{j,0} = \Delta_{j+1,q} > \epsilon^{\Delta}_j \Delta_{j+1,q}$, covering the case where $\ell = 0$. Combining these bounds with the very successful nature of each iteration at level j, we obtain that, for each (j,p) with $p = 0, \ldots, \ell - 1$,

$$\begin{aligned}
h_j(x_{j,p}) - h_j(x_{j,p} + s_{j,p}) &\geq \eta_2 \delta_{j,p} \\
&\geq \tfrac{1}{2}\eta_2 \kappa_{\rm red} \kappa_{\rm g}^r \kappa_\epsilon^{j-d(i,k)} \|g_{i,k}\| \Delta_{j,p} \\
&\geq \tfrac{1}{2}\kappa_{\rm red} \kappa_{\rm g}^r \kappa_\epsilon^{j-d(i,k)} \eta_2 \epsilon^{\Delta}_j \|g_{i,k}\| \Delta_{j+1,q} \\
&\geq \tfrac{1}{2}\kappa_{\rm red} \kappa_{\rm g}^r \kappa_\epsilon^{j-d(i,k)+1} \|g_{i,k}\| \Delta_{j+1,q},
\end{aligned}$$

where we used (4.6) and (4.27) to obtain the last inequality. Summing now over iterations $p = 0, \ldots, \ell - 1$ at level j, we obtain that

$$\begin{aligned}
h_j(x_{j,0}) - h_j(x_{j,\ell}) &= \sum_{p=0}^{\ell-1} [h_j(x_{j,p}) - h_j(x_{j,p} + s_{j,p})] \\
&\geq \tfrac{1}{2}\kappa_{\rm red} \kappa_{\rm g}^r \kappa_\epsilon^{j-d(i,k)+1} \ell \, \|g_{i,k}\| \Delta_{j+1,q},
\end{aligned}$$

yielding (4.30).

4. In order to prove item 4, we start by proving that the total decrease in h_j (the objective function for the considered minimization sequence at the jth level) is bounded above by some multiple of $\|g_{i,k}\|$ and $\Delta_{j+1,q}$. We first note that the mean-value theorem gives that

$$h_j(x_{j,0} + s_{j,\min}) = h_j(x_{j,0}) + \langle g_{j,0}, s_{j,\min}\rangle + \tfrac{1}{2}\langle s_{j,\min}, \nabla^2 h_j(\xi_j)\, s_{j,\min}\rangle$$

for some $\xi_j \in [x_{j,0}, x_{j,0} + s_{j,\min}]$, where we have defined

$$s_{j,\min} = \arg \min_{\|s_j\|_j \leq \Delta_{j+1,q}} h_j(x_{j,0} + s_j).$$

Hence, we obtain that, for all s_j such that $\|s_j\|_j \leq \Delta_{j+1,q}$,

$$h_j(x_{j,0}) - h_j(x_{j,0} + s_j) \leq h_j(x_{j,0}) - h_j(x_{j,0} + s_{j,\min}) \leq \frac{\|g_{j,0}\|}{\sqrt{\kappa_\sigma}} \Delta_{j+1,q} + \frac{\kappa_{\rm H}}{2\kappa_\sigma} \Delta^2_{j+1,q}.$$

But we have that $\|x_{j,\ell} - x_{j,0}\|_j \leq \Delta_{j+1,q}$ because of (4.11), and therefore the right inequalities of (4.17), (4.8), and (4.28) now give that

$$h_j(x_{j,0}) - h_j(x_{j,\ell}) \leq \left[\frac{\kappa_{\rm PR}^r + \frac{1}{2}\kappa_{\rm PR}^r \kappa_{\rm g}^r}{\sqrt{\kappa_\sigma}} + \frac{\kappa_2 \kappa_{\rm H}}{2\kappa_\sigma} \right] \|g_{i,k}\| \Delta_{j+1,q} \tag{4.40}$$

for all (j,ℓ) with $\ell \geq 0$. Combining now this bound with (4.30) and remembering that $\kappa_\epsilon < 1$, we deduce that item 4 must hold with (4.31).

5. Finally, since the minimization sequence at level j is guaranteed to terminate after a finite number of iterations $1 \leq \ell \leq p_*$, we deduce from (4.30) and the definition of $\delta_{j+1,q}$ that

$$\delta_{j+1,q} \geq \tfrac{1}{2}\kappa_{\rm red}\kappa_{\rm g}^r \kappa_\epsilon^{j+1-d(i,k)} \|g_{i,k}\| \Delta_{j+1,q},$$

and (4.32) then immediately follows. ∎

We may deduce the following important corollary from this theorem.

COROLLARY 4.5. *Assume* (4.28) *holds for some iteration* (i,k) *for which* $\|g_{i,k}\| > 0$. *Then all iterations* $(j,\ell) \in \mathcal{R}(i,k)$ *are very successful. Moreover, the total number of iterations in* $\mathcal{R}(i,k)$ *is finite and* $\Delta_{i,k}^+ \geq \Delta_{i,k}$.

Proof. As suggested above, we proceed by showing that $\mathcal{V}(i,k) = \mathcal{R}(i,k)$, working from the deepest recursion level upward. Thus consider level $j = d(i,k)$ first. At this level, all iterations (j,ℓ) belong to $\mathcal{T}(i,k)$ and thus, by (4.29), to $\mathcal{V}(i,k)$. If $j = i$, we have achieved our objective. Assume, therefore, that $j < i$ and consider level $j+1$. Using (4.32), we see that all iterations involving a recursion to level j must belong to $\mathcal{V}(i,k)$, while the other (Taylor) iterations again belong to $\mathcal{V}(i,k)$ by (4.29). If $j+1 = i$, we have thus proved that $\mathcal{V}(i,k) = \mathcal{R}(i,k)$. If $j+1 < i$, we may then apply the same reasoning to level $j+2$, and so on, until level i is reached. We may thus conclude that $\mathcal{V}(i,k)$ and $\mathcal{R}(i,k)$ always coincide and, because of item 2 of Theorem 4.4, contain only very successful iterations. Furthermore, using item 4 of Theorem 4.4, we see that the total number of iterations in $\mathcal{R}(i,k)$ is bounded above by

$$\sum_{l=0}^{r} p_*^\ell \leq r p_*^r + 1.$$

Finally, the fact that $\Delta_{i,k}^+ \geq \Delta_{i,k}$ then results from the mechanism of Step 6 of the algorithm and the very successful nature of iteration $(i,k) \in \mathcal{R}(i,k)$. ∎

This last result guarantees the finiteness of the recursion at iteration (i,k) (and thus the finiteness of the computation of $s_{i,k}$) if $\Delta_{i,k}$ is small enough. It also ensures the following useful consequence.

LEMMA 4.6. *Each minimization sequence contains at least one successful iteration.*

Proof. This follows from the fact that unsuccessful iterations cause the trust-region radius to decrease, until (4.28) is eventually satisfied and a (very) successful iteration occurs because of Corollary 4.5. ∎

We now investigate the consequence of the above results on the trust-region radius at each minimization level.

LEMMA 4.7. *For every iteration* (j,ℓ), *with* $j = 0, \ldots, r$ *and* $\ell \geq 0$, *we have that*

$$\Delta_{j,\ell} \geq \gamma_1 \min\left[\Delta_{\min}^s,\, \kappa_2 \epsilon_j^g,\, \epsilon_j^\Delta \Delta_{j+1,q}\right], \tag{4.41}$$

where $(j+1,q) = \pi(j,\ell)$.

Proof. Consider the minimization sequence at level $j \leq r$ initiated from iteration $(j+1, q)$, and assume, for the purpose of obtaining a contradiction, that iteration (j, ℓ) is the first such that

$$\Delta_{j,\ell} < \gamma_1 \min\left[\Delta^{\rm s}_{\rm min},\, \kappa_2 \epsilon^{\rm g}_j,\, \epsilon^{\Delta}_j \Delta_{j+1,q}\right]. \tag{4.42}$$

Note that, because $\epsilon^{\Delta}_j < 1$ and $\gamma_1 < 1$,

$$\Delta_{j,0} = \min[\Delta^{\rm s}_j,\, \Delta_{j+1,q}] \geq \min[\Delta^{\rm s}_{\rm min},\, \epsilon^{\Delta}_j \Delta_{j+1,q}] > \gamma_1 \min\left[\Delta^{\rm s}_{\rm min},\, \kappa_2 \epsilon^{\rm g}_j,\, \epsilon^{\Delta}_j \Delta_{j+1,q}\right],$$

which ensures that $\ell > 0$ and hence that $\Delta_{j,\ell}$ is computed by applying Step 6 of the algorithm at iteration $(j, \ell-1)$. Suppose now that

$$\Delta_{j,\ell} = \Delta_{j+1,q} - \|x_{j,\ell} - x_{j,0}\|_j; \tag{4.43}$$

i.e., the second term is active in (2.17). Our definition of $\Delta_{r+1,0} = \infty$ and (4.42) then ensure that $j < r$. Then, using (4.15), the definition of γ_1, and (4.42), we deduce that, for $j < r$,

$$\Delta_{j,\ell} \geq \Delta_{j+1,q} - (1 - \epsilon^{\Delta}_j)\Delta_{j+1,q} = \epsilon^{\Delta}_j \Delta_{j+1,q} > \gamma_1 \epsilon^{\Delta}_j \Delta_{j+1,q} > \Delta_{j,\ell},$$

which is impossible. Hence (4.43) cannot hold, and we obtain from (2.17) that $\Delta_{j,\ell} = \Delta^+_{j,\ell-1} \geq \gamma_1 \Delta_{j,\ell-1}$, where the last inequality results from (2.16). Combining this bound with (4.42) and (4.14), we deduce that

$$\Delta_{j,\ell-1} \leq \min\left[\Delta^{\rm s}_{\rm min},\, \kappa_2 \epsilon^{\rm g}_j,\, \epsilon^{\Delta}_j \Delta_{j+1,q}\right] \leq \min\left[\Delta^{\rm s}_{\rm min},\, \kappa_2 \|g_{j,\ell-1}\|\right].$$

Hence we may apply Corollary 4.5 and conclude that iteration $(j, \ell-1)$ is very successful and that $\Delta_{j,\ell-1} \leq \Delta^+_{j,\ell-1} = \Delta_{j,\ell}$. As a consequence, iteration (j, ℓ) cannot be the first such that (4.42) holds. This contradiction now implies that (4.42) is impossible, which completes the proof. ∎

Thus trust-region radii are bounded away from zero by a level-dependent factor. We now verify that this factor may be made independent of the level.

THEOREM 4.8. *There exists a constant* $\Delta_{\rm min} \in (0, \min[\Delta^s_{\rm min}, 1])$ *such that*

$$\Delta_{j,\ell} \geq \Delta_{\rm min} \tag{4.44}$$

for every iteration (j, ℓ).

Proof. Observe first that Lemma 4.7 ensures the bound

$$\Delta_{r,k} \geq \gamma_1 \min[\Delta^{\rm s}_{\rm min},\, \kappa_2 \epsilon^{\rm g}_r] \stackrel{\rm def}{=} \gamma_1 \mu \tag{4.45}$$

for all $k \geq 0$, because we have assumed that the call to the uppermost level is made with an infinite trust-region radius. Note that $\mu \in (0, 1)$ because κ_2 and $\epsilon^{\rm g}_r$ both belong to $(0, 1)$. Suppose now that, for some iteration (j, ℓ),

$$\Delta_{j,\ell} < \gamma_1^{r+2} (\epsilon^{\Delta}_{\rm min})^r \mu. \tag{4.46}$$

If $j = r$, this contradicts (4.45); hence $0 \leq j < r$. Lemma 4.7 and the definition of μ in (4.45) then imply that

$$\min[\mu,\, \epsilon^{\Delta}_j \Delta_{j+1,q}] < \gamma_1^{r+1} (\epsilon^{\Delta}_{\rm min})^r \mu,$$

where, as above, iteration $(j+1,q) = \pi(j,\ell)$. If $\min[\mu,\, \epsilon_j^{\Delta}\Delta_{j+1,q}] = \mu$, then $\mu < \gamma_1^{r+1}(\epsilon_{\min}^{\Delta})^r\mu$, which is impossible because $\gamma_1^{r+1}(\epsilon_{\min}^{\Delta})^r < 1$. As a consequence,

$$\epsilon_j^{\Delta}\Delta_{j+1,q} = \min[\mu,\, \epsilon_j^{\Delta}\Delta_{j+1,q}] < \gamma_1^{r+1}(\epsilon_{\min}^{\Delta})^r\mu \le \gamma_1^{r+1}(\epsilon_{\min}^{\Delta})^{r-1}\epsilon_j^{\Delta}\mu,$$

because of (4.6), and hence

$$\Delta_{j+1,q} < \gamma_1^{r+1}(\epsilon_{\min}^{\Delta})^{r-1}\mu.$$

This condition is entirely similar to (4.46) but one level higher. We may therefore repeat the reasoning at levels $j+1,\ldots,r-1$, yielding the bound

$$\Delta_{r,k} < \gamma_1^{r+2-(r-j)}(\epsilon_{\min}^{\Delta})^{r-(r-j)}\mu = \gamma_1^{j+2}(\epsilon_{\min}^{\Delta})^{j}\mu < \gamma_1\mu.$$

But this last inequality contradicts (4.45), and we therefore deduce that (4.46) never holds. This proves (4.44) with

$$\Delta_{\min} \stackrel{\text{def}}{=} \gamma_1^{r+2}(\epsilon_{\min}^{\Delta})^r \min[\Delta_{\min}^{\text{s}},\, \kappa_2\epsilon_r^{\text{g}}], \tag{4.47}$$

and the bounds $\gamma_1 \in (0,1)$, $\epsilon_{\min}^{\Delta} \in (0,1)$, $\kappa_2 \in (0,1)$, and $\epsilon_r^{\text{g}} \in (0,1)$ together imply that $\Delta_{\min} \in (0, \min[\Delta_{\min}^{\text{s}},\, 1])$, as requested. ∎

This result must be compared to Theorem 6.4.3 on p. 135 of [13], keeping (4.14) in mind along with the fact that we have called the uppermost minimization level with some nonzero tolerance ϵ_r^{g}. Also note in (4.47) that $\Delta_{\min}$ is linearly proportional to ϵ_r^{g} for small enough values of this threshold. The next crucial step of our analysis is to show that the algorithm is well defined in that all the recursions are finite.

THEOREM 4.9. *The number of iterations at each level is finite. Moreover, there exists $\kappa_h \in (0,1)$ such that, for every minimization sequence at level $i = 0,\ldots,r$,*

$$h_i(x_{i,0}) - h_i(x_{i,p+1}) \ge \tau_{i,p}\,\eta_1^{i+1}\kappa_h,$$

where $\tau_{i,p}$ is the total number of successful iterations in $\bigcup_{\ell=0}^{p}\mathcal{T}(i,\ell)$.

Proof. We prove the desired result by induction on higher and higher levels from 0 to r. We start by defining $\omega_{i,\ell}$ to be the number of successful iterations in $\mathcal{T}(i,\ell)$, as well as the number of successful iterations in the set $\bigcup_{\ell=0}^{p}\mathcal{T}(i,\ell)$:

$$\tau_{i,p} = \sum_{\ell=0}^{p}\omega_{i,\ell}. \tag{4.48}$$

Note that $\omega_{i,\ell} \ge 1$ if iteration (i,ℓ) is successful.

Consider first an arbitrary minimization sequence at level 0 (if any), and assume, without loss of generality, that it belongs to $\mathcal{R}(r,k)$ for some $k \ge 0$. Every iteration in this minimization sequence must be a Taylor iteration, which means that every successful iteration in the sequence satisfies

$$\begin{aligned} h_0(x_{0,\ell}) - h_0(x_{0,\ell+1}) &\ge \eta_1\kappa_{\text{red}}\epsilon_0^{\text{g}}\min\left[\frac{\epsilon_0^{\text{g}}}{\kappa_{\text{H}}},\, \Delta_{\min}\right] \\ &\ge \omega_{0,\ell}\,\eta_1\kappa_{\text{red}}\epsilon_{\min}^{\text{g}}\min\left[\frac{\epsilon_{\min}^{\text{g}}}{\kappa_{\text{H}}},\, \Delta_{\min}\right], \end{aligned} \tag{4.49}$$

where we have used (2.9), (4.14), (4.2), Theorem 4.8, (4.6), and the fact that $\omega_{0,\ell} = 1$ for every successful iteration $(0,\ell)$ because $\mathcal{T}(0,\ell) = \{(0,\ell)\}$. Since we know from

Lemma 4.6 that there is at least one such iteration for every minimization sequence, we may sum the objective decreases at level 0 and obtain from (4.49) that

$$h_0(x_{0,0}) - h_0(x_{0,p+1}) = \sum_{\ell=0}^{p} {}^{(\mathrm{S})}[h_0(x_{0,\ell}) - h_0(x_{0,\ell+1})] \geq \tau_{0,p}\eta_1\kappa_h, \tag{4.50}$$

where the sum with superscript (S) is restricted to successful iterations and where

$$\kappa_h \stackrel{\mathrm{def}}{=} \kappa_{\mathrm{red}}\epsilon_{\mathrm{min}}^{\mathrm{g}} \min\left[\frac{\epsilon_{\mathrm{min}}^{\mathrm{g}}}{\kappa_{\mathrm{H}}}, \Delta_{\mathrm{min}}\right] \in (0,1). \tag{4.51}$$

If $r = 0$, we know that $h_0 = f$ is bounded below by assumption, and (4.50) implies that $\tau_{0,p}$ must be finite. If $r > 0$, our assumption that f_0 is continuous implies that h_0 is also continuous and hence bounded below on the set $\{x \in \Re^{n_0} \mid \|x - x_{0,0}\|_0 \leq \Delta_{r,k}\}$. The relation (4.50), Lemma 4.1, and (4.8) therefore again impose the finiteness of $\tau_{0,p}$. Since $\tau_{0,p}$ accounts for all successful iterations in the minimization sequence, we obtain that there must be a last finite successful iteration $(0, \ell_0)$. If the sequence were nevertheless infinite, this would mean that every iteration $(0, \ell)$ is unsuccessful for all $\ell > \ell_0$, causing $\Delta_{j,\ell}$ to converge to zero, which is impossible in view of Theorem 4.8. Hence the minimization sequence is finite. The same reasoning may be applied to every such sequence at level 0.

Now consider an arbitrary minimization sequence at level i (again, without loss of generality, within $\mathcal{R}(r,k)$ for some $k \geq 0$) and assume that each minimization sequence at level $i-1$ is finite and also that each successful iteration $(i-1,u)$ in every minimization sequence at this lower level satisfies

$$h_{i-1}(x_{i-1,u}) - h_{i-1}(x_{i-1,u+1}) \geq \omega_{i-1,u}\,\eta_1^i\kappa_h, \tag{4.52}$$

which is the direct generalization of (4.49) at level $i-1$. Consider a successful iteration (i,ℓ), whose existence is ensured by Lemma 4.6. If it is a Taylor iteration (i.e., if $(i,\ell) \in \mathcal{T}(i,\ell)$), we obtain as above that

$$h_i(x_{i,\ell}) - h_i(x_{i,\ell+1}) \geq \eta_1\kappa_h \geq \eta_1^{i+1}\kappa_h = \omega_{i,\ell}\,\eta_1^{i+1}\kappa_h \tag{4.53}$$

since $\eta_1 \in (0,1)$ and $\omega_{i,\ell} = 1$ for every successful Taylor iteration. If, on the other hand, iteration (i,ℓ) uses Step 2, then, assuming $x_{i-1,*} = x_{i-1,t+1}$, we obtain that

$$\begin{aligned} h_i(x_{i,\ell}) - h_i(x_{i,\ell+1}) &\geq \eta_1[h_{i-1}(x_{i-1,0}) - h_{i-1}(x_{i-1,*})] \\ &= \eta_1 \sum_{u=0}^{t} {}^{(\mathrm{S})}[h_{i-1}(x_{i-1,u}) - h_{i-1}(x_{i-1,u+1})]. \end{aligned}$$

Observing that $\omega_{i,\ell} = \tau_{i-1,t}$, (4.52) and (4.48) then give that

$$h_i(x_{i,\ell}) - h_i(x_{i,\ell+1}) \geq \eta_1^{i+1}\kappa_h \sum_{u=0}^{t} \omega_{i-1,u} = \tau_{i-1,t}\,\eta_1^{i+1}\kappa_h = \omega_{i,\ell}\,\eta_1^{i+1}\kappa_h. \tag{4.54}$$

Combining (4.53) and (4.54), we see that (4.52) again holds at level i instead of $i-1$. Moreover, as above,

$$h_i(x_{i,0}) - h_i(x_{i,p+1}) = \sum_{\ell=0}^{p} {}^{(\mathrm{S})}[h_i(x_{i,\ell}) - h_i(x_{i,\ell+1})] \geq \tau_{i,p}\,\eta_1^{i+1}\kappa_h \tag{4.55}$$

for the minimization sequence including iteration (i, ℓ). If $i = r$, $h_i = f$ is bounded below by assumption and (4.55) imposes that the number of successful iterations in this sequence must again be finite. The same conclusion holds if $i < r$, since h_i is continuous and hence bounded below on the set $\{x \in \Re^{n_i} \mid \|x - x_{i,0}\|_i \leq \Delta_{r,k}\}$, which contains $x_{i,p+1}$ because of Lemma 4.1 and (4.8). As for level 0, we may then conclude that the number of iterations (both successful and unsuccessful) in the minimization sequence is finite. Moreover, the same reasoning holds for every minimization sequence at level i, and the induction is complete. ∎

A first remarkable consequence of this theorem is an upper bound on the number of iterations needed by the trust-region algorithm to reduce the gradient norm at level r below a given threshold value.

THEOREM 4.10. *Assume that one knows a constant $f_{\rm low}$ such that $h_r(x_r) = f(x) \geq f_{\rm low}$ for every $x \in \Re^n$. Then Algorithm RMTR needs at most*

$$\left\lceil \frac{f(x_{r,0}) - f_{\rm low}}{\theta(\epsilon^g_{\rm min})} \right\rceil$$

successful Taylor iterations at any level to obtain an iterate $x_{r,k}$ such that $\|g_{r,k}\| \leq \epsilon^g_r$, where

$$\theta(\epsilon) = \eta_1^{r+1} \kappa_{\rm red} \epsilon \min\left[\frac{\epsilon}{\kappa_{\rm H}}, \gamma_1^{r+2} (\epsilon^{\Delta}_{\rm min})^r \min[\Delta^s_{\rm min}, \kappa_2 \epsilon] \right].$$

Proof. The desired bound directly follows from Theorem 4.9, (4.51), (4.47), and the definition of $\epsilon^{\rm g}_{\rm min}$. (To keep the expression manageable, we have refrained from substituting the value of κ_2 from (4.28) and, in this value, that of κ_1 from (4.16), all these values being independent of ϵ.) ∎

Of course, the bound provided by this theorem may be very pessimistic and not all the constants in the definition of $\theta(\epsilon)$ may be known in practice, but this loose complexity result is nevertheless theoretically interesting as it applies to general nonconvex problems. One should note that the bound is in terms of iteration numbers, and implicitly accounts only for the cost of computing a Taylor step satisfying (2.9). Theorem 4.10 suggests several comments.

1. The bound involves the number of successful Taylor iterations, that is, successful iterations where the trial step is computed without resorting to further recursion. This provides an adequate measure of the linear algebra effort for all successful iterations, since successful iterations using the recursion of Step 2 cost little beyond the evaluation of the level-dependent objective function and its gradient. Moreover, the number of such iterations is, by construction, at most equal to r times that of Taylor iterations (in the worst case, where each iteration at level r includes a full recursion to level 0 with a single successful iteration at each level $j > 0$). Hence the result shows that the number of necessary successful iterations, all levels included, is of order $1/\epsilon^2$ for small values of ϵ. This order is not qualitatively altered by the inclusion of unsuccessful iterations either, provided we replace the very successful trust-region radius update (top case in (2.16)) by

$$\Delta^+_{i,k} \in [\Delta_{i,k}, \gamma_3 \Delta_{i,k}] \quad \text{if } \rho_{i,k} \geq \eta_2$$

for some $\gamma_3 > 1$. Indeed, Theorem 4.8 imposes that the decrease in radius caused by unsuccessful iterations must asymptotically be compensated for by

an increase at successful iterations, irrespective of the fact that $\Delta_{\min}$ depends on ϵ by (4.47). This is to say that, if α is the average number of unsuccessful iterations per successful one at any level, then one must have that $\gamma_3\gamma_2^{\alpha} \geq 1$ and therefore that $\alpha \leq -\log(\gamma_3)/\log(\gamma_2)$. Thus the complexity bound in $1/\epsilon^2$ for small ϵ is modified by a constant factor only if all iterations (successful and unsuccessful) are considered. This therefore also gives a worst-case upper bound on the number of function and gradient evaluations.

2. This complexity bound is of the same order as the corresponding bound for the pure gradient method (see [31, p. 29]). This is not surprising given that it is based on the Cauchy condition, which itself results from a step in the steepest-descent direction.
3. The bound involves the number of successful Taylor iterations *summed up on all levels* (as a result of Theorem 4.9). Thus successful such iterations at cheap low levels decrease the number of necessary expensive ones at higher levels, and the multiscale algorithm requires (at least in the theoretical worst case) fewer Taylor iterations at the upper level than the single-level variant. This provides theoretical backing for the practical observation that the structure of multiscale unconstrained optimization problems can be used to advantage.
4. The constants involved in the definition of $\theta(\epsilon)$ do not depend on the problem dimension but rather on the properties of the problem (r, $\kappa_{\rm H}$, κ_σ) or of the algorithm itself ($\kappa_{\rm red}$, $\kappa_{\rm g}$, γ_1, η_1, η_2, $\epsilon^{\Delta}_{\min}$, $\Delta^{\rm s}_{\min}$). If we consider the case where different levels correspond to different discretization meshes and make the mild assumption that r and $\kappa_{\rm H}$ are uniformly bounded above and that κ_σ is uniformly bounded below, we observe that our complexity bound is mesh-independent.

A second important consequence of Theorem 4.9 is that the algorithm is globally convergent in the sense that it generates a subsequence of iterates whose gradients converge to zero if run with $\epsilon^{\rm g}_r = 0$.

COROLLARY 4.11. *Assume that Algorithm RMTR is called at the uppermost level with $\epsilon^g_r = 0$. Then*

$$\liminf_{k\to\infty} \|g_{r,k}\| = 0. \tag{4.56}$$

Proof. We first observe that the sequence of iterates $\{x_{r,k}\}$ generated by the algorithm called with $\epsilon^{\rm g}_r = 0$ is identical to that generated as follows. We consider, at level r, a sequence of gradient tolerances $\{\epsilon^{\rm g}_{r,j}\} \in (0,1)$ monotonically converging to zero, start the algorithm with $\epsilon^{\rm g}_r = \epsilon^{\rm g}_{r,0}$, and slightly alter the mechanism of Step 5 (at level r only) to reduce $\epsilon^{\rm g}_r$ from $\epsilon^{\rm g}_{r,j}$ to $\epsilon^{\rm g}_{r,j+1}$ as soon as $\|g_{r,k+1}\| \leq \epsilon^{\rm g}_{r,j}$. The calculation is then continued with this more stringent threshold until it is also attained, $\epsilon^{\rm g}_r$ is then again reduced, and so on. Since $\Delta_{r+1,0} = \infty$, each successive minimization at level r can stop at iteration k only if

$$\|g_{r,k+1}\| \leq \epsilon^{\rm g}_{r,j}. \tag{4.57}$$

Theorem 4.9 then implies that there are only finitely many successful iterations between two reductions of $\epsilon^{\rm g}_r$. We therefore obtain that for each $\epsilon^{\rm g}_{r,j}$ there is an arbitrarily large k such that (4.57) holds. The desired result then follows immediately from our assumption that $\{\epsilon^{\rm g}_{r,j}\}$ converges to zero. ∎

The interest of this result is mostly theoretical, since most practical applications of Algorithm RMTR consider a nonzero gradient tolerance $\epsilon^{\rm g}_r$.

The reader may have noticed that our theory still applies when we modify the technique described at the start of Corollary 4.11 by allowing a reduction of all the $\epsilon_i^{\rm g}$ to zero at the same time,[4] instead of merely reducing the uppermost one. If this modified technique is used, and assuming the trust region becomes asymptotically inactive at every level (as is most often the case in practice), each minimization sequence in the algorithm becomes infinite (as if it were initiated with a zero gradient threshold and an infinite initial radius). Recursion to lower levels then remains possible for arbitrarily small gradients and may therefore occur arbitrarily far in the sequence of iterates. Moreover, we may still apply Corollary 4.11 at each level and deduce that, if the trust region becomes asymptotically inactive,

$$\liminf_{k\to\infty} \|g_{i,k}\| = 0 \tag{4.58}$$

for all $i = 0, \ldots, r$.

As is the case for single-level trust-region algorithms, we now would like to prove that the limit inferior in (4.56) (and possibly (4.58)) can be replaced by a true limit, while still allowing recursion for very small gradients. We start by deriving a variant of Theorem 4.9 that does not assume that *all* gradient norms remain above some threshold to obtain a measure of the predicted decrease at some iteration (i, k).

LEMMA 4.12. *There exists a constant $\kappa_3 \in (0, 1)$ such that, for all (i, k) such that $\|g_{i,k}\| > 0$,*

$$\delta_{i,k} \geq \kappa_{\rm red}\eta_1^r\gamma_1^r\kappa_{\rm g}^r\|g_{i,k}\| \min\left[\, \Delta_{\rm min}^s,\, \kappa_3\|g_{i,k}\|,\, \Delta_{i,k} \,\right]. \tag{4.59}$$

Proof. Consider iteration (i, k). If it is a Taylor iteration, then, if we set

$$\kappa_3 = \min\left[\frac{\kappa_{\rm g}^r}{\kappa_{\rm H}},\, \kappa_2\kappa_{\rm g}^r \right] = \kappa_2\kappa_{\rm g}^r \in (0, 1), \tag{4.60}$$

(4.59) immediately follows from (2.9), (4.2), and the bounds $\kappa_{\rm g} \in (0, 1)$, $\eta_1 \in (0, 1)$, and $\gamma_1 \in (0, 1)$. Otherwise, define the iteration (j, ℓ) (with $j < i$) to be the deepest successful iteration in $\mathcal{R}(i, k)$ such that $g_{j,0} = g_{j,1} = \cdots = g_{j,\ell} = R_{j+1} \ldots R_i\, g_{i,k}$ and such that all iterations $(j + 1, t_{j+1})$, $(j + 2, t_{j+2})$, $\ldots$, up to $(i - 1, t_{i-1})$ of the path from (j, ℓ) to (i, k), are successful (meaning that iterations (j, u) are unsuccessful for $u = 0, \ldots, \ell - 1$, if any, and that iterations (p, u) are also unsuccessful for $p = j + 1, \ldots, i - 1$ and $u = 0, \ldots, t_p - 1$, if any). Note that such a path is guaranteed to exist because of Lemma 4.6. Using the first part of (2.14), we then obtain that

$$\|g_{j,0}\| = \|g_{j,1}\| = \cdots = \|g_{j,\ell}\| = \|R_{j+1} \ldots R_i\, g_{i,k}\| \geq \kappa_{\rm g}^r\|g_{i,k}\| > 0. \tag{4.61}$$

If $\ell = 0$, then

$$\Delta_{j,\ell} = \min[\Delta_j^{\rm s},\, \Delta_{j+1,t_{j+1}}] \geq \min[\Delta_{\rm min}^{\rm s},\, \Delta_{j+1,t_{j+1}}]. \tag{4.62}$$

If, on the other hand, $\ell > 0$, we know that iterations $(j, 0)$ to $(j, \ell-1)$ are unsuccessful. Corollary 4.5 then implies that (4.28) cannot hold for iteration $(j, \ell - 1)$, and thus that

$$\Delta_{j,\ell-1} > \min[\, \Delta_{\rm min}^{\rm s},\, \kappa_2\|g_{j,\ell-1}\| \,] = \min[\, \Delta_{\rm min}^{\rm s},\, \kappa_2\|g_{j,0}\| \,].$$

[4] The ratios $\epsilon_i^{\rm g}/\epsilon_r^{\rm g}$ could, for instance, be fixed or kept within prescribed bounds.

But this inequality, (2.16), (2.17), the unsuccessful nature of the first ℓ iterations at level j, (4.61), and the bound $\gamma_1 < 1$ then yield that

$$\begin{aligned}\Delta_{j,\ell} &\geq \min[\,\gamma_1\Delta_{j,\ell-1},\, \Delta_{j+1,t_{j+1}} - \|x_{j,0} - x_{j,\ell}\|_j\,]\\ &= \min[\,\gamma_1\Delta_{j,\ell-1},\, \Delta_{j+1,t_{j+1}}\,]\\ &\geq \min[\,\gamma_1 \min(\Delta^{\rm s}_{\min},\, \kappa_2\|g_{j,0}\|),\, \Delta_{j+1,t_{j+1}}\,]\\ &\geq \min[\,\gamma_1 \min(\Delta^{\rm s}_{\min},\, \kappa_2\kappa_{\rm g}^r\|g_{i,k}\|),\, \Delta_{j+1,t_{j+1}}\,]\\ &\geq \gamma_1 \min[\,\Delta^{\rm s}_{\min},\, \kappa_2\kappa_{\rm g}^r\|g_{i,k}\|,\, \Delta_{j+1,t_{j+1}}\,].\end{aligned}$$

Combining this last inequality with (4.62), we conclude that, for $\ell \geq 0$,

$$\Delta_{j,\ell} \geq \gamma_1 \min[\,\Delta^{\rm s}_{\min},\, \kappa_2\kappa_{\rm g}^r\|g_{i,k}\|,\, \Delta_{j+1,t_{j+1}}\,].$$

Our choice of iteration (j,ℓ) also ensures that the same reasoning can now be applied not only to iteration (j,ℓ) but also to every iteration in the path $(j+1,t_{j+1})$, $\ldots$, $(i-1,t_{i-1})$, because the first part of (2.14) implies that $\|g_{p,0}\| = \|R_{p+1}\ldots R_i\, g_{i,k}\| \geq \kappa_{\rm g}^r\|g_{i,k}\|$ for all $j \leq p < i$. Thus we obtain that

$$\Delta_{j+u,t_{j+u}} \geq \gamma_1 \min[\,\Delta^{\rm s}_{\min},\, \kappa_2\kappa_{\rm g}^r\|g_{i,k}\|,\, \Delta_{j+u+1,t_{j+u+1}}\,]$$

for $u = 0,\ldots,i-j-1$ (where we identify $t_i = k$ for $u = i-j-1$). We may then use these bounds recursively level by level and deduce that

$$\begin{aligned}\Delta_{j,\ell} &\geq \gamma_1 \min[\,\Delta^{\rm s}_{\min},\, \kappa_2\kappa_{\rm g}^r\|g_{j,k}\|,\, \Delta_{j+1,t_{j+1}}\,]\\ &\geq \gamma_1 \min[\,\Delta^{\rm s}_{\min},\, \kappa_2\kappa_{\rm g}^r\|g_{i,k}\|,\, \gamma_1 \min(\Delta^{\rm s}_{\min},\, \kappa_2\kappa_{\rm g}^r\|g_{i,k}\|,\, \Delta_{j+2,t_{j+2}})\,]\\ &\geq \gamma_1^2 \min[\,\Delta^{\rm s}_{\min},\, \kappa_2\kappa_{\rm g}^r\|g_{i,k}\|,\, \Delta_{j+2,t_{j+2}}\,]\\ &\geq \gamma_1^r \min[\,\Delta^{\rm s}_{\min},\, \kappa_2\kappa_{\rm g}^r\|g_{i,k}\|,\, \Delta_{i,k}\,]\end{aligned} \tag{4.63}$$

because $\gamma_1 < 1$. On the other hand, $(j,\ell) \in \mathcal{T}(i,k)$ by construction, and we therefore obtain from (2.9) and (4.2) that

$$\delta_{j,\ell} \geq \kappa_{\rm red}\|g_{j,\ell}\| \min\left[\,\frac{\|g_{j,\ell}\|}{\kappa_{\rm H}},\, \Delta_{j,\ell}\right]. \tag{4.64}$$

Gathering now (4.61), (4.63), and (4.64), we obtain that

$$\delta_{j,\ell} \geq \kappa_{\rm red}\kappa_{\rm g}^r\|g_{i,k}\| \min\left[\,\frac{\kappa_{\rm g}^r\|g_{i,k}\|}{\kappa_{\rm H}},\, \gamma_1^r \min[\Delta^{\rm s}_{\min},\, \kappa_2\kappa_{\rm g}^r\|g_{i,k}\|,\, \Delta_{i,k}]\right],$$

and thus, using (4.60), that

$$\delta_{j,\ell} \geq \kappa_{\rm red}\kappa_{\rm g}^r\gamma_1^r\|g_{i,k}\| \min\left[\,\Delta^{\rm s}_{\min},\, \kappa_3\|g_{i,k}\|,\, \Delta_{i,k}\,\right]. \tag{4.65}$$

But the fact that all iterations on the path from (j,ℓ) to (i,k) are successful also implies that

$$\begin{aligned}\delta_{i,k} &= h_{i-1}(x_{i-1,0}) - h_{i-1}(x_{i-1,*}) \geq h_{i-1}(x_{i-1,t_{i-1}}) - h_{i-1}(x_{i-1,t_{i-1}+1})\\ &\geq \eta_1\delta_{i-1,t_{i-1}} = \eta_1[h_{i-2}(x_{i-2,0}) - h_{i-2}(x_{i-2,*})]\\ &\geq \eta_1[h_{i-2}(x_{i-2,t_{i-2}}) - h_{i-2}(x_{i-2,t_{i-2}+1})] \geq \eta_1^2\delta_{i-2,t_{i-2}} \geq \eta_1^r\delta_{j,\ell}.\end{aligned}$$

The bound (4.59) then follows from this last inequality and (4.65). ∎

All the elements are now in place to show that, if the algorithm is run with $\epsilon_r^{\rm g} = 0$, then gradients at level r converge to zero.

THEOREM 4.13. *Assume that Algorithm RMTR is called at the uppermost level with $\epsilon_r^g = 0$. Then*

$$\lim_{k\to\infty} \|g_{r,k}\| = 0. \tag{4.66}$$

Proof. The proof is identical to that of Theorem 6.4.6 on p. 137 of [13], with (4.59) (with $i = r$) now playing the role of the sufficient model reduction condition AA.1 at level r. ∎

This last result implies, in particular, that any limit point of the infinite sequence $\{x_{r,k}\}$ is first-order critical for problem (2.1). But we may draw stronger conclusions. If we assume that the trust region becomes asymptotically inactive at all levels and that all $\epsilon_i^{\rm g}$ $(i = 0, \ldots, r-1)$ are driven down to zero together with $\epsilon_r^{\rm g}$ (thus allowing recursion even for very small gradients), then, as explained above, each minimization sequence in the algorithm becomes infinite, and we may apply Theorem 4.13 to each of them, concluding that, if the trust region becomes asymptotically inactive,

$$\lim_{k\to\infty} \|g_{i,k}\| = 0$$

for every level $i = 0, \ldots, r$. The behavior of Algorithm RMTR is therefore truly coherent with its multiscale formulation, since the same convergence results hold for each level.

The convergence results at the upper level are unaffected if minimization sequences at lower levels are "prematurely" terminated, provided each such sequence contains at least one successful iteration. Indeed, Lemmas 4.1 and 4.2 do not depend on the actual stopping criterion used, and all subsequent proofs do not depend on it either. Thus, one might think of stopping a minimization sequence after a preset number of successful iterations: in combination with the freedom left at Step 1 to choose the model whenever (2.14) holds, this strategy allows a straightforward implementation of fixed lower-iterations patterns, like the V or W cycles in multigrid methods. This is what we have done in section 3.

Our theory also remains essentially unchanged if we merely insist on first-order coherence (i.e., conditions (2.5) and (2.6)) to hold only for small enough trust-region radii $\Delta_{i,k}$, or only up to a perturbation of the order of $\Delta_{i,k}$ or $\|g_{i,k}\|\Delta_{i,k}$. Other generalizations may be possible. Similarly, although we have assumed for motivation purposes that each f_i is "more costly" to minimize that f_{i-1}, we have not used this feature in the theory presented above, nor have we used the form of the lower levels objective functions. In particular, our choice of section 3 to define f_i as identically zero for $i = 0, \ldots, r-1$ satisfies all our assumptions. Nonconstant prolongation and restriction operators of the form $P_i(x_{i,k})$ and $R_i(x_{i,k})$ may also be considered, provided the singular values of these operators remain uniformly bounded.

In its full generality, convergence to second-order critical points appears to be out of reach unless one is able to guarantee some "eigenpoint condition." Such a condition imposes that, if $\tau_{i,k}$, the smallest eigenvalue of $H_{i,k}$, is negative, then

$$m_{i,k}(x_{i,k}) - m_{i,k}(x_{i,k} + s_{i,k}) \geq \kappa_{\rm eip} |\tau_{i,k}| \min[\tau_{i,k}^2, \Delta_{i,k}^2]$$

for some constant $\kappa_{\rm eip} \in (0, \frac{1}{2})$ (see AA.2 in [13, p. 153]). This is easy to obtain at relatively coarse levels, where the cost of an eigenvalue computation or of a factorization remains acceptable. For instance, the algorithm considered in section 3 is convergent

to critical points that satisfy second-order optimality conditions *at the coarsest level*. This results from the application of the Moré–Sorensen exact trust-region subproblem solver at that level, for which this property is well known (see section 6.6 of [13], for instance). The idea of imposing an eigenpoint condition at the coarsest level to obtain second-order criticality at that level is also at the core of the globalization proposal in [9], but it can be verified [21] that this technique does not enforce second-order convergence at finer levels. However, imposing an eigenpoint condition at fine levels may be judged impractical: for instance, the SCM smoothing strategy described above does not guarantee such a condition but merely that

$$m_{i,k}(x_{i,k}) - m_{i,k}(x_{i,k} + s_{i,k}) \geq \tfrac{1}{2}|\mu_{i,k}|\Delta_{i,k}^2,$$

where $\mu_{i,k}$ is the most negative diagonal element of $H_{i,k}$. This weaker result is caused by the fact that SCM limits its exploration of the model's curvature to the coordinate axes, at variance with the TCG and GLTR methods, which implicitly construct Lanczos approximations to Hessian eigenvalues. Convergence to fine-level first-order critical points satisfying a weak version of second-order optimality can, however, be expected in this case. In particular, the diagonal elements of the objective function's Hessian have to be nonnegative at such limit points (see [21]).

5. Comments and perspectives. We have defined a class of recursive trust-region algorithms whose members are able to exploit cheap lower-level models in a multiscale optimization problem. This class has been proved to be well defined and globally convergent to first order; preliminary numerical experience suggests that it may have strong potential. We have also presented a theoretical complexity result giving a bound on the number of iterations that are required by the algorithms of our class to find an approximate critical point of the objective function within prescribed accuracy. This last result also shows that the total complexity of solving an unconstrained multiscale problem can be shared amongst the levels, exploiting the structure to advantage.

Although the example of discretized problems has been used as a major motivation for our work, this is not the only case where our theory can be applied. We think, in particular, of cases where different models of the true objective function might live in the same space but involve different levels of complexity and/or cost. This is of interest, for instance, in a number of problems arising from physics, like data assimilation in weather forecasting [15], where different models may involve different levels of sophistication in the physical modeling itself. More generally, the algorithms and theory presented here are relevant in most areas where simplified models are considered, such as multidisciplinary optimization [1, 2, 3] or PDE-constrained problems [4, 14].

We may also think of investigating even more efficient algorithms combining the trust-region framework developed here with other globalization techniques, like line-searches [17, 32, 39], nonmonotone techniques [40, 42, 44], or filter methods [20]. While this might add yet another level of technicality to the convergence proofs, we expect such extensions to be possible and the resulting algorithms to be of practical interest.

Another important research direction is to investigate what kinds of Hessian (and possibly gradient) approximations are practically efficient within our framework, especially at the fine levels. Various options are possible, ranging from specialized finite differences to secant approximations.

Applying recursive trust-region methods of the type discussed here to constrained problems is another obvious avenue of research. Although we anticipate the associated

convergence theory to be again more technically difficult, intuition and limited numerical experience suggest that the power of such methods should also be exploitable in this case.

A number of practical issues related to Algorithm RMTR (such as alternative gradient smoothing and choice of cycle patterns) have not been discussed, although they may be crucial in practice. We investigate these issues in a forthcoming paper describing (so far encouraging) numerical experience with Algorithm RMTR.

Acknowledgments. The authors are indebted to Nick Gould for his comments on a draft of the manuscript and to Natalia Alexandrov for stimulating discussion.

REFERENCES

[1] N. M. Alexandrov, J. E. Dennis, R. M. Lewis, and V. Torczon, *A trust region framework for managing the use of approximation models*, Structural Optimization, 15 (1998), pp. 16–23.

[2] N. M. Alexandrov and R. L. Lewis, *An overview of first-order model management for engineering optimization*, Optimization and Engineering, 2 (2001), pp. 413–430.

[3] N. M. Alexandrov, R. L. Lewis, C. R. Gumbert, L. L. Green, and P. A. Newman, *Approximation and model management in aerodynamic optimization with variable fidelity models*, J. Aircraft, 38 (2001), pp. 1093–1101.

[4] E. Arian, M. Fahl, and E. W. Sachs, *Trust-Region Proper Orthogonal Decomposition for Flow Control*, Technical report 2000-25, Institute for Computer Applications in Science and Engineering, NASA Langley Research Center, Hampton, VA, 2000.

[5] R. E. Bank, P. E. Gill, and R. F. Marcia, *Interior point methods for a class of elliptic variational inequalities*, in High Performance Algorithms and Software for Nonlinear Optimization, Biegler et al., eds., Springer-Verlag, New York, 2003, pp. 218–235.

[6] S. J. Benson, L. C. McInnes, J. J. Moré, and J. Sarich, *Scalable Algorithms in Optimization: Computational Experiments*, Preprint ANL/MCS-P1175-0604, Mathematics and Computer Science, Argonne National Laboratory, Argonne, IL, 2004; in Proceedings of the 10th AIAA/ISSMO Multidisciplinary Analysis and Optimization (MA&O) Conference, Albany, NY, 2004.

[7] J. T. Betts and S. O. Erb, *Optimal low thrust trajectory to the moon*, SIAM J. Appl. Dyn. Syst., 2 (2003), pp. 144–170.

[8] T. Biegler, O. Ghattas, M. Heinkenschloss, and B. Van Bloemen Waanders, Eds., *High Performance Algorithms and Software for Nonlinear Optimization*, Springer-Verlag, New York, 2003.

[9] A. Borzi and K. Kunisch, *A globalisation strategy for the multigrid solution of elliptic optimal control problems*, Optim. Methods Softw., 21 (2006), pp. 445–459.

[10] J. H. Bramble, *Multigrid Methods*, Longman Scientific and Technical, Harlow, UK, 1993.

[11] A. Brandt, *Multi-level adaptative solutions to boundary value problems*, Math. Comp., 31 (1977), pp. 333–390.

[12] W. L. Briggs, V. E. Henson, and S. F. McCormick, *A Multigrid Tutorial*, 2nd ed., SIAM, Philadelphia, 2000.

[13] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Trust-Region Methods*, MPS-SIAM Ser. Optim. 1, SIAM, Philadelphia, 2000.

[14] M. Fahl and E. Sachs, *Reduced order modelling approaches to PDE-constrained optimization based on proper orthogonal decomposition*, in High Performance Algorithms and Software for Nonlinear Optimization, Biegler et al., eds., Springer-Verlag, New York, 2003, pp. 268–281.

[15] M. Fisher, *Minimization algorithms for variational data assimilation*, in Proceedings of the ECMWF Seminar on Recent Developments in Numerical Methods for Atmospheric Modelling, Reading, UK, 1998, pp. 364–385.

[16] E. Gelman and J. Mandel, *On multilevel iterative methods for optimization problems*, Math. Programming, 48 (1990), pp. 1–17.

[17] E. M. Gertz, *Combination Trust-Region Line-Search Methods for Unconstrained Optimization*, Ph.D. thesis, Department of Mathematics, University of California, San Diego, CA, 1999.

[18] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed., Johns Hopkins University Press, Baltimore, 1989.

[19] N. I. M. Gould, S. Lucidi, M. Roma, and Ph. L. Toint, *Solving the trust-region subproblem using the Lanczos method*, SIAM J. Optim., 9 (1999), pp. 504–525.
[20] N. I. M. Gould, C. Sainvitu, and Ph. L. Toint, *A filter-trust-region method for unconstrained optimization*, SIAM J. Optim., 16 (2005), pp. 341–357.
[21] S. Gratton, A. Sartenaer, and Ph. L. Toint, *Second-order convergence properties of trust-region methods using incomplete curvature information, with an application to multigrid optimization*, J. Comput. Appl. Math., 24 (2006),pp. 676–692.
[22] A. Griewank and Ph. L. Toint, *Local convergence analysis for partitioned quasi-Newton updates*, Numer. Math., 39 (1982), pp. 429–448.
[23] W. Hackbusch, *Iterative Solution of Large Sparse Systems of Equations*, Appl. Math. Sci. 95, Springer-Verlag, New York, 1994.
[24] W. Hackbusch and A. Reusken, *Analysis of a damped nonlinear multilevel method*, Numer. Math., 55 (1989), pp. 225–246.
[25] P. W. Hemker and G. M. Johnson, *Multigrid approach to Euler equations*, in Multigrid Methods, Frontiers Appl. Math. 3, S. F. McCormick, ed., SIAM, Philadelphia, 1987, pp. 57–72.
[26] R. M. Lewis and S. G. Nash, *Practical aspects of multiscale optimization methods for VLSI-CAD*, in Multiscale Optimization and VLSI/CAD, J. Cong and J. R. Shinnerl, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002, pp. 265–291.
[27] R. M. Lewis and S. G. Nash, *Model problems for the multigrid optimization of systems governed by differential equations*, SIAM J. Sci. Comput., 26 (2005), pp. 1811–1837.
[28] J. J. Moré, *Terascale optimal PDE solvers*, in ICIAM 2003 Conference, Sydney, Australia, 2003.
[29] J. J. Moré and D. C. Sorensen, *On the use of directions of negative curvature in a modified Newton method*, Math. Programming, 16 (1979), pp. 1–20.
[30] S. G. Nash, *A multigrid approach to discretized optimization problems*, Optim. Methods Softw., 14 (2000), pp. 99–116.
[31] Y. Nesterov, *Introductory Lectures on Convex Optimization*, Appl. Optim. 87, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2004.
[32] J. Nocedal and Y. Yuan, *Combining trust region and line search techniques*, in Advances in Nonlinear Programming, Y. Yuan, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 153–176.
[33] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, London, 1970.
[34] M. J. D. Powell, *A Fortran Subroutine for Unconstrained Minimization Requiring First Derivatives of the Objective Function*, Technical report R-6469, AERE Harwell Laboratory, Harwell, Oxfordshire, England, 1970.
[35] M. J. D. Powell, *A new algorithm for unconstrained optimization*, in Nonlinear Programming, J. B. Rosen, O. L. Mangasarian, and K. Ritter, eds., Academic Press, London, 1970, pp. 31–65.
[36] A. Sartenaer, *Automatic determination of an initial trust region in nonlinear programming*, SIAM J. Sci. Comput., 18 (1997), pp. 1788–1803.
[37] T. Steihaug, *The conjugate gradient method and trust regions in large scale optimization*, SIAM J. Numer. Anal., 20 (1983), pp. 626–637.
[38] Ph. L. Toint, *Towards an efficient sparsity exploiting Newton method for minimization*, in Sparse Matrices and Their Uses, I. S. Duff, ed., Academic Press, London, 1981, pp. 57–88.
[39] Ph. L. Toint, *VE*08*AD, a routine for partially separable optimization with bounded variables*, Harwell Subroutine Library, 2 (1983).
[40] Ph. L. Toint, *Non-monotone trust-region algorithms for nonlinear optimization subject to convex constraints*, Math. Programming, 77 (1997), pp. 69–94.
[41] U. Trottenberg, C. W. Oosterlee, and A. Schüller, *Multigrid*, Elsevier, Amsterdam, The Netherlands, 2001.
[42] M. Ulbrich, *Nonmonotone trust-region methods for bound-constrained semismooth equations with applications to nonlinear mixed complementarity problems*, SIAM J. Optim., 11 (2001), pp. 889–917.
[43] P. Wesseling, *An Introduction to Multigrid Methods*, J. Wiley and Sons, Chichester, UK, 1992.
[44] Y. Xiao and F. Zhou, *Nonmonotone trust region methods with curvilinear path in unconstrained optimization*, Computing, 48 (1992), pp. 303–317.
[45] I. Yavneh and G. Dardyk, *A multilevel nonlinear method*, SIAM J. Sci. Comput., 28 (2006), pp. 24–46.

SIAM J. OPTIM.
Vol. 19, No. 1, pp. 445–471

ON THE GLOBAL SOLUTION OF LINEAR PROGRAMS WITH LINEAR COMPLEMENTARITY CONSTRAINTS*

JING HU†, JOHN E. MITCHELL†, JONG-SHI PANG‡, KRISTIN P. BENNETT†, AND GAUTAM KUNAPULI†

Abstract. This paper presents a parameter-free integer-programming-based algorithm for the global resolution of a linear program with linear complementarity constraints (LPCCs). The cornerstone of the algorithm is a minimax integer program formulation that characterizes and provides certificates for the three outcomes—infeasibility, unboundedness, or solvability—of an LPCC. An extreme point/ray generation scheme in the spirit of Benders decomposition is developed, from which valid inequalities in the form of satisfiability constraints are obtained. The feasibility problem of these inequalities and the carefully guided linear-programming relaxations of the LPCC are the workhorses of the algorithm, which also employs a specialized procedure for the sparsification of the satifiability cuts. We establish the finite termination of the algorithm and report computational results using the algorithm for solving randomly generated LPCCs of reasonable sizes. The results establish that the algorithm can handle infeasible, unbounded, and solvable LPCCs effectively.

Key words. linear programs with linear complimentarity constraints, global resolution, cutting-plane methods

AMS subject classifications. Primary, 90C33; Secondary, 90C26, 90C10

DOI. 10.1137/07068463x

1. Introduction. Forming a subclass of the class of mathematical programs with equilibrium/complementarity constraints (MPECs/MPCCs) [37, 39, 11], linear programs with linear complementarity constraints (LPCCs) are disjunctive linear optimization problems that contain a set of complementarity conditions. In turn, a large subclass of LPCCs are bilevel linear/quadratic programs [10] that provide a broad modeling framework for parameter identification in convex quadratic programming; an example of such an application was proposed recently for the cross validation of a host of machine-learning problems [6, 33, 32]. While there have been significant recent advances on nonlinear-programming- (NLP-) based computational methods for solving MPECs and the closely related MPCCs [1, 2, 3, 8, 14, 15, 19, 20, 29, 30, 25, 35, 36, 42, 43], many of which have nevertheless focused on obtaining stationary solutions [12, 13, 37, 39, 38, 40, 42, 48, 47, 46], the global solution of an LPCC remains elusive. Particularly impressive among these advances is the suite of NLP solvers publicly available on the NEOS system [49]; many of them, such as FILTER and KNITRO, are capable of producing a solution of some sort to an LPCC very efficiently. Yet they are incapable of ascertaining the quality of the computed solution. This is the major deficiency of these numerical solvers. Continuing our foray into the subject of computing global solutions of LPCCs, which begins with the recent article [41] that pertains to a special problem arising from the optimization of the value at risk, the present

*Received by the editors September 21, 2007; accepted for publication (in revised form) December 17, 2007; published electronically May 9, 2008. This work was supported in part by the Office of Naval Research under grant N00014-06-1-0014.

http://www.siam.org/journals/siopt/19-1/68463.html

†Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180-1590 (huj@rpi.edu, mitchj@rpi.edu, bennek@rpi.edu, kunapg@rpi.edu). The second author was supported by the National Science Foundation under grant DMS-0317323.

‡Department of Industrial and Enterprise Systems Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801 (jspang@uiuc.edu).

paper proposes a parameter-free integer-programming-based cutting-plane algorithm for globally resolving a general LPCC.

As a disjunctive linear optimization problem, the global solution of an LPCC has been the subject of sustained, but not particularly focused, investigation since the early work of Ibaraki [26, 27] and Jeroslow [28], who pioneered some cutting-plane methods for solving a "complementary program," which is a historical and not widely used name for an LPCC. Over the years, various integer-programming-based methods [4, 5, 21] and global-optimization-based methods [16, 17, 44, 45] have been developed that are applicable to an LPCC. In this paper, we present a new cutting-plane method that will successfully resolve a general LPCC in finite time; i.e., the method will terminate with one of the following three mutually exclusive conclusions: the LPCC is infeasible, the LPCC is feasible but has an unbounded objective, or the LPCC attains a finite optimal solution. We also leverage the advances of the NLP solvers and use two of them to benchmark our algorithm. In addition, we propose a simple linear-programming-based preprocessor whose effectiveness will be demonstrated via computational results.

The proposed method begins with an equivalent formulation of an LPCC as a 0-1 integer program (IP) involving a conceptually very large parameter, whose existence is not guaranteed unless a certain boundedness condition holds. Via dualization of the linear-programming relaxation of the IP, we obtain a minimax 0-1 integer program, which yields a certificate for the three states of the LPCC, without any a priori boundedness assumption. The original 0-1 IP with the conceptual parameter provides the formulation for the application of Benders decomposition [34], which we show can be implemented without involving the parameter in any way. Thus, the resulting algorithm is reminiscent of the well-known phase I implementation of the "big-M" method for solving linear programs, wherein the big-M formulation is only conceptual whose practical solution does not require the knowledge of the scalar M.

The implementation of our parameter-free algorithm is accomplished by solving integer subprograms defined solely by *satisfiability constraints* [7, 31]; in turn, each such constraint corresponds to a "piece" of the LPCC. By using this interpretation, the overall algorithm can be considered as solving the LPCC by searching on its (finitely many) linear-programming pieces, with the search guided by solving the satisfiability IPs. The implementation of the algorithm is aided by valid upper bounds on the LPCC optimal objective value that are being updated as the algorithm progresses, which also serve to provide the desired certificates at the termination of the algorithm.

Hooker [22, 23] and Hooker and Ottosson [24] have presented a general Benders decomposition framework for integer-programming problems, where the subproblems and the master problem may be solved by various techniques, for example, constraint programming. The constraints returned by the subproblems may well be satisfiability constraints similar to those we derive from solving a linear-programming subproblem, leading to a satisfiability master problem. Hooker develops a broad framework for integrating different solution methodologies, with the solution approaches driven by the types of constraints. Different decomposition methods are available depending on the particular mix of types of constraints. Codato and Fischetti [9] have specialized Hooker's approach to solve integer programs arising from linear programs with conditional constraints of the kind "if then," of which the complementarity condition is a special case. Such constraints are modeled with the introduction of integer variables together with big-M coefficients. The cited article presents computational results for feasible bounded problems where only the binary variables associated with the big-M coefficients appear in the objective function. In our problem (2.3), the objective func-

tion is a linear function of the continuous variables; the objective function could be regarded as a nonlinear function of the binary variables (see $\varphi(z)$ defined in (2.9)). Our algorithmic approach can successfully characterize infeasible and unbounded LPCC problems as well as solve problems with a finite optimal value. Hooker [23] states that "the success of a Benders method often rests on finding strong Benders cuts that rule out as many infeasible solutions as possible." The sparsification methodology we present in section 4 is an approach to generate strong Benders cuts.

The organization of the rest of the paper is as follows. Section 2 presents the formal statement of the LPCC, summarizes the three states of the LPCC, and introduces the new minimax IP formulation. Section 3 reformulates the minimax IP formulation in terms of the extreme points and rays of the key polyhedron Ξ (see (2.6)) and establishes the theoretical foundation for the cutting-plane algorithm to be presented in section 5. The key steps of the algorithm, which involve solving linear programs (LPs) to sparsify the satisfiability constraints, are explained in section 4. The sixth and last section reports the computational results and completes the paper with some concluding remarks.

2. Preliminary discussion. Let $c \in \Re^n$, $d \in \Re^m$, $f \in \Re^k$, $q \in \Re^m$, $A \in \Re^{k \times n}$, $B \in \Re^{k \times m}$, $M \in \Re^{m \times m}$, and $N \in \Re^{m \times n}$ be given. Consider the LPCC [18] of finding $(x, y) \in \Re^n \times \Re^m$ in order to

$$\begin{array}{ll} \underset{(x,y)}{\text{minimize}} & c^T x + d^T y \\ \text{subject to} & Ax + By \geq f \\ \text{and} & 0 \leq y \perp q + Nx + My \geq 0, \end{array} \tag{2.1}$$

where $a \perp b$ means that the two vectors are orthogonal; i.e., $a^T b = 0$. It is well known that the LPCC is equivalent to the minimization of a large number of linear programs, each defined on one *piece* of the feasible region of the LPCC. That is, for each subset α of $\{1, \ldots, m\}$ with complement $\bar{\alpha}$, we may consider the LP(α):

$$\begin{array}{ll} \underset{(x,y)}{\text{minimize}} & c^T x + d^T y \\ \text{subject to} & Ax + By \geq f, \\ & (\, q + Nx + My \,)_\alpha \geq 0 = y_\alpha, \\ \text{and} & (\, q + Nx + My \,)_{\bar{\alpha}} = 0 \leq y_{\bar{\alpha}}. \end{array} \tag{2.2}$$

The following facts are consequences of the disjunctive property of the complementarity condition:

(a) The LPCC (2.1) is infeasible if and only if the LP(α) is infeasible for *all* $\alpha \subseteq \{1, \ldots, m\}$;

(b) the LPCC (2.1) is feasible and has an unbounded objective if and only if the LP(α) is feasible and has an unbounded objective for *some* $\alpha \subseteq \{1, \ldots, m\}$;

(c) the LPCC (2.1) is feasible and attains a finite optimal objective value if and only if (i) a subset α of $\{1, \ldots, m\}$ exists such that the LP(α) is feasible and (ii) every such feasible LP(α) has a finite optimal objective value; in this case, the optimal objective value of the LPCC (2.1), denoted $\text{LPCC}_{\min}$, is the minimum of the optimal objective values of all such feasible LPs.

The first step in our development of an IP-based algorithm for solving the LPCC (2.1) without any a priori assumption is to derive results parallel to the above three

facts in terms of some parameter-free integer problems. For this purpose, we recall the standard approach of solving (2.1) as an IP containing a large parameter. This approach is based on the following "equivalent" IP formulation of (2.1) wherein the complementarity constraint is reformulated in terms of the binary vector $z \in \{0,1\}^m$ via a conceptually very large scalar $\theta > 0$:

$$
\begin{array}{lll}
\underset{(x,y,z)}{\text{minimize}} & & c^T x + d^T y \\
\text{subject to} & & Ax + By \geq f, \\
& \theta\, z \;\geq\; q + Nx + My \geq 0, & \\
& & \theta(\,\mathbf{1} - z\,)\, y \geq 0, \\
\text{and} & & z \in \{\,0,1\,\}^m,
\end{array}
\tag{2.3}
$$

where $\mathbf{1}$ is the m-vector of all ones. In the standard approach, we first derive a valid value on θ by solving LPs to obtain bounds on all of the variables and constraints of (2.1). We then solve the fixed IP (2.3) by using the so-obtained θ by, for example, the Benders approach. There are two drawbacks of such an approach: One is the limitation of the approach to problems with bounded feasible regions; the other drawback is the nontrivial computation to derive the required bounds even if they are known to exist implicitly. In contrast, our new approach removes such a theoretical restriction and eliminates the front-end computation of bounds. The price of the new approach is that it solves a (finite) family of IPs of a special type, each defined solely by constraints of the *satisfiability* type. The following discussion sets the stage for the approach. [A referee suggests that "another drawback of attacking the integer program (2.3) is the probably (very) weak LP relaxations (which will affect the convergence of branch and cut methods as well as approaches based on Benders decomposition)."]

For a given binary vector z and a positive scalar θ, we associate with (2.3) the linear program below, which we denote $\mathrm{LP}(\theta; z)$:

$$
\begin{array}{lll}
\underset{(x,y)}{\text{minimize}} & c^T x + d^T y & \\
\text{subject to} & Ax + By \geq f & (\,\lambda\,), \\
& Nx + My \geq -q & (\,u^-\,), \\
& -Nx - My \geq q - \theta\, z & (\,u^+\,), \\
& -y \geq -\theta\,(\,\mathbf{1} - z\,) & (\,v\,), \\
\text{and} & y \geq 0, &
\end{array}
\tag{2.4}
$$

where the dual variables of the respective constraints are given in the parentheses. The dual of (2.4), which we denote $\mathrm{DLP}(\theta, z)$, is

$$
\begin{array}{lll}
\underset{(\lambda,u^\pm,v)}{\text{maximize}} & f^T\lambda + q^T(\,u^+ - u^-\,) - \theta\,\left[\, z^T u^+ + (\,\mathbf{1} - z\,)^T v \,\right] \\
\text{subject to} & A^T\lambda - N^T(\,u^+ - u^-\,) \;=\; c, \\
& B^T\lambda - M^T(\,u^+ - u^-\,) - v \;\leq\; d, \\
\text{and} & (\,\lambda, u^\pm, v\,) \;\geq\; 0.
\end{array}
\tag{2.5}
$$

Let $\Xi \subseteq \Re^{k+3m}$ be the feasible region of the $\mathrm{DLP}(\theta, z)$; i.e.,

$$
\Xi \equiv \left\{ (\,\lambda, u^\pm, v\,) \;\geq\; 0 \;:\; \begin{array}{r} A^T\lambda - N^T(\,u^+ - u^-\,) \;=\; c \\ B^T\lambda - M^T(\,u^+ - u^-\,) - v \;\leq\; d \end{array} \right\}.
\tag{2.6}
$$

Note that Ξ is a fixed polyhedron independent of the pair (θ, z); Ξ has at least one extreme point if it is nonempty. Let $\mathrm{LP}_{\min}(\theta; z)$ and $d(\theta; z)$ denote the optimal objective value of (2.4) and (2.5), respectively. Throughout, we adopt the standard convention that the optimal objective value of an infeasible maximization (minimization) problem is defined to be $-\infty$ (∞). We summarize some basic relations between the above programs in the following result.

PROPOSITION 2.1. *The following three statements hold:*

(a) *Any feasible solution (x^0, y^0) of (2.1) induces a pair (θ_0, z^0), where $\theta_0 > 0$ and $z^0 \in \{0,1\}^m$, such that the tuple (x^0, y^0, z^0) is feasible to (2.3) for all $\theta \geq \theta_0$; such a z^0 has the property that*

$$\begin{aligned} (\, q + Nx^0 + My^0 \,)_i \, > \, 0 \Rightarrow z_i^0 \, = \, 1, \\ (\, y^0 \,)_i \, > \, 0 \Rightarrow z_i^0 \, = \, 0. \end{aligned} \tag{2.7}$$

(b) *Conversely, if (x^0, y^0, z^0) is feasible to (2.3) for some $\theta \geq 0$, then (x^0, y^0) is feasible to (2.1).*

(c) *If (x^0, y^0) is an optimal solution to (2.1), then it is optimal to the $\mathrm{LP}(\theta, z^0)$ for all pairs (θ, z^0) such that $\theta \geq \theta_0$ and z^0 satisfies (2.7); moreover, for each $\theta > \theta_0$, any optimal solution $(\widehat{\lambda}, \widehat{u}^{\pm}, \widehat{v})$ of the $\mathrm{DLP}(\theta, z^0)$ satisfies $(z^0)^T \widehat{u}^+ + (\, \mathbf{1} - z^0 \,)^T \widehat{v} = 0$.*

Proof. Only (c) requires a proof. Suppose that (x^0, y^0) is optimal to (2.1). Let (θ, z^0) be such that $\theta \geq \theta_0$ and $z^0 \in \{0,1\}^m$ satisfies (2.7). Then (x^0, y^0) is feasible to the $\mathrm{LP}(\theta, z^0)$; hence

$$c^T x^0 + d^T y^0 \, \geq \, \mathrm{LP}_{\min}(\theta, z^0). \tag{2.8}$$

But the reverse inequality must hold because of (b) and the optimality of (x^0, y^0) to (2.1). Consequently, equality holds in (2.8). For $\theta > \theta_0$, if i is such that $z_i^0 > 0$, then

$$(\, q + Nx^0 + My^0 \,) \, \leq \, \theta_0 \, z_i^0 \, < \theta \, z_i^0,$$

and complementary slackness implies that $(\widehat{u}^+)_i = 0$. Similarly, we can show that $z_i^0 = 0 \Rightarrow v_i = 0$. Hence (c) follows. ∎

2.1. The parameter-free dual programs. Property (c) of Proposition 2.1 suggests that the inequality constraint $z^T u^+ + (\mathbf{1} - z)^T v \leq 0$, or, equivalently, the equality constraint $z^T u^+ + (\mathbf{1} - z)^T v = 0$ (because all variables are nonnegative and $z \in \{0,1\}^m$), should have an important role to play in an IP approach to the LPCC. This motivates us to define two value functions on the binary vectors. Specifically, for any $z \in \{0,1\}^m$, define

$$\begin{aligned} \Re \cup \{ \pm\infty \} \ni \varphi(z) \equiv \quad & \underset{(\lambda, u^{\pm}, v)}{\text{maximum}} \quad f^T\lambda + q^T(\, u^+ - u^- \,) \\ & \text{subject to} \quad A^T\lambda - N^T(\, u^+ - u^- \,) \, = \, c, \\ & \qquad\qquad\quad B^T\lambda - M^T(\, u^+ - u^- \,) - v \, \leq \, d, \\ & \qquad\qquad\quad (\, \lambda, u^{\pm}, v \,) \, \geq \, 0, \\ & \text{and} \qquad\quad z^T u^+ + (\mathbf{1} - z)^T v \, \leq \, 0 \end{aligned} \tag{2.9}$$

and its homogenization:

$$
\begin{array}{rlll}
\{\,0,\infty\,\} \ni \varphi_0(z) \equiv & \underset{(\lambda,u^{\pm},v)}{\text{maximum}} & f^T\lambda + q^T(\,u^+ - u^-\,) \\
& \text{subject to} & A^T\lambda - N^T(\,u^+ - u^-\,) = 0, \\
& & B^T\lambda - M^T(\,u^+ - u^-\,) - v \le 0, \\
& & (\,\lambda, u^{\pm}, v\,) \ge 0, \\
& \text{and} & z^T u^+ + (\,\mathbf{1} - z\,)^T v \le 0.
\end{array} \tag{2.10}
$$

Clearly, (2.10) is always feasible, and $\varphi_0(z)$ takes on the values 0 or ∞ only. Unlike (2.10), which is independent of the pair (c,d), (2.9) depends on (c,d) and is not guaranteed to be feasible; thus $\varphi(z) \in \Re \cup \{\pm\infty\}$. For any pair (c,d) for which (2.9) is feasible, we have

$$\varphi(z) < \infty \;\Leftrightarrow\; \varphi_0(z) = 0.$$

To this equivalence we add the following proposition that describes a one-to-one correspondence between (2.10) and the feasible pieces of the LPCC. The *support* of a vector z, denoted $\mathrm{supp}(z)$, is the index set of the nonzero components of z.

PROPOSITION 2.2. *For any $z \in \{0,1\}^m$, $\varphi_0(z) = 0$ if and only if the* LP(α) *is feasible, where $\alpha \equiv \mathrm{supp}(z)$.*

Proof. The dual of (2.10) is

$$
\begin{array}{ll}
\underset{(x,y)}{\text{minimize}} & 0^T x + 0^T y \\
\text{subject to} & Ax + By \ge f, \\
& \theta\, z \ge q + Nx + My \ge 0, \\
\text{and} & \theta\,(\,\mathbf{1} - z\,) \ge y \ge 0.
\end{array} \tag{2.11}
$$

By LP duality, it follows that if $\varphi_0(z) = 0$, then (2.11) is feasible for any $\theta > 0$; conversely, if (2.11) is feasible for some $\theta > 0$, then $\varphi_0(z) = 0$. In turn, (2.11) is feasible for some $\theta > 0$ if and only if the LP(α) is feasible for $\alpha \equiv \mathrm{supp}(z)$. □

For subsequent purposes, it would be useful to record the following equivalence between the extreme points/rays of the feasible region of (2.9) and those of the feasible set Ξ.

PROPOSITION 2.3. *For any $z \in [0,1]^m$, a feasible solution $(\lambda^p, u^{\pm,p}, v^p)$ of* (2.9) *is an extreme point in this region if and only if it is extreme in Ξ; a feasible ray $(\lambda^r, u^{\pm,r}, v^r)$ of* (2.9) *is extreme in this region if and only if it is extreme in Ξ.*

Proof. We prove only the first assertion; that for the second is similar. The sufficiency holds because the feasible region of (2.9) is a subset of Ξ. To prove the converse, suppose that $(\lambda^p, u^{\pm,p}, v^p)$ is an extreme solution of (2.9). Then this triple must be an element of Ξ. If it lies on the line segment of two other feasible solutions of Ξ, then the latter two solutions must satisfy the additional constraint $z^T u^+ + (\mathbf{1} - z)^T v \le 0$. Therefore, $(\lambda^p, u^{\pm,p}, v^p)$ is also extreme in Ξ. □

2.2. The set $\mathcal{Z}$ and a minimax formulation. We now define the key set of binary vectors:

$$\mathcal{Z} \equiv \{\, z \in \{\,0,1\,\}^m : \varphi_0(z) = 0\,\},$$

which, by Proposition 2.2, is the feasibility descriptor of the feasible region of the LPCC (2.1). Note that $\mathcal{Z}$ is a finite set. We also define the minimax integer program:

$$\text{(2.12)}\qquad \underset{z\in\mathcal{Z}}{\text{minimize}}\ \varphi(z) \equiv \left[\begin{array}{ll} \underset{(\lambda,u^{\pm},v)}{\text{maximum}} & f^T\lambda + q^T(\,u^+ - u^-\,) \\ \text{subject to} & A^T\lambda - N^T(\,u^+ - u^-\,) \,=\, c, \\ & B^T\lambda - M^T(\,u^+ - u^-\,) - v \,\le\, d, \\ & (\,\lambda, u^{\pm}, v\,) \,\ge\, 0, \\ \text{and} & z^T u^+ + (\mathbf{1} - z)^T v \,\le\, 0 \end{array}\right].$$

Since $\mathcal{Z}$ is a finite set, and since $\varphi(z) \in \Re \cup \{-\infty\}$ for $z \in \mathcal{Z}$, it follows that $\operatorname{argmin}_{z\in\mathcal{Z}} \varphi(z) \neq \emptyset$ if and only if $\mathcal{Z} \neq \emptyset$. The following result rephrases the three basic facts connecting the LPCC (2.1) and its LP pieces in terms of the IP (2.12).

THEOREM 2.4. *The following three statements hold:*

(a) *The LPCC* (2.1) *is infeasible if and only if* $\min_{z\in\mathcal{Z}} \varphi(z) = \infty$ *(i.e.,* $\mathcal{Z} = \emptyset$*);*

(b) *the LPCC* (2.1) *is feasible and has an unbounded objective value if and only if* $\min_{z\in\mathcal{Z}} \varphi(z) = -\infty$ *(i.e.,* $z \in \mathcal{Z}$ *exists such that* $\varphi(z) = -\infty$*);*

(c) *the LPCC* (2.1) *attains a finite optimal objective value if and only if* $-\infty < \min_{z\in\mathcal{Z}} \varphi(z) < \infty$.

In all cases, $\text{LPCC}_{\min} = \min_{z\in\mathcal{Z}} \varphi(z)$*; moreover, for any* $z \in \{0,1\}^m$ *for which* $\varphi(z) > -\infty$, $\text{LPCC}_{\min} \le \varphi(z)$.

Proof. Statement (a) is an immediate consequence of Proposition 2.2. Statement (b) is equivalent to saying that the LPCC (2.1) is feasible and has an unbounded objective if and only if $z \in \{0,1\}^m$ exists such that $\varphi_0(z) = 0$ and $\varphi(z) = -\infty$. Suppose that the LPCC (2.1) is feasible and unbounded. Then an index set $\alpha \subseteq \{1,\ldots,m\}$ exists such that the LP(α) is feasible and unbounded. By letting $z \in \{0,1\}^m$ be such that $\operatorname{supp}(z) = \alpha$ and $\bar{\alpha}$ be the complement of α in $\{1,\ldots,m\}$, we have $\varphi_0(z) = 0$. Moreover, the dual of the (unbounded) LP(α) is

$$\text{(2.13)}\qquad \left[\begin{array}{ll} \underset{(\lambda,u_{\bar{\alpha}},u_{\alpha}^-)}{\text{maximize}} & f^T\lambda + (\,q_{\bar{\alpha}}\,)^T u_{\bar{\alpha}} - (\,q_{\alpha}\,)^T u_{\alpha}^- \\ \text{subject to} & A^T\lambda - (\,N_{\bar{\alpha}\bullet}\,)^T u_{\bar{\alpha}} + (\,N_{\alpha\bullet}\,)^T u_{\alpha}^- \,=\, c, \\ & (\,B_{\bullet\bar{\alpha}}\,)^T\lambda - (\,M_{\bar{\alpha}\bar{\alpha}}\,)^T u_{\bar{\alpha}} + (\,M_{\alpha\bar{\alpha}}\,)^T u_{\alpha}^- \,\le\, d_{\bar{\alpha}}, \\ \text{and} & (\,\lambda, u_{\alpha}^-\,) \,\ge\, 0, \end{array}\right.$$

which is equivalent to the problem (2.9) corresponding to the binary vector z defined here. (Note that the $\bullet$ in the subscripts is the standard notation in linear programming, denoting rows/columns of matrices.) Therefore, since (2.13) is infeasible, it follows that $\varphi(z) = -\infty$ by convention. Conversely, suppose that $z \in \{0,1\}^m$ exists such that $\varphi_0(z) = 0$ and $\varphi(z) = -\infty$. Let $\alpha \equiv \operatorname{supp}(z)$ and $\bar{\alpha} \equiv$ complement of α in $\{1,\ldots,m\}$. It then follows that (2.11), and thus the LP(α), is feasible. Moreover, since $\varphi(z) = -\infty$, it follows that (2.13), being equivalent to (2.9), is infeasible; thus the LP(α) is unbounded. Statement (c) follows readily from (a) and (b). The equality between $\text{LPCC}_{\min}$ and $\min_{z\in\mathcal{Z}} \varphi(z)$ is due to the fact that the maximizing LP defining $\varphi(z)$ is essentially the dual of the piece LP(α). To prove the last assertion of the theorem, let $z \in \{0,1\}^m$ be such that $\varphi(z) > -\infty$. Without loss of generality, we may assume that $\varphi(z) < \infty$. Thus the LP (2.9) attains a finite maximum; hence $\varphi_0(z) = 0$. Therefore $z \in \mathcal{Z}$, and the bound $\text{LPCC}_{\min} \le \varphi(z)$ holds readily. ∎

3. The Benders approach. In essence, our strategy for solving the LPCC (2.1) is to apply a Benders approach to the minimax IP (2.12). For this purpose, we let $\{(\lambda^{p,i}, u^{\pm,p,i}, v^{p,i})\}_{i=1}^{K}$ and $\{(\lambda^{r,j}, u^{\pm,r,j}, v^{r,j})\}_{j=1}^{L}$ be the finite set of extreme points and extreme rays of the polyhedron Ξ. Note that $K \geq 1$ if and only if $\Xi \neq \emptyset$. (These extreme points and rays will be generated as needed. For the discussion in this section, we take them as available.) In what follows, we derive a restatement of Theorem 2.4 in terms of these extreme points and rays.

The IP (2.12) can be written as

$$\underset{z \in \mathcal{Z}}{\text{minimize}} \left[\begin{array}{ll} \underset{(\rho^p, \rho^r) \geq 0}{\text{maximum}} & \displaystyle\sum_{i=1}^{K} \rho_i^p \left[f^T \lambda^{p,i} + q^T (u^{+,p,i} - u^{-,p,i}) \right] \\ & \qquad + \displaystyle\sum_{j=1}^{L} \rho_j^r \left[f^T \lambda^{r,j} + q^T (u^{+,r,j} - u^{-,r,j}) \right] \\ \text{subject to} & \displaystyle\sum_{i=1}^{K} \rho_i^p \left[z^T u^{+,p,i} + (\mathbf{1} - z)^T v^{p,i} \right] \\ & \qquad + \displaystyle\sum_{j=1}^{L} \rho_j^r \left[z^T u^{+,r,j} + (\mathbf{1} - z)^T v^{r,j} \right] \leq 0 \\ \text{and} & \displaystyle\sum_{i=1}^{K} \rho_i^p = 1 \end{array} \right], \tag{3.1}$$

which is the *master IP*. It turns out that the set $\mathcal{Z}$ can be completely described in terms of certain *ray cuts*, whose definition requires the index set:

$$\mathcal{L} \equiv \left\{ j \in \{1, \ldots, L\} : f^T \lambda^{r,j} + q^T (u^{+,r,j} - u^{-,r,j}) > 0 \right\}.$$

The following proposition shows that the set $\mathcal{Z}$ can be described in terms of satisfiability inequalities by using the extreme rays in $\mathcal{L}$.

PROPOSITION 3.1.

$$\mathcal{Z} = \left\{ z \in \{0,1\}^m : \sum_{\ell : u_\ell^{+,r,j} > 0} z_\ell + \sum_{\ell : v_\ell^{r,j} > 0} (1 - z_\ell) \geq 1 \ \ \forall j \in \mathcal{L} \right\}.$$

Proof. Since a tuple $(\lambda, u^{\pm}, v)$ is feasible to (2.10) if and only if it is a nonnegative combination of the extreme rays of (2.9), which are necessarily extreme rays of Ξ by Proposition 2.3, it follows that a tuple $(\lambda, u^{\pm}, v)$ is feasible to (2.10) if and only if there exist nonnegative coefficients $\{\rho_j^r\}_{j=1}^{L}$ such that

$$(\lambda, u^{\pm}, v) = \sum_{j=1}^{L} \rho_j^r (\lambda^{r,j}, u^{\pm,r,j}, v^{r,j})$$

and $\sum_{j=1}^{L} \rho_j^r [z^T u^{+,r,j} + (\mathbf{1} - z)^T v^{r,j}] \leq 0$. Therefore, $\varphi_0(z)$ is equal to

$$\begin{array}{ll} \underset{\rho^r \geq 0}{\text{maximize}} & \displaystyle\sum_{j=1}^{L} \rho_j^r \left[f^T \lambda^{r,j} + q^T (u^{+,r,j} - u^{-,r,j}) \right] \\ \text{subject to} & \displaystyle\sum_{j=1}^{L} \rho_j^r \left[z^T u^{+,r,j} + (\mathbf{1} - z)^T v^{r,j} \right] \leq 0, \end{array}$$

and the latter maximization problem has a finite optimal solution if and only if

$$f^T\lambda^{r,j} + q^T(\,u^{+,r,j} - u^{-,r,j}\,) \;>\; 0 \quad\Longrightarrow\quad z^T u^{+,r,j} + (\mathbf{1}-z)^T v^{r,j} \;>\; 0$$
$$\Longleftrightarrow\quad \sum_{\ell:u_\ell^{+,r,j}>0} z_\ell + \sum_{\ell:v_\ell^{r,j}>0} (\,1-z_\ell\,) \;\geq\; 1.$$

Therefore, the equality between $\mathcal{Z}$ and the right-hand set is immediate. □

An immediate corollary of Proposition 3.1 is that it provides a certificate of infeasibility for the LPCC.

COROLLARY 3.2. *If $\mathcal{R} \subseteq \mathcal{L}$ exists such that*

$$\left\{ z \in \{0,1\}^m \,:\, \sum_{\ell:u_\ell^{+,r,j}>0} z_\ell + \sum_{\ell:v_\ell^{r,j}>0} (\,1-z_\ell\,) \;\geq\; 1 \;\; \forall\, j \;\in\; \mathcal{R} \right\} = \emptyset,$$

then the LPCC (2.1) *is infeasible.*

Proof. The assumption implies that $\mathcal{Z} = \emptyset$. Thus the infeasibility of the LPCC follows from Theorem 2.4(a). □

In view of Proposition 3.1, (3.1) is equivalent to:

$$\underset{z\in\mathcal{Z}}{\text{minimize}} \left[\begin{array}{ll} \underset{\rho^p\geq 0}{\text{maximum}} & \displaystyle\sum_{i=1}^{K} \rho_i^p \left[f^T\lambda^{p,i} + q^T(u^{+,p,i} - u^{-,p,i}) \right] \\ \text{subject to} & \displaystyle\sum_{i=1}^{K} \rho_i^p \left[\, z^T u^{+,p,i} + (\mathbf{1}-z)^T v^{p,i} \,\right] \;\leq\; 0 \\ \text{and} & \displaystyle\sum_{i=1}^{K} \rho_i^p \;=\; 1 \end{array}\right]. \tag{3.2}$$

Note that the $\text{LPCC}_{\min}$ is equal to the minimum objective value of (3.2). Similar to the inequality

$$\sum_{\ell:u_\ell^{+,r,j}>0} z_\ell + \sum_{\ell:v_\ell^{r,j}>0} (\,1-z_\ell\,) \;\geq\; 1,$$

which we call a *ray cut* (because it is induced by an extreme ray), we will make use of a *point cut*

$$\sum_{\ell:u_\ell^{+,p,i}>0} z_\ell + \sum_{\ell:v_\ell^{p,i}>0} (\,1-z_\ell\,) \;\geq\; 1,$$

that is induced by an extreme point $(\lambda^{p,i}, u^{\pm,p,i}, v^{p,i})$ of Ξ chosen from the following collection:

$$\mathcal{K} \equiv \left\{\, i \;\in\; \{\,1,\ldots,K\,\} \,:\, f^T\lambda^{p,i} + q^T(\,u^{+,p,i} - u^{-,p,i}\,) \;=\; \varphi(z) \text{ for some } z \in \mathcal{Z} \,\right\}.$$

Note that $\mathcal{K} \neq \emptyset \Rightarrow \mathcal{Z} \neq \emptyset$, which in turn implies that the LPCC (2.1) is feasible. Moreover,

$$\min_{i\in\mathcal{K}} \left[\, f^T\lambda^{p,i} + q^T(\,u^{+,p,i} - u^{-,p,i}\,) \,\right] \;\geq\; \text{LPCC}_{\min}.$$

For a given pair of subsets $\mathcal{P} \times \mathcal{R} \subseteq \mathcal{K} \times \mathcal{L}$, let

$$\mathcal{Z}(\mathcal{P},\mathcal{R}) \equiv \left\{ z \in \{0,1\}^m : \begin{array}{c} \sum\limits_{\ell: u_\ell^{+,r,j}>0} z_\ell + \sum\limits_{\ell: v_\ell^{r,j}>0} (1-z_\ell) \geq 1 \ \forall j \in \mathcal{R} \\ \sum\limits_{\ell: u_\ell^{+,p,i}>0} z_\ell + \sum\limits_{\ell: v_\ell^{p,i}>0} (1-z_\ell) \geq 1 \ \forall i \in \mathcal{P} \end{array} \right\}.$$

We have the following result.

PROPOSITION 3.3. *If there exists $\mathcal{P} \times \mathcal{R} \subseteq \mathcal{K} \times \mathcal{L}$ such that*

$$\min_{i \in \mathcal{P}} \left[f^T \lambda^{p,i} + q^T (u^{+,p,i} - u^{-,p,i}) \right] > \mathrm{LPCC}_{\min},$$

then $\operatorname{argmin}_{z \in \mathcal{Z}} \varphi(z) \subseteq \mathcal{Z}(\mathcal{P},\mathcal{R})$.

Proof. Let $\widetilde{z} \in \mathcal{Z}$ be a minimizer of $\varphi(z)$ on $\mathcal{Z}$. (The proposition is clearly valid if no such minimizer exists.) If $\widetilde{z} \notin \mathcal{Z}(\mathcal{P},\mathcal{R})$, then there exists $i \in \mathcal{P}$ such that

$$\sum_{\ell: u_\ell^{+,p,i}>0} \widetilde{z}_\ell + \sum_{\ell: v_\ell^{p,i}>0} (1-\widetilde{z}_\ell) = 0.$$

Hence, $(\lambda^{p,i}, u^{\pm,p,i}, v^{p,i})$ is feasible to the LP (2.9) corresponding to $\varphi(\widetilde{z})$; thus

$$\mathrm{LPCC}_{\min} = \varphi(\widetilde{z}) \geq f^T \lambda^{p,i} + q^T (u^{+,p,i} - u^{-,p,i}) > \mathrm{LPCC}_{\min},$$

which is a contradiction. ∎

Analogous to Corollary 3.2, we have the following corollary of Proposition 3.3.

COROLLARY 3.4. *If there exists $\mathcal{P} \times \mathcal{R} \subseteq \mathcal{K} \times \mathcal{L}$, with $\mathcal{P} \neq \emptyset$, such that $\mathcal{Z}(\mathcal{P},\mathcal{R}) = \emptyset$, then*

$$\mathrm{LPCC}_{\min} = \min_{i \in \mathcal{P}} \left[f^T \lambda^{p,i} + q^T (u^{+,p,i} - u^{-,p,i}) \right] \in (-\infty, \infty). \tag{3.3}$$

Proof. Indeed, if the claimed equality does not hold, then $\operatorname{argmin}_{z \in \mathcal{Z}} \varphi(z) = \emptyset$. But this implies that $\mathcal{Z} = \emptyset$, which contradicts the assumption that $\mathcal{P} \neq \emptyset$. ∎

Combining Corollaries 3.2 and 3.4, we obtain the desired restatement of Theorem 2.4 in terms of the extreme points and rays of Ξ.

THEOREM 3.5. *The following three statements hold:*

(a) *The LPCC* (2.1) *is infeasible if and only if a subset $\mathcal{R} \subseteq \mathcal{L}$ exists such that $\mathcal{Z}(\emptyset,\mathcal{R}) = \emptyset$;*
(b) *the LPCC* (2.1) *is feasible and has an unbounded objective if and only if $\mathcal{Z}(\mathcal{K},\mathcal{L}) \neq \emptyset$;*
(c) *the LPCC* (2.1) *attains a finite optimal objective value if and only if a pair $\mathcal{P} \times \mathcal{R} \subseteq \mathcal{K} \times \mathcal{L}$ exists, with $\mathcal{P} \neq \emptyset$, such that $\mathcal{Z}(\mathcal{P},\mathcal{R}) = \emptyset$.*

Proof. Statement (a) follows from Corollary 3.2 by noting that a subset $\mathcal{R} \subseteq \mathcal{L}$ exists such that $\mathcal{Z}(\emptyset,\mathcal{R}) = \emptyset$ if and only if $\mathcal{Z} = \mathcal{Z}(\emptyset,\mathcal{L}) = \emptyset$. To prove (b), suppose first that $\mathcal{Z}(\mathcal{K},\mathcal{L}) \neq \emptyset$. Let $\widehat{z} \in \mathcal{Z}(\mathcal{K},\mathcal{L})$. Then $\widehat{z} \in \mathcal{Z}$. We claim that $\varphi(\widehat{z}) = -\infty$; i.e., the LP (2.9) corresponding to $\widehat{z}$ is infeasible. Assume otherwise, and then since $\varphi_0(\widehat{z}) = 0$, it follows that $\varphi(\widehat{z})$ is finite. Hence there exists an extreme point $(\lambda^{p,i}, u^{\pm,p,i}, v^{p,i})$ of the LP (2.9) corresponding to $\widehat{z}$ such that $f^T \lambda^{p,i} + q^T (u^{+,p,i} - u^{-,p,i}) = \varphi(\widehat{z})$; thus the index $i \in \mathcal{K}$, which implies that

$$\sum_{\ell: u_\ell^{+,p,i}>0} \widehat{z}_\ell + \sum_{\ell: v_\ell^{p,i}>0} (1-\widehat{z}_\ell) \geq 1,$$

because $\widehat{z} \in \mathcal{Z}(\mathcal{K}, \mathcal{L})$. But this contradicts the feasibility of $(\lambda^{p,i}, u^{\pm,p,i}, v^{p,i})$ to the LP (2.9) corresponding to $\widehat{z}$. Therefore, the LPCC (2.1) is feasible and has an unbounded objective value; thus, the "if" statement in (b) holds. Conversely, suppose that $\text{LPCC}_{\min} = -\infty$. By Theorem 2.4, it follows that $\widehat{z} \in \mathcal{Z}$ exists such that $\varphi(\widehat{z}) = -\infty$; i.e., the LP (2.9) corresponding to $\widehat{z}$ is infeasible. In turn, this means that

$$\widehat{z}^T u^{+,p,i} + (\mathbf{1} - \widehat{z})^T v^{p,i} > 0$$

for all $i = 1, \ldots, K$, or, equivalently,

$$\sum_{\ell : u_\ell^{+,p,i} > 0} \widehat{z}_\ell + \sum_{\ell : v_\ell^{p,i} > 0} (1 - \widehat{z}_\ell) \geq 1$$

for all $i = 1, \ldots, K$. Consequently, $\widehat{z} \in \mathcal{Z}(\mathcal{K}, \mathcal{L})$. Hence, statement (b) holds. Finally, the "if" statement in (c) follows from Corollary 3.4. Conversely, if the LPCC (2.1) has a finite optimal solution, then by (b), it follows that $\mathcal{Z}(\mathcal{K}, \mathcal{L}) = \emptyset$. Since the LPCC (2.1) is feasible, $\mathcal{K} \neq \emptyset$ by (a), establishing the "only if" statement in (c). □

Theorem 3.5 constitutes the theoretical basis for the algorithm to be presented in section 5 for resolving the LPCC. Through the successive generation of extreme points and rays of Ξ, the algorithm searches for a pair of subsets $\mathcal{P} \times \mathcal{R}$ such that $\mathcal{Z}(\mathcal{P}, \mathcal{R}) = \emptyset$. If such a pair can be successfully identified, then the LPCC is either infeasible ($\mathcal{P} = \emptyset$) or attains a finite optimal solution ($\mathcal{P} \neq \emptyset$). If no such pair is found, then the LPCC is unbounded. In the algorithm, the last case is identified with a binary vector $z \in \mathcal{Z}$, with $\varphi(z) = -\infty$; i.e., the LP (2.9) is infeasible. Based on the value function $\varphi(z)$ and the point/ray cuts, the algorithm will be shown to terminate in finite time.

4. Simple cuts and sparsification. In this section, we explain several key steps in the main algorithm to be presented in the next section. The first idea is a version of the well-known Gomory cut in integer programming specialized to the LPCC and which has previously been employed for bilevel LPs; see [5]. The second idea aims at "sparsifying" the ray/point cuts to facilitate the computation of elements of the working sets $\mathcal{Z}(\mathcal{P}, \mathcal{R})$. Specifically, a satisfiability constraint

$$\sum_{i \in \mathcal{I}'} z_i + \sum_{j \in \mathcal{J}'} (1 - z_j) \geq 1 \quad \text{is } \textit{sparser} \text{ than} \quad \sum_{i \in \mathcal{I}} z_i + \sum_{j \in \mathcal{J}} (1 - z_j) \geq 1$$

if $\mathcal{I}' \subseteq \mathcal{I}$ and $\mathcal{J}' \subseteq \mathcal{J}$. In general, a satisfiability inequality cuts off certain LP pieces of the LPCC; the sparser the inequality is, the more pieces it cuts off. Thus, it is desirable to sparsify a cut as much as possible. Nevertheless, sparsification requires the solution of linear subprograms; thus one needs to balance the work required with the benefit of the process.

4.1. Simple cuts. The following discussion is a minor variant of that presented in [5] for bilevel LPs. Consider the LP relaxation of the LPCC (2.1):

$$\begin{array}{lll} \underset{(x,y,w)}{\text{minimize}} & c^T x + d^T y & \\ \text{subject to} & Ax + By \geq f & \\ \text{and} & 0 \leq y, \quad w \equiv q + Nx + My \geq 0, & \end{array} \tag{4.1}$$

where the orthogonal condition $y^T w = 0$ is dropped. Assume that, by solving this LP, an optimal solution is obtained that fails the latter orthogonality condition, say,

$y_i w_i > 0$ in this solution. Thus, y_i and w_i must be basic variables in a basic optimal solution of the LP; in such a solution, w_i and y_i can be expressed in terms of the nonbasic variables, which we denote by the generic variables s_j, as follows: For some constants a_j and b_j,

$$w_i \;=\; w_{i0} - \sum_{s_j:\text{nonbasic}} a_j\, s_j \quad \text{and} \quad y_i \;=\; y_{i0} - \sum_{s_j:\text{nonbasic}} b_j\, s_j,$$

where w_{i0} and y_{i0} are the current values of the variables w_i and y_i, respectively, with $\min(w_{i0}, y_{i0}) > 0$. It is not difficult to show that the following inequality must be satisfied by all feasible solutions of the LPCC (2.1):

$$\sum_{\substack{s_j:\text{nonbasic} \\ \max(a_j,b_j)>0}} \max\left(\frac{a_j}{w_{i0}}, \frac{b_j}{y_{i0}}\right) s_j \;\geq\; 1. \tag{4.2}$$

Note that if $a_j \leq 0$ for all nonbasic j, then $w_i > 0 = y_i$ for every feasible solution of the LPCC (2.1). A similar remark can be made if $b_j \leq 0$ for all nonbasic j.

Following the terminology in [5], we call the inequality (4.2) a *simple cut*. Multiple such cuts can be added to the constraint $Ax + By \geq f$, resulting in a modified inequality $\widetilde{A}x + \widetilde{B}y \geq \widetilde{f}$. We can generate and add even more simple cuts by repeating the above step. This strategy turns out to be a very effective preprocessor for the algorithm to be described in the next section. At the end of this preprocessor, we obtain an optimal solution $(\bar{x}, \bar{y}, \bar{w})$ of (4.1) that remains infeasible to the LPCC (otherwise, this solution would be optimal for the LPCC); the optimal objective value $c^T\bar{x} + d^T\bar{y}$ provides a valid lower bound for $\text{LPCC}_{\min}$. (Note that if (4.1) is unbounded, then the preprocessor does not produce any cuts or a finite lower bound.)

LPCC feasibility recovery. Occurring in many applications of the LPCC, the special case $B = 0$ deserves a bit more discussion. First note that in this case the modified matrix $\widetilde{B}$ is not necessarily zero. Nevertheless, the solution $(\bar{x}, \bar{y}, \bar{w})$ obtained from the simple-cut preprocessor can be used to produce a feasible solution to the LPCC (2.1) by simply solving the linear complementarity problem (LCP): $0 \leq y \perp q + N\bar{x} + My \geq 0$ (assuming that the matrix M has favorable properties so that this step is effective). By letting $\bar{y}'$ be a solution to the latter LCP, the objective value $c^T\bar{x} + d^T\bar{y}'$ yields a valid upper bound to $\text{LPCC}_{\min}$. This recovery procedure of an LPCC feasible solution can be extended to the case where $B \neq 0$. (Incidentally, this class of LPCCs is generally "more difficult" than the class where $B = 0$, where the difficulty is determined by our empirical experience from the computational tests.) Indeed, from any feasible solution $(\bar{x}, \bar{y}, \bar{w})$ to the LP relaxation of the LPCC (2.1) but not to the LPCC itself, we could attempt to recover a feasible solution to the LPCC along with an element in $\mathcal{Z}$ either by solving the LP(α), where $\alpha \equiv \{i : \bar{y}_i \leq \bar{w}_i\}$, or by solving $\varphi(z)$, where $z_\alpha = 1$ and $z_{\bar{\alpha}} = 0$. A feasible solution to this LP piece yields a feasible solution to the LPCC and a finite upper bound. In general, there is no guarantee that this procedure will always be successful; nevertheless, it is very effective when it works.

4.2. Cut management. A key step in our algorithm involves the selection of elements in the sets $\mathcal{Z}(\mathcal{P}, \mathcal{R})$ for various index pairs $(\mathcal{P}, \mathcal{R})$. Generally speaking, this involves solving integer subprograms. By recognizing that the constraints in each $\mathcal{Z}(\mathcal{P}, \mathcal{R})$ are of the satisfiability type, we could in principle employ special algorithms

for implementing this step (see [7, 31], and the references therein for some such algorithms). To facilitate such a selection, we have developed a special heuristic that utilizes a valid upper bound of $\mathrm{LPCC}_{\min}$ to sparsify the terms in the ray/point cuts in a working set. In what follows, we describe how the algorithm manages these cuts.

There are three pools of cuts, labeled $\mathcal{Z}_{\mathrm{work}}$, the working pool, $\mathcal{Z}_{\mathrm{wait}}$, the wait pool, and $\mathcal{Z}_{\mathrm{cand}}$, the candidate pool. Inequalities in $\mathcal{Z}_{\mathrm{work}}$ are valid sparsifications of those in $\mathcal{Z}(\mathcal{P}, \mathcal{R})$ corresponding to a current pair $(\mathcal{P}, \mathcal{R})$. Thus, the set of binary vectors satisfying the inequalities in $\mathcal{Z}_{\mathrm{work}}$, which we denote $\widehat{\mathcal{Z}}_{\mathrm{work}}$, is a subset of $\mathcal{Z}(\mathcal{P}, \mathcal{R})$. Inequalities in $\mathcal{Z}_{\mathrm{cand}}$ are candidates for sparsification; the sparsification procedure described below always ends with this set empty. The decision of whether or not to sparsify a valid inequality is made according to a current LPCC upper bound and a small scalar $\delta > 0$. In essence, the sparsification is an effective way to facilitate the search for a feasible element in $\widehat{\mathcal{Z}}_{\mathrm{work}}$. At one extreme, a sparsest inequality with only one term in it automatically fixes one complementarity (e.g., $z_1 \geq 1$ fixes $w_1 = 0$); at another extreme, it is computationally more difficult to find feasible points satisfying many dense inequalities.

We sparsify an inequality

$$\sum_{i \in \mathcal{I}} z_i + \sum_{j \in \mathcal{J}} (\, 1 - z_j \,) \geq 1 \tag{4.3}$$

in the following way. Let $\mathcal{I} = \mathcal{I}_1 \cup \mathcal{I}_2$ be a partition of $\mathcal{I}$ into two disjoint subsets $\mathcal{I}_1$ and $\mathcal{I}_2$; similarly, let $\mathcal{J} = \mathcal{J}_1 \cup \mathcal{J}_2$. We split (4.3), which we call the *parent*, into two subinequalities:

$$\sum_{i \in \mathcal{I}_1} z_i + \sum_{j \in \mathcal{J}_1} (\, 1 - z_j \,) \geq 1 \quad \text{and} \quad \sum_{i \in \mathcal{I}_2} z_i + \sum_{j \in \mathcal{J}_2} (\, 1 - z_j \,) \geq 1 \tag{4.4}$$

and test both to see if they are valid for the LPCC. To test the left-hand inequality, we consider the LP relaxation (4.1) of the LPCC (2.1) with the additional constraints $w_i = (q + Nx + My)_i = 0$ for $i \in \mathcal{I}_1$ and $y_i = 0$ for $i \in \mathcal{J}_1$, which we call a *relaxed LP with restriction*. If this LP has an objective value greater than the current $\mathrm{LPCC}_{\mathrm{ub}}$, then we have successfully sparsified the inequality (4.3) into the sparser inequality:

$$\sum_{i \in \mathcal{I}_1} z_i + \sum_{j \in \mathcal{J}_1} (\, 1 - z_j \,) \geq 1, \tag{4.5}$$

which must be valid for the LPCC. (In this situation, any dual solution to the relaxed LP with restriction is feasible in the dual LP (2.9) for any binary vector z that violates (4.5). Hence, the value $\varphi(z)$ of the LPCC on this piece must be at least $\mathrm{LPCC}_{\mathrm{ub}}$, implying that such a piece cannot contain an optimal solution of the LPCC.) Otherwise, by using the feasible solution to the relaxed LP, we employ the LPCC feasibility recovery procedure to compute an LPCC feasible solution along with a binary $z \in \mathcal{Z}$. If successful, one of two cases happen: If $\varphi(z) \geq \mathrm{LPCC}_{\mathrm{ub}}$, then a new cut can be generated; otherwise, we have reduced the LPCC upper bound. In either case, we obtain positive progress in the algorithm. If no LPCC feasible solution is recovered, then we save the cut (4.5) in the wait pool $\mathcal{Z}_{\mathrm{wait}}$ for later consideration. In essence, cuts in the wait pool are not yet proven to be valid for the LPCC; they will be revisited when there is a reduction in $\mathrm{LPCC}_{\mathrm{ub}}$. Note that every inequality in $\mathcal{Z}_{\mathrm{wait}}$ has an LP optimal objective value associated with it that is less than the current LPCC upper bound.

In our experiment, we randomly divide the sets $\mathcal{I}$ and $\mathcal{J}$ roughly into two equal halves each and adopt a strategy that attempts to sparsify the *root inequality* (4.3) as much as possible via a random branching rule. The following illustrates one such division:

$$\begin{array}{ccc} & z_1 + z_3 + z_4 + (1 - z_2) + (1 - z_6) \geq 1 & \\ & \swarrow \qquad\qquad \searrow & \\ z_1 + z_3 + (1 - z_2) \geq 1 & & z_4 + (1 - z_6) \geq 1. \end{array}$$

We use a small scalar $\delta > 0$ to help decide on the subsequent branching. In essence, we branch only if the inequality appears strong. By solving LPs, the procedure below sparsifies a given valid inequality for the LPCC, called the *root* of the procedure.

SPARSIFICATION PROCEDURE. Let (4.3) be the root inequality to be sparsified, $\mathrm{LPCC}_{\mathrm{ub}}$ be the current LPCC upper bound, and $\delta > 0$ be a given scalar. Branch (4.3) into two subinequalities (4.4), both of which we put in the set $\mathcal{Z}_{\mathrm{cand}}$.

Main step. If $\mathcal{Z}_{\mathrm{cand}}$ is empty, terminate. Otherwise pick a *candidate* inequality in $\mathcal{Z}_{\mathrm{cand}}$, say, the left one in (4.4) with the corresponding pair of index sets $(\mathcal{I}_1, \mathcal{J}_1)$. Solve the LP relaxation (4.1) of the LPCC (2.1) with the additional constraints $w_i = (q + Nx + My)_i = 0$ for $i \in \mathcal{I}_1$ and $y_i = 0$ for $i \in \mathcal{J}_1$, obtaining an LP optimal objective value, say, $\mathrm{LP}_{\mathrm{rlx}} \in \Re \cup \{\pm\infty\}$. We have the following three cases:

- If $\mathrm{LP}_{\mathrm{rlx}} \in [\,\mathrm{LPCC}_{\mathrm{ub}}, \mathrm{LPCC}_{\mathrm{ub}} + \delta\,]$, move the candidate inequality from $\mathcal{Z}_{\mathrm{cand}}$ into $\mathcal{Z}_{\mathrm{work}}$ and remove its parent; return to the main step.
- If $\mathrm{LP}_{\mathrm{rlx}} < \mathrm{LPCC}_{\mathrm{ub}}$, apply the LPCC feasibility recovery procedure to the feasible solution at termination of the current relaxed LP with restriction. If the procedure is successful, return to the main step with either a new cut or a reduced $\mathrm{LPCC}_{\mathrm{ub}}$. Otherwise, move the incumbent candidate inequality from $\mathcal{Z}_{\mathrm{cand}}$ into $\mathcal{Z}_{\mathrm{wait}}$; return to the main step.
- If $\delta + \mathrm{LPCC}_{\mathrm{ub}} < \mathrm{LP}_{\mathrm{rlx}}$, move the candidate inequality from $\mathcal{Z}_{\mathrm{cand}}$ into $\mathcal{Z}_{\mathrm{work}}$ and remove its parent; further branch the candidate inequality into two subinequalities, both of which we put into the candidate pool $\mathcal{Z}_{\mathrm{cand}}$; return to the main step.

During the procedure, the set $\mathcal{Z}_{\mathrm{cand}}$ may grow from the initial size of 2 inequalities when the root of the procedure is first split. Nevertheless, by solving finitely many LPs, this set will eventually shrink to empty; when that happens, either we have successfully sparsified the root inequality and placed multiple sparser cuts into $\mathcal{Z}_{\mathrm{work}}$, or some sparser cuts are added to the pool $\mathcal{Z}_{\mathrm{wait}}$, waiting to be proven valid for the LPCC in subsequent iterations. Note that associated with each inequality in $\mathcal{Z}_{\mathrm{wait}}$ is the value $\mathrm{LP}_{\mathrm{rlx}}$.

5. The IP algorithm. We are now ready to present the parameter-free IP-based algorithm for resolving an arbitrary LPCC (2.1). Subsequently, we will establish that the algorithm will successfully terminate in a finite number of *iterations* with a definitive resolution of the LPCC in one of its three states. Referring to a return to Step 1, each iteration consists of solving one feasibility IP of the satisfiability kind, a couple of LPs to compute $\varphi(\widehat{z})$ and possibly $\varphi_0(\widehat{z})$ corresponding to a binary vector $\widehat{z}$ obtained from the IP, and multiple LPs within the sparsification procedure associated with an induced point/ray cut.

ALGORITHM.
Step 0 (preprocessing and initialization). Generate multiple simple cuts to tighten the complementarity constraints. If any of the LPs encountered in this step is infeasible, then so is the LPCC (2.1). In general, let LPCC_{lb} ($-\infty$ allowed) and LPCC_{ub} (∞ allowed) be valid lower and upper bounds of $\text{LPCC}_{\min}$, respectively. Let $\delta > 0$ be a small scalar. [A finite optimal solution to a relaxed LP provides a finite lower bound, and a feasible solution to the LPCC, which could be obtained by the LPCC feasibility recovery procedure, provides a finite upper bound.] Set $\mathcal{P} = \mathcal{R} = \emptyset$ and $\mathcal{Z}_{\text{work}} = \mathcal{Z}_{\text{wait}} = \emptyset$. (Thus, $\widehat{\mathcal{Z}}_{\text{work}} = \{0,1\}^m$.)
Step 1 (solving a satisfiability IP). Determine a vector $\widehat{z} \in \widehat{\mathcal{Z}}_{\text{work}}$. If this set is empty, go to Step 2. Otherwise go to Step 3.
Step 2 (termination: infeasibility or finite solvability). If $\mathcal{P} = \emptyset$, we have obtained a certificate of infeasibility for the LPCC (2.1); stop. If $\mathcal{P} \neq \emptyset$, we have obtained a certificate of global optimality for the LPCC (2.1) with $\text{LPCC}_{\min}$ given by (3.3); stop.
Step 3 (solving dual LP). Compute $\varphi(\widehat{z})$ by solving the LP (2.9). If $\varphi(\widehat{z}) \in (-\infty, \infty)$, go to Step 4a. If $\varphi(\widehat{z}) = \infty$, proceed to Step 4b. If $\varphi(\widehat{z}) = -\infty$, proceed to Step 4c.
Step 4a (adding an extreme point). Let $(\lambda^{p,i}, u^{\pm,p,i}, v^{p,i}) \in \mathcal{K}$ be an optimal extreme point of Ξ. There are 3 cases:

- If $\varphi(\widehat{z}) \in [\ \text{LPCC}_{\text{ub}},\ \text{LPCC}_{\text{ub}} + \delta]$, let $\mathcal{P} \leftarrow \mathcal{P} \cup \{i\}$, and add the corresponding point cut to $\mathcal{Z}_{\text{work}}$; return to Step 1.
- If $\varphi(\widehat{z}) > \text{LPCC}_{\text{ub}} + \delta$, let $\mathcal{P} \leftarrow \mathcal{P} \cup \{i\}$, and add the corresponding point cut to $\mathcal{Z}_{\text{work}}$. Apply the sparsification procedure to the new point cut, obtaining an updated $\mathcal{Z}_{\text{work}}$ and $\mathcal{Z}_{\text{wait}}$ and possibly a reduced LPCC_{ub}. If the LPCC upper bound is reduced during the sparsification procedure, go to Step 5 to activate some of the cuts in the wait pool; otherwise, return to Step 1.
- If $\varphi(\widehat{z}) < \text{LPCC}_{\text{ub}}$, let $\text{LPCC}_{\text{ub}} \leftarrow \varphi(\widehat{z})$, and go to Step 5.

Step 4b (adding an extreme ray). Let $(\lambda^{r,j}, u^{\pm,r,j}, v^{r,j}) \in \mathcal{L}$ be an extreme ray of Ξ. Set $\mathcal{R} \longleftarrow \mathcal{R} \cup \{j\}$, and add the corresponding ray cut to $\mathcal{Z}_{\text{work}}$. Apply the sparsification procedure to the new ray cut, obtaining an updated $\mathcal{Z}_{\text{work}}$ and $\mathcal{Z}_{\text{wait}}$ and possibly a reduced LPCC_{ub}. If the LPCC upper bound is reduced during the sparsification procedure, go to Step 5 to activate some of the cuts in the wait pool; otherwise, return to Step 1.
Step 4c (determining LPCC unboundedness). Solve the LP (2.10) to determine $\varphi_0(\widehat{z})$. If $\varphi_0(\widehat{z}) = 0$, then the vector $\widehat{z}$ and its support provide a certificate of unboundedness for the LPCC (2.1). Stop. If $\varphi_0(\widehat{z}) = \infty$, go to Step 4b.
Step 5 (LPCC_{ub} is reduced). Move all inequalities in $\mathcal{Z}_{\text{wait}}$ with values LP_{rlx} greater than (the just-reduced) LPCC_{ub} into $\mathcal{Z}_{\text{work}}$. Apply the sparsification procedure to each newly moved inequality with $\text{LP}_{\text{rlx}} > \text{LPCC}_{\text{ub}} + \delta$. Reapply this step to the cuts in $\mathcal{Z}_{\text{wait}}$ each time the LPCC upper bound is reduced from the sparsification procedure. Return to Step 1 when no more cuts in $\mathcal{Z}_{\text{wait}}$ are eligible for sparsification.

We have the following finiteness result.

THEOREM 5.1. *The algorithm terminates in a finite number of iterations.*

Proof. The finiteness is due to several observations: (a) The set of m-dimensional binary vectors is finite, (b) each iteration of the algorithm generates a new binary vector that is distinct from all of those previously generated, and (c) there are only

finitely many cuts, sparsified or not. In turn, (a) and (c) are obvious, and (b) follows from the operation of the algorithm: Whenever $\varphi(\widehat{z}) \geq \text{LPCC}_{\text{ub}}$, the new point cut or ray cut will cut off all binary vectors generated so far, including $\widehat{z}$; if $\varphi(\widehat{z}) < \text{LPCC}_{\text{ub}}$, then $\widehat{z}$ cannot be one of previously generated binary vectors because its φ-value is smaller than those of the other vectors. □

5.1. A numerical example. We use the following simple example to illustrate the algorithm:

$$
\begin{array}{ll}
\underset{(x,y)}{\text{minimize}} & x_1 + 2\,y_1 - y_3 \\
\text{subject to} & x_1 + x_2 \geq 5, \\
& x_1,\, x_2 \geq 0, \\
& 0 \leq y_1 \perp x_1 - y_3 + 1 \qquad \geq 0, \\
& 0 \leq y_2 \perp x_2 + y_1 + y_2 \qquad \geq 0, \\
& 0 \leq y_3 \perp x_1 + x_2 - y_2 + 2 \geq 0.
\end{array} \tag{5.1}
$$

Note that the LCP in the variable y is not derived from a convex quadratic program; in fact, the matrix

$$
M \equiv \begin{bmatrix} 0 & 0 & -1 \\ 1 & 1 & 0 \\ 0 & -1 & 0 \end{bmatrix}
$$

has all principal minors nonnegative, but the LCPs defined by this matrix may have zero or unbounded solutions.

Initialization: Set the upper bound as infinity: $\text{LPCC}_{\text{ub}} = \infty$. Set the working set $\mathcal{Z}_{\text{work}}$ and the waiting set $\mathcal{Z}_{\text{wait}}$ both equal to empty.

Iteration 1: Since $\widehat{\mathcal{Z}}_{\text{work}} = \{0,1\}^3$, we can pick an arbitrary binary vector z. We choose $z = (0,0,0)$ and solve the dual LP (2.9):

$$
\begin{array}{ll}
\underset{(\lambda,u^{\pm},v)}{\text{maximize}} & 5\,\lambda + u_1^+ + 2\,u_3^+ - u_1^- - 2\,u_3^- \\
\text{subject to} & \lambda - u_1^+ + u_1^- - u_3^+ + u_3^- \leq 1, \\
& \lambda - u_2^+ + u_2^- - u_3^+ + u_3^- \leq 0, \\
& -u_2^+ + u_2^- - v_1 \leq 2, \\
& -u_2^+ + u_2^- + u_3^+ - u_3^- - v_2 \leq 0, \\
& u_1^+ - u_1^- - v_3 \leq -1, \\
& v_1 + v_2 + v_3 \leq 0, \\
& (\,\lambda, u^{\pm}, v\,) \geq 0,
\end{array} \tag{5.2}
$$

which is unbounded, yielding an extreme ray with $u^+ = (0, 10/7, 10/7)$ and $v = (0,0,0)$ and a corresponding ray cut: $z_2 + z_3 \geq 1$. (Briefly, this cut is valid since $z_2 = z_3 = 0$ implies that both $x_2 + y_1 + y_2 = 0$ and $x_1 + x_2 - y_2 + 2 = 0$, which can't both hold for nonnegative x and y.) Add this cut to $\mathcal{Z}_{\text{work}}$, and initiate the sparsification procedure. This inequality $z_2 + z_3 \geq 1$ can be branched into $z_2 \geq 1$ or $z_3 \geq 1$. To test if $z_2 \geq 1$ is a valid cut, we form the following relaxed LP of (5.1) by

restricting $x_2 + y_1 + y_2 = 0$:

$$\begin{array}{lll} \underset{(x,y)}{\text{minimize}} & x_1 + 2\,y_1 - y_3 & \\ \text{subject to} & x_1 + x_2 & \geq 5, \\ & x_1 - y_3 + 1 & \geq 0, \\ & x_2 + y_1 + y_2 & = 0, \\ & x_1 + x_2 - y_2 + 2 & \geq 0, \\ & x,\, y \geq 0. & \end{array} \tag{5.3}$$

An optimal solution of the LP (5.3) is $(x_1, x_2, y_1, y_2, y_3) = (5, 0, 0, 0, 6)$ with the optimal objective value $\mathrm{LP}_{\mathrm{rlx}} = -1$. This is not a feasible solution of the LPCC (5.1) because the third complementarity is violated. The inequality $z_2 \geq 1$ is therefore placed in the waiting set $\mathcal{Z}_{\mathrm{wait}}$. We then use $(x_1, x_2) = (5, 0)$ to recover an LPCC feasible solution by solving the LCP in the variable y. This yields $y = (0, 0, 0)$, $w = (6, 0, 7)$, and hence a corresponding vector $z = (1, 0, 1)$. By using this z in (2.9), we get another dual problem:

$$\begin{array}{lr} \underset{(\lambda,u^{\pm},v)}{\text{maximize}} & 5\,\lambda + u_1^+ + 2\,u_3^+ - u_1^- - 2\,u_3^- \\ \text{subject to} & \lambda - u_1^+ + u_1^- - u_3^+ + u_3^- \leq 1, \\ & \lambda - u_2^+ + u_2^- - u_3^+ + u_3^- \leq 0, \\ & -u_2^+ + u_2^- - v_1 \leq 2, \\ & -u_2^+ + u_2^- + u_3^+ - u_3^- - v_2 \leq 0, \\ & u_1^+ - u_1^- - v_3 \leq -1, \\ & u_1^+ + v_2 + u_3^+ \leq 0, \\ & (\,\lambda, u^{\pm}, v\,) \geq 0, \end{array} \tag{5.4}$$

which has an optimal value 5 that is smaller than the current upper bound $\mathrm{LPCC}_{\mathrm{ub}}$. So we update the upper bound as $\mathrm{LPCC}_{\mathrm{ub}} = 5$. Note that this update occurs during the sparsification step. A corresponding optimal solution to (5.4) is $u^+ = (0, 1, 0)$ and $v = (0, 0, 1)$. Hence we can add the point cut $z_2 + (1 - z_3) \geq 1$ to $\mathcal{Z}_{\mathrm{work}}$.

When we next proceed to the other branch $z_3 \geq 1$, we have a relaxed LP:

$$\begin{array}{lll} \underset{(x,y)}{\text{minimize}} & x_1 + 2\,y_1 - y_3 & \\ \text{subject to} & x_1 + x_2 & \geq 5, \\ & x_1 - y_3 + 1 & \geq 0, \\ & x_2 + y_1 + y_2 & \geq 0, \\ & x_1 + x_2 - y_2 + 2 & = 0, \\ & x,\, y \geq 0. & \end{array} \tag{5.5}$$

Solving (5.5) gives an optimal value $\mathrm{LP}_{\mathrm{rlx}} = -1$, which is smaller than $\mathrm{LPCC}_{\mathrm{ub}}$, and a violated complementarity with $w_2 = 12$ and $y_2 = 7$. By adding $z_3 \geq 1$ to $\mathcal{Z}_{\mathrm{wait}}$, we apply the LPCC feasibility recovering procedure to $x = (0, 5)$ and get a new LPCC

feasible piece with $z = (1, 1, 1)$. By substituting z into (2.9), we get another LP:

$$
\begin{array}{lrcl}
\underset{(\lambda, u^{\pm}, v)}{\text{maximize}} & 5\,\lambda + u_1^+ + 2\,u_3^+ - u_1^- - 2\,u_3^- & & \\
\text{subject to} & \lambda - u_1^+ + u_1^- - u_3^+ + u_3^- & \le & 1, \\
& \lambda - u_2^+ + u_2^- - u_3^+ + u_3^- & \le & 0, \\
& -u_2^+ + u_2^- - v_1 & \le & 2, \\
& -u_2^+ + u_2^- + u_3^+ - u_3^- - v_2 & \le & 0, \\
& u_1^+ - u_1^- - v_3 & \le & -1, \\
& u_1^+ + u_2^+ + u_3^+ & \le & 0, \\
& (\lambda, u^{\pm}, v) & \ge & 0,
\end{array}
\tag{5.6}
$$

which has an optimal objective value 0. So a better upper bound is found; thus $\mathrm{LPCC}_{\mathrm{ub}} = 0$. A point cut $1 - z_3 \ge 1$ is derived from an optimal solution of (5.6). This cut obviously implies the previous cut $z_2 + (1 - z_3) \ge 1$. In order to reduce the work load of the IP solver, we can delete $z_2 + (1 - z_3) \ge 1$ from $\mathcal{Z}_{\mathrm{work}}$ and add in $1 - z_3 \ge 1$ instead. So far, we have the updated upper bound $\mathrm{LPCC}_{\mathrm{ub}} = 0$ and the working set $\mathcal{Z}_{\mathrm{work}}$ defined by the two inequalities:

$$
z_2 + z_3 \ge 1 \quad \text{and} \quad 1 - z_3 \ge 1. \tag{5.7}
$$

This completes iteration 1. During this iteration, we have solved 5 LPs, the $\mathrm{LPCC}_{\mathrm{ub}}$ has improved twice, and we have obtained 2 valid cuts.

Iteration 2: Solving a satisfiability IP yields a $z = (0, 1, 0) \in \widehat{\mathcal{Z}}_{\mathrm{work}}$. Indeed, any element in $\widehat{\mathcal{Z}}_{\mathrm{work}}$, which is defined by the two inequalities in (5.7), must have $z_2 = 1$ and $z_3 = 0$; thus it remains to determine z_1. As it turns out, z_1 is irrelevant. To see this, we substitute $z = (0, 1, 0)$ into (2.9), obtaining

$$
\begin{array}{lrcl}
\underset{(\lambda, u^{\pm}, v)}{\text{maximize}} & 5\,\lambda + u_1^+ + 2\,u_3^+ - u_1^- - 2\,u_3^- & & \\
\text{subject to} & \lambda - u_1^+ + u_1^- - u_3^+ + u_3^- & \le & 1, \\
& \lambda - u_2^+ + u_2^- - u_3^+ + u_3^- & \le & 0, \\
& -u_2^+ + u_2^- - v_1 & \le & 2, \\
& -u_2^+ + u_2^- + u_3^+ - u_3^- - v_2 & \le & 0, \\
& u_1^+ - u_1^- - v_3 & \le & -1, \\
& u_2^+ + v_1 + v_3 & \le & 0, \\
& (\lambda, u^{\pm}, v) & \ge & 0.
\end{array}
\tag{5.8}
$$

The LP (5.8) is unbounded and has an extreme ray where $u^+ = (0, 0, 10/7)$ and $v = (0, 10/7, 0)$. So we can add a valid ray cut $(1 - z_2) + z_3 \ge 1$ to $\mathcal{Z}_{\mathrm{work}}$.

Termination: The updated working set $\mathcal{Z}_{\mathrm{work}}$ consists of 3 inequalities:

$$
\left\{ \begin{array}{r} z_2 + z_3 \ge 1 \\ 1 - z_3 \ge 1 \\ (1 - z_2) + z_3 \ge 1 \end{array} \right\},
$$

which can be seen to be inconsistent. Hence we get a certificate of termination. Since there is one point cut in $\mathcal{Z}_{\rm work}$, the LPCC (5.1) has an optimal objective value 0, which happens on the piece $z = (1, 1, 1)$. (This termination can be expected from the fact that $z_2 = 1$ and $z_3 = 0$ for elements in the set $\widehat{\mathcal{Z}}_{\rm work}$ prior to the last ray cut; these values of z imply that $y_2 = w_3 = 0$, which is not consistent with the nonnegativity of x. This inconsistency is detected by the algorithm through the generation of a ray cut that leaves $\widehat{\mathcal{Z}}_{\rm work}$ empty.)

6. Computational results. To test the effectiveness of the algorithm, we have implemented and compared it with benchmark algorithms from NEOS, which for the purpose here were chosen to be the FILTER solver and the KNITRO solver. As expected, these two solvers consistently produce high-quality LPCC feasible solutions. For the test problems we used, both often found solutions that turned out to be globally optimal, as was proved by our algorithm. (The details can be seen in Tables 6.1, 6.2, and 6.3.) We coded our algorithm in MATLAB and used CPLEX 9.1 to solve the LPs and the satisfiability IPs. The experiments were run on a DELL desktop computer with 1.40 GHz Pentium 4 processor and 1.00 GB of RAM.

Our goal in this computational study is threefold: (A) to test the practical ability of the algorithm to provide a certificate of global optimality for LPCCs with finite optimal solutions; (B) to determine the quality of the solutions obtained by using the simple-cut preprocessor; and (C) to demonstrate that the algorithm is capable of detecting infeasibility and unboundedness for LPCCs of these kinds. All problems are randomly generated. One at a time, a total of $\lfloor m/3 \rfloor$ simple cuts are generated in the preprocessing step for each problem. To test (A) and (B), the problems are generated to have optimal solutions; for (C), the problems are generated to be either infeasible or have unbounded objective values. The algorithm does not make use of such information in any way; instead, it is up to the algorithm to verify the prescribed

TABLE 6.1

Special LPCCs with $B = 0$, $A \in \Re^{90\times 100}$, and 100 complementarities. Remark: The first column "#" is the problem counter; the second column "$\mathrm{LPCC}_{\rm lb}$" contains the objective values of LP relaxations before and after the preprocessing. The column "$\mathrm{LPCC}_{\rm ub}$" reports the objective values of the LPCC feasible solutions. The right subcolumn contains the verifiably optimal $\mathrm{LPCC}_{\rm min}$. The left subcolumn contains the values obtained after preprocessing with the LPCC feasibility recovery procedure. The objective values obtained from FILTER *and* KNITRO *are reported in the next columns. (Note that these values are very comparable and practically optimal in all problems except* 1 *and* 7 *for both and* 8 *for* KNITRO.*) The total number of LPs solved (excluding the $\lfloor m/3 \rfloor$ relaxed LPs in the preprocessing step) and the number of IPs solved in the run are reported in the last two columns. At the suggestion of a referee, we also reported the number of "major iterations" in the two* NEOS *solvers; these are placed as subscripts in the objective values of the respective solvers. It should be noted that such iterations refer to different procedures in the two solvers.*

	$\mathrm{LPCC}_{\rm lb}$		$\mathrm{LPCC}_{\rm ub}$					
#	rxLPCC	Preprocess	Preprocess	$\mathrm{LPCC}_{\rm min}$	FILTER	KNITRO	LP	IP
1	1094.6041	1106.8297	1146.7550	1127.4885	1140.5614_{5}	1141.6696_{52}	396	18
2	1172.1830	1176.6893	1185.0192	1182.2146	1182.2145_{5}	1182.2147_{47}	57	1
3	820.2584	823.8912	823.9099	823.9055	823.9055_{10}	823.9058_{55}	14	1
4	796.9560	813.9752	840.4828	833.9718	833.9717_{6}	833.9718_{41}	611	22
5	841.1786	849.0122	850.2416	849.8451	849.8451_{5}	849.8452_{44}	66	1
6	924.7529	926.1028	926.5924	926.5000	926.5000_{5}	926.5000_{56}	21	1
7	1536.1748	1541.4464	1543.8863	1541.9443	1543.1950_{6}	1543.1951_{55}	35	1
8	1076.8760	1090.2155	1109.1441	1106.3617	1106.3616_{5}	1113.8938_{70}	363	25
9	1232.7912	1239.7156	1239.8283	1239.8283	1239.8284_{7}	1239.8285_{62}	10	1
10	1217.1191	1229.2734	1250.6693	1249.9884	1249.9884_{8}	1249.9886_{67}	832	46

TABLE 6.2

Special LPCCs with $B = 0$, $A \in \Re^{200\times 300}$, and 300 complementarities. Remark: The explanation of this table is the same as Table 6.1. Note that in problem 7, the solution obtained after the preprocessing step is immediately verified to be globally optimal. For these runs, the KNITRO *solutions are practically optimal in all cases, but the* FILTER *solution in problem 2 is noticeably suboptimal.*

	LPCC_{lb}		LPCC_{ub}					
#	rxLPCC	Preprocess	Preprocess	LPCC_{min}	FILTER	KNITRO	LP	IP
1	2469.4400	2474.3166	2479.1835	2478.2254	2478.2256_{19}	2478.2264_{66}	125	1
2	3213.7176	3229.1930	3299.1115	3270.1842	3280.1865_{8}	3270.1844_{72}	4071	62
3	3639.4490	3651.5714	3671.6385	3660.5407	3660.5412_{42}	3660.5412_{79}	350	2
4	3127.3708	3140.3119	3265.7213	3176.4109	3176.4108_{11}	3176.4115_{69}	1249	15
5	2958.9147	2959.9381	2959.9498	2959.9498	2959.9495_{6}	2959.9529_{66}	5	1
6	2630.3282	2645.6771	2703.0018	2672.5706	2684.5288_{30}	2672.5710_{60}	4511	70
7	2616.9852	2617.2640	2617.2640	2617.2640	2617.2638_{14}	2617.2673_{65}	0	0
8	2766.9544	2770.1510	2771.3315	2771.2374	2771.2372_{27}	2771.2379_{70}	26	1
9	2842.4480	2846.7174	2847.9806	2847.6923	2847.6926_{7}	2847.6929_{48}	319	2
10	3207.6861	3220.3810	3235.4082	3230.9893	3230.9896_{6}	3230.9897_{66}	1569	16

TABLE 6.3

General LPCCs with $B \neq 0$, $A \in \Re^{55\times 50}$, and 50 complementarities. Remark: For these runs, there are more instances where the two NEOS *solutions are noticeably suboptimal.*

	LPCC_{lb}		LPCC_{ub}					
#	rxLPCC	Preprocess	Preprocess	LPCC_{min}	FILTER	KNITRO	LP	IP
1	28.7739	29.0318	29.0502	29.0501	29.0501_{8}	30.0155_{32}	21	2
2	36.1885	36.8258	39.1063	37.5509	37.5509_{7}	37.5510_{25}	229	9
3	33.8630	34.4988	39.1285	37.0022	38.3216_{7}	38.7521_{28}	4842	696
4	33.7618	34.1479	34.3034	34.2228	34.6057_{5}	34.2398_{54}	102	7
5	21.4187	21.9246	22.9642	22.2835	22.2945_{5}	22.2837_{35}	209	24
6	29.8919	29.9681	30.1085	30.0829	30.0829_{6}	30.0830_{26}	108	13
7	37.6712	37.9972	38.0405	38.0405	38.0419_{6}	38.0419_{29}	92	7
8	20.8210	21.4586	27.9618	22.3969	22.7453_{7}	22.4164_{37}	187	21
9	39.0227	39.4792	40.7839	40.3380	44.7872_{8}	44.3173_{26}	321	14
10	40.0135	40.7994	41.6865	41.3957	41.5810_{5}	41.5810_{37}	190	19

problem status. In all of the experiments, optimality of the LPCC is declared if the difference between the lower and upper bounds is less than or equal to 1e-6; this tolerance is also employed to determine the LPCC feasibility of the relaxed LP solutions. The parameter δ for the sparsification step is selected to be 0.2.

All problems have the nonnegativity constraint $x \geq 0$. The computational results for the problems with finite optima are reported in Figures 6.1, 6.2, and 6.3 and Tables 6.1, 6.2, and 6.3. Each figure contains one set of ten randomly generated problems with the same characteristics. Figures 6.1, 6.2, and 6.3 correspond to problems with $[n, m, k] = [100, 100, 90]$, $[300, 300, 200]$, and $[50, 50, 55]$, respectively. These sizes and the data density are dictated by the limitations of MATLAB that is the environment where our experiments were performed. All data are randomly generated with uniform distributions. The objectives vectors c and d are generated from the intervals [0 1] and [1 3], respectively. For Figures 6.1 and 6.2, the matrix $B = 0$, and the matrix M is generated with up to 2000 nonzero entries and of the form:

$$M \equiv \begin{bmatrix} D_1 & E^T \\ -E & D_2 \end{bmatrix}, \tag{6.1}$$

where D_1 and D_2 are positive diagonal matrices of random order and with elements

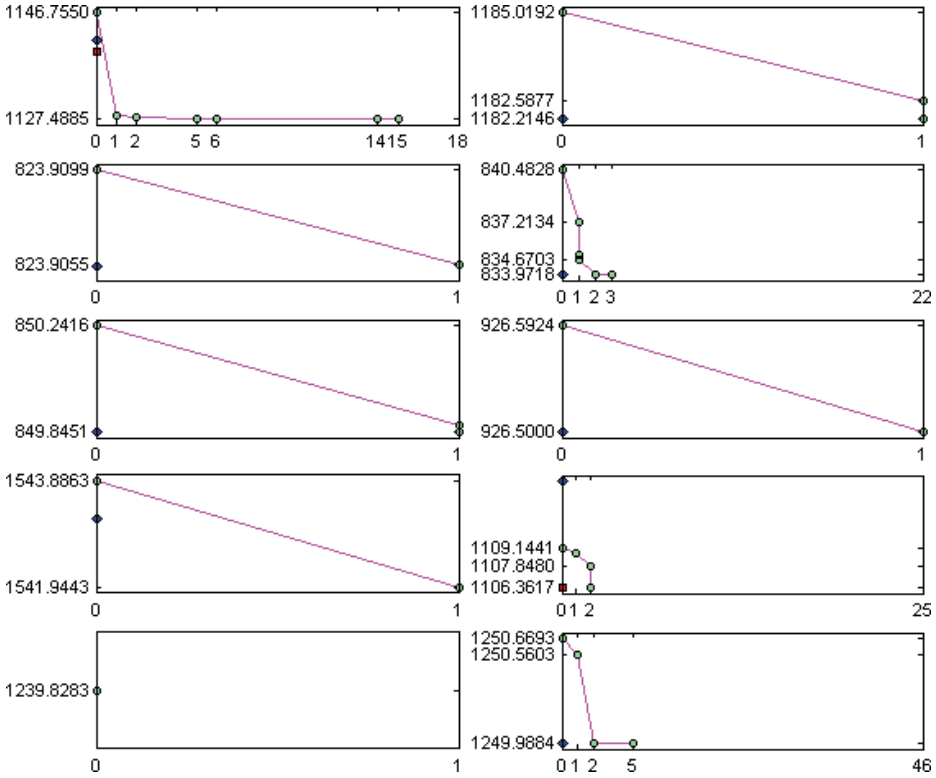


FIG. 6.1. *Special LPCCs with* $B = 0$, $A \in \Re^{90\times 100}$, *and* 100 *complementarities. Remark: Each circle signifies that a better feasible LPCC solution is found. The circle's horizontal coordinate indicates the iteration where* $\mathrm{LPCC}_{\mathrm{ub}}$ *is updated; its vertical coordinate gives the value of updated* $\mathrm{LPCC}_{\mathrm{ub}}$ *(we omitted some values if they are not significantly improved). Note that it is possible for* $\mathrm{LPCC}_{\mathrm{ub}}$ *to improve within one iteration by the sparsification step; see the example in subsection* 5.1 *and also the top run in the right column. In the fifth run in the left column, both of the* FILTER *and* KNITRO *results coincide with* $\mathrm{LPCC}_{\mathrm{min}}$*, which is obtained after preprocessing and verified to be optimal after one iteration.*

chosen from [0 2] and E is arbitrary with elements in $[-11]$. The vector q is randomly generated with elements in the interval $[-20 - 10]$. Note that M is positive definite, albeit not symmetric. This property of M and the choice of $B = 0$ ensure LPCC feasibility and thus optimality (because c and d are nonnegative and the variables are nonnegative). For Figure 6.3, $B \neq 0$, and the matrix M has no special structure but has only 10% density. The rest of the data A, f, q, and N are generated to ensure LPCC feasibility and thus optimality. Details of the data generation and the resulting data can be found in [50].

Figures 6.1, 6.2, and 6.3 detail the progress of the runs, showing in particular how $\mathrm{LPCC}_{\mathrm{ub}}$ decreases with the number of iterations. The vertical axis refers to the LPCC objective values, and the horizontal axis labels the number of iterations as defined in the opening paragraph of section 5. The top value on the vertical axis is the LPCC objective value obtained at termination of the preprocessor with the LPCC feasibility recovery step. The bottom value is verifiably $\mathrm{LPCC}_{\mathrm{min}}$. The vertical axis is scaled differently in each run with respect to the difference between the top and the bottom values. As comparison, the objective values obtained from FILTER (marked by the red square) and KNITRO (marked by the blue diamond) are also shown on the vertical axis;

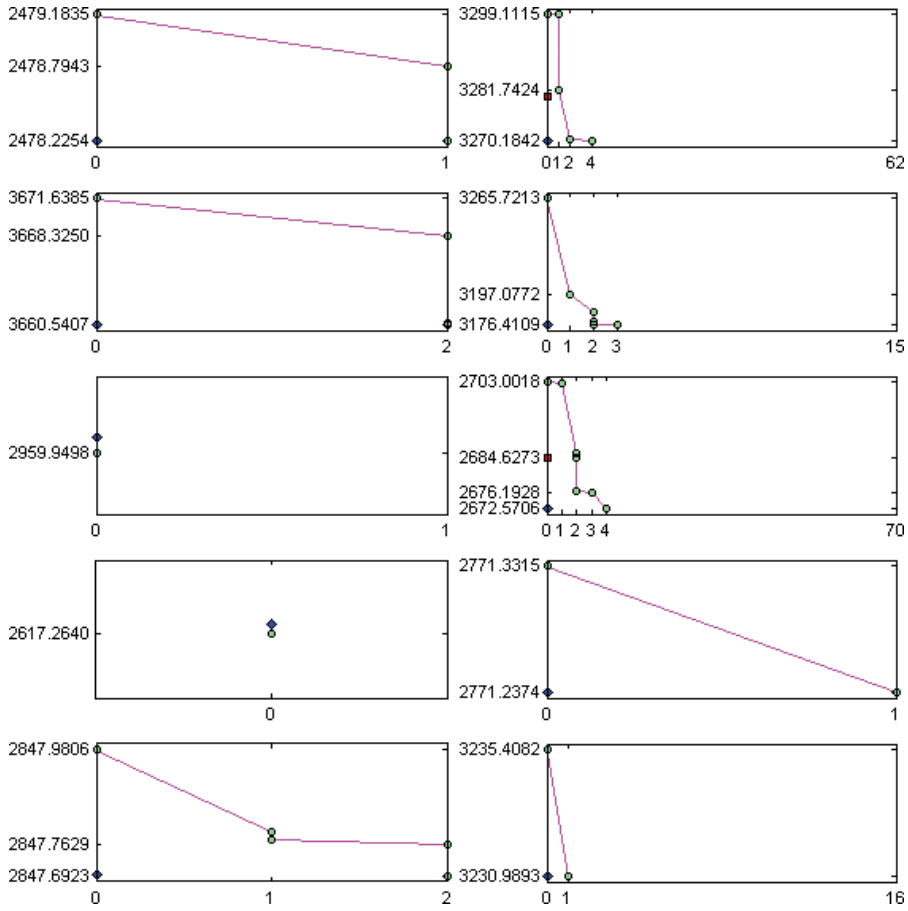


FIG. 6.2. *Special LPCCs with* $B = 0$, $A \in \Re^{200\times 300}$, *and* 300 *complementarities. Remark: The explanation for the figure is similar to that of Figure* 6.1. *Note that in the third and fourth runs in the left column* $\mathrm{LPCC}_{\mathrm{ub}}$ *is obtained right after preprocessing. In the third run, the solution's global optimality is verified after* 1 *iteration; while in the fourth run, the solution is immediately verified to be globally optimal (the difference between the upper and lower bounds of the LPCC is within* 1e-6*).*

if the difference between the FILTER and KNITRO values in a run is within 1e-3, we mark only the KNITRO result (the exact values from these two solvers can be found in Tables 6.1, 6.2, and 6.3). The upper limit of the horizontal axis indicates the number of IPs needed to be solved in each run. Note that in some runs a globally optimal solution might have been obtained in an earlier iteration without certification, and the algorithm needs more subsequent iterations to verify its global optimality. For example, in the fourth run of the right-hand column in Figure 6.1, a globally optimal solution is first obtained at iteration 2, but the certificate is established only after 23 more iterations. Other details about the figures are summarized in the remarks below the figures.

Corresponding to the problems in Figures 6.1, 6.2, and 6.3, respectively, Tables 6.1, 6.2, and 6.3 report more details about the runs, which are indexed by counting first rowwise and then columnwise in the figures (for example, the fourth run in Table 6.1 is the second row on the right column in Figure 6.1). In addition to the objective values obtained in our algorithm and from the NEOS solvers, these tables also

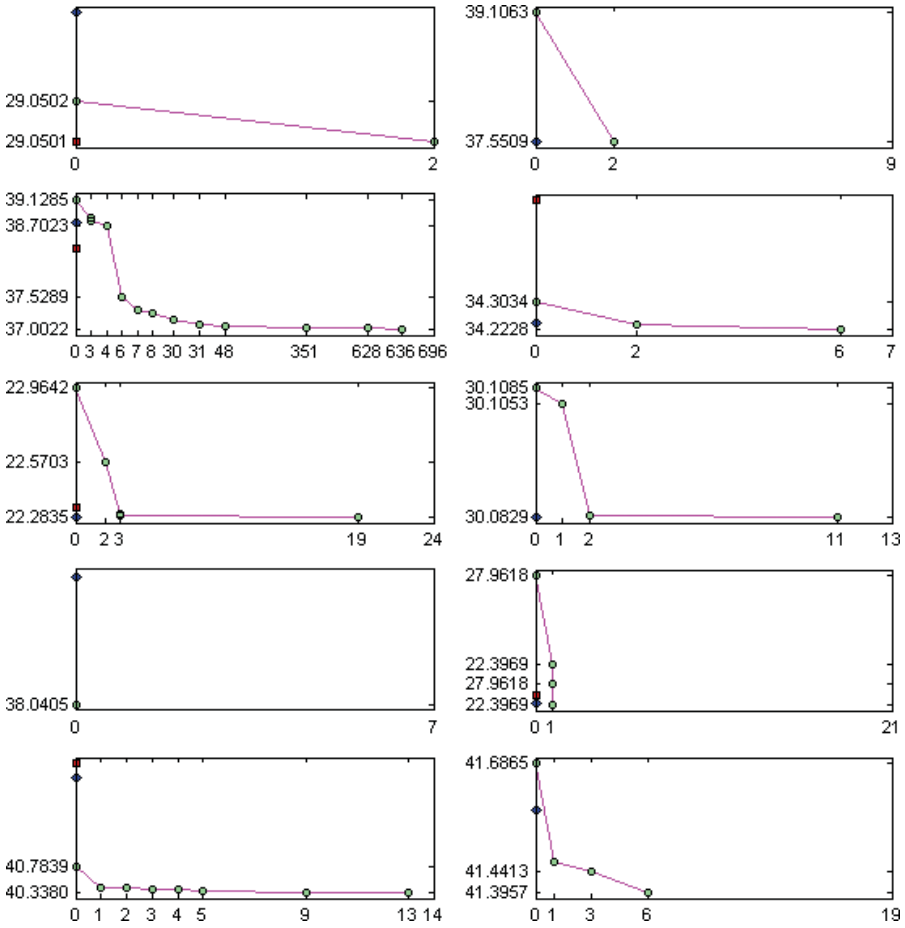


FIG. 6.3. *General LPCCs with* $B \neq 0$, $A \in \Re^{55\times 50}$, *and* 50 *complementarities.*

report the numbers of IPs and LPs (excluding the $\lfloor m/3 \rfloor$ relaxed LPs solved in the preprocessor), solved in the solution process. These numbers, which are independent of the computational platform and machine, provide a good indicator of the efforts required by the algorithm in processing the LPCCs. We did not report computational times for two reasons: (i) The MATLAB results are computer-dependent, and the runs involve interfaces between MATLAB and CPLEX, and (ii) our runs are experimental, and our coding is at an amateur level.

The computational results for the infeasible and unbounded LPCCs are reported in Table 6.4, which contains 3 subtables (a), (b), and (c). The first two subtables (a) and (b) pertain to feasible but unbounded LPCCs. For the unbounded problems, we set $B = 0$, q is arbitrary, and we generate A with a nonnegative column, M given by (6.1), and f such that $\{x \geq 0 : Ax \geq f\}$ is feasible. Problems in (a) and (b) have the same parameters except for the objective vectors c and d and matrix A. For the problems in (a), we simply maximize one single x-variable whose A column is nonnegative. For the problems in (b), the objective vectors c and d are both negative, and the matrix A is the same as it is in group (a) except that a small number 0.005 is added to its nonnegative column (see the discussion in the first conclusion below for why this is done). The third subtable (c) pertains to a class of infeasible LPCCs

TABLE 6.4

Infeasible and unbounded LPCCs with 50 complementarities. # iters = number of returns to Step 1 = number of IPs solved; # cuts = number of satisfiability constraints in $\mathcal{Z}_{work}$ at termination; # LPs = number of LPs solved, excluding the $\lfloor m/3 \rfloor$ relaxed LPs in the preprocessing step.

Prob	# iters	# cuts	# LPs	# iters	# cuts	# LPs	# iters	# cuts	# LPs
1	50	47	195	4	3	9	14	14	28
2	6	4	14	5	4	12	2	2	4
3	1081	828	2604	2	1	5	38	38	76
4	166	144	424	42	39	120	7	7	14
5	436	305	991	735	621	1860	47	49	100
6	18	17	54	498	379	1125	48	48	96
7	3	4	11	352	127	663	20	20	40
8	426	356	1191	489	373	1158	13	13	26
9	9	9	26	5	4	12	50	50	100
10	4	3	11	9	7	22	6	6	12
	(a)			(b)			(c)		

TABLE 6.5

General LPCCs with $B \neq 0$, $A \in \Re^{25\times 25}$, and 25 complementarities. Column A = number of IPs or LPs solved in the run without sparsification; column B = number of IPs or LPs solved in the run with sparsification. Remark: In column A, the number of IPs solved in the run is equal to the number of solved LPs. Except for problems 3, 9, and 10, the B approach (with the sparsification step implemented) is doing much better than the A approach. Especially for problems 1 and 5–8, the numbers of IPs and LPs are dramatically reduced. For the remaining problems, the computational effort with sparsification is at least comparable to, if not better than, the approach without sparsification. When the number of complementarities in the LPCCs grows, we expect more computational savings with the sparsification step implemented.

	A		B	
Prob	# LPs	# IPs	# LPs	# IPs
1	122	122	61	18
2	17	17	33	11
3	7	7	51	9
4	16	16	41	12
5	280	280	70	21
6	598	598	85	23
7	195	195	90	26
8	65	65	43	10
9	9	9	41	9
10	8	8	33	8

generated as follows: q, N, and M are all positive so that the only solution to the LCP $0 \leq y \perp q + Nx + My \geq 0$ for $x \geq 0$ is $y = 0$; $Ax + By \geq f$ is feasible for some $(x, y) \geq 0$, with $y \neq 0$, but $Ax \geq f$ has no solution in $x \geq 0$.

To illustrate the effectiveness of the sparsification step, we generated some LPCCs with $n = m = k = 25$ and the same characteristics as the problems in Figure 6.3. Table 6.5 reports the numbers of LPs and IPs that are needed to be solved in both runs with or without this step.

The main conclusions from the experiments are summarized below.

• The algorithm successfully terminates with the correct status of all of the LPCCs reported. In fact, we have tested many more problems than those reported and obtained similar success. There are, nevertheless, a few instances where the LPCCs are apparently unbounded but the algorithm fails to terminate after 6000 iterations without the definitive conclusion, even though the LPCC objective is noticeably tending to $-\infty$. We cannot explain these exceptional cases which we suspect are due to round-off errors in the computations. This suspicion led us to add the small 0.005

adjustment in the unbounded set of runs reported above; with this small adjustment, the algorithm successfully terminated with the desired certificate of unboundedness.

• For the special LPCCs with $B = 0$, the results from the 2 NEOS algorithms FILTER and KNITRO are proved to be suboptimal in 2 out of the 20 runs (the first and fourth runs on the left column in Figure 6.1). In the other 18 runs, our algorithm is able to obtain an optimal solution with little computational effort (within 5 iterations) but requires significant additional computations to produce the desired certificate of global optimality. For the general LPCCs with $B \neq 0$, the objective values obtained from FILTER and KNITRO are suboptimal in 6 out of 10 runs. In the other 4 runs, only 5 iterations are needed to derive either a globally optimal solution or an LPCC feasible solution whose objective value is within 3% of the optimal value. These results confirm that the verification of global optimality is generally much more demanding than the computation of the solution without proof of optimality.

• Except for one problem (problem 7 in Table 6.3), the solutions obtained by the simple-cut preprocessor for all LPCCs with finite optima are within 5% of the globally optimal solutions. In fact, some of the solutions obtained from the preprocessing are immediately verified to be optimal. This suggests that very high-quality LPCC feasible solutions can be produced efficiently by solving a reasonable number of LPs.

• The sparsification procedure is quite effective; so is the LPCC feasibility recovery step. Indeed without the latter, there is a significant percentage of problems where the algorithm fails to make progress after 3000 iterations. With this step installed, all problems are resolved satisfactorily.

• While the numbers of IPs solved are quite reasonable in most cases, there are several runs where the numbers of relaxed LPs solved are unusually large, especially when the problem size increases. This suggests that stronger cuts are needed for both general LPCCs and for specialized problems arising from large-scale applications. The implementation of a dedicated solver for satisfiability problems, such as those described in [7, 31], could considerably improve the overall solution times of the LPCC algorithm. These refinements of the algorithm are presently being investigated.

Concluding remarks. In this paper, we have presented a parameter-free IP-based algorithm for the global resolution of an LPCC and reported computational results with the application of the algorithm for solving a set of randomly generated LPCCs of moderate sizes. Continued research on refining the algorithm and applying it to realistic classes of LPCCs, such as the bilevel machine-learning problems described in [6, 32, 33] and other applied problems, is currently underway.

Acknowledgments. The authors are grateful to 2 referees for their constructive comments that have significantly improved the presentation of the paper. In particular, one of them alerted the authors to [9], which discusses Benders techniques for conditional linear constraints without requiring the big-M.

REFERENCES

[1] M. ANITESCU, *On using the elastic mode in nonlinear programming approaches to mathematical programs with complementarity constraints*, SIAM J. Optim., 15 (2005), pp. 1203–1236.

[2] M. ANITESCU, *Global convergence of an elastic mode approach for a class of mathematical programs with complementarity constraints*, SIAM J. Optim., 16 (2005), pp. 120–145.

[3] M. ANITESCU, P. TSENG, AND S.J. WRIGHT, *Elastic-mode algorithms for mathematical programs with equilibrium constraints: Global convergence and stationarity properties*, Math. Program., 110 (2007), pp. 337–371.

[4] C. AUDET, P. HANSEN, B. JAMMARD, AND G. SAVARD, *Links between linear bilevel and mixed* 0–1 *programming problems*, J. Optim. Theory Appl., 93 (1997), pp. 273–300.

[5] C. Audet, G. Savard, and W. Zghal, *New branch-and-cut algorithm for bilevel linear programming*, J. Optim. Theory Appl., 134 (2007), pp. 353–370.
[6] K. Bennett, X. Ji, J. Hu, G. Kunapuli, and J.S. Pang, *Model selection via bilevel programming*, in Proceedings of the International Joint Conference on Neural Networks (IJCNN'06) Vancouver, B.C., Canada, IEEE Press, Piscataway, NJ, 2006.
[7] B. Borchers and J. Furman, *A two-phase exact algorithm for MAX-SAT and weighted MAX-SAT Problems*, J. Comb. Optim., 2 (1998), pp. 299–306.
[8] L. Chen and D. Goldfarb, *An Active-Set Method for Mathematical Programs with Linear Complementarity Constraints*, SIAM J. Optim., submitted.
[9] G. Codato and M. Fischetti, *Combinatorial Benders' cuts for mixed-integer linear programming*, Oper. Res., 54 (2006), pp. 758–766.
[10] S. Dempe, *Foundations of Bilevel Programming*, Kluwer Academic, Dordrecht, 2002.
[11] S. Dempe, *Annotated bibliography on bilevel programming and mathematical programs with equilibrium constraints*, Optimization, 52 (2003), pp. 33–359.
[12] S. Dempe, V.V. Kalashnikov, and N. Kalashnykova, *Optimality conditions for bilevel programming problems*, in Optimization with Multivalued Mappings: Theory, Applications and Algorithms, S. Dempe and N. Kalashnykova, eds., Springer-Verlag, New York, 2006, pp. 11–36.
[13] M.L. Flegel and C. Kanzow, *On the Guignard constraint qualification for mathematical programs with equilibrium constraints*, Optimization, 54 (2005), pp. 517–534.
[14] R. Fletcher and S. Leyffer, *Solving mathematical program with complementarity constraints as nonlinear programs*, Optim. Methods Softw., 19 (2004), pp. 15–40.
[15] R. Fletcher, S. Leyffer, D. Ralph, and S. Scholtes, *Local convergence of SQP methods for mathematical programs with equilibrium constraints*, SIAM J. Optim., 17 (2006), pp. 259–286.
[16] C.A. Floudas and V. Visweswaran, *A global optimization algorithm (GOP) for certain classes of nonconvex NLPs*-I. *Theory*, Comput. Chem. Engrg., 14 (1990), pp. 1397–1417.
[17] C.A. Floudas and V. Visweswaran, *Primal-relaxed dual global optimization approach*, J. Optim. Theory Appl., 78 (1993), pp. 187–225.
[18] M. Fukushima and J.S. Pang, *Some feasibility issues in mathematical programs with equilibrium constraints*, SIAM J. Optim., 8 (1998), pp. 673–681.
[19] M. Fukushima and J.S. Pang, *Convergence of a smoothing continuation method for mathematical programs with complementarity constraints*, in Ill-Posed Variational Problems and Regularization Techniques, Lecture Notes in Econom. and Math. Systems 477, M. Thera and R. Tichatschke, eds., Springer-Verlag, Berlin/Heidelberg, 1999, pp. 99–110.
[20] M. Fukushima and P. Tseng, *An implementable active-set algorithm for computing a B-stationary point of a mathematical program with linear complementarity constraints*, SIAM J. Optim., 12 (2002), pp. 724–739 [with erratum].
[21] P. Hansen, B. Jaumard, and G. Savard, *New branch-and-bound rules for linear bilevel programming*, SIAM J. Sci. Comput., 13 (1992), pp. 1194–1217.
[22] J.N. Hooker, *Logic-Based Methods for Optimization*, Wiley, New York, 2000.
[23] J.N. Hooker, *Integrated Methods for Optimization*, Springer-Verlag, New York, 2006.
[24] J.N. Hooker and G. Ottosson, *Logic-based Benders decomposition*, Math. Program., 96 (2003), pp. 33–60.
[25] X.M. Hu and D. Ralph, *Convergence of a penalty method for mathematical programming with complementarity constraints*, J. Optim. Theory Appl., 123 (2004), pp. 365–390.
[26] T. Ibaraki, *Complementary programming*, Oper. Res., 19 (1971), pp. 1523–1529.
[27] T. Ibaraki, *The use of cuts in complementary programming*, Oper. Res., 21 (1973), pp. 353–359.
[28] R.G. Jeroslow, *Cutting-planes for complementarity constraints*, SIAM J. Control Optim., 16 (1978), pp. 56–62.
[29] H. Jiang and D. Ralph, *Smooth SQP methods for mathematical programs with nonlinear complementarity constraints*, SIAM J. Optim., 10 (2000), pp. 779–808.
[30] H. Jiang and D. Ralph, *Extension of quasi-Newton methods to mathematical programs with complementarity constraints*, Comput. Optim. Appl., 25 (2003), pp. 123–150.
[31] S. Joy, J.E. Mitchell, and B. Borchers, *A branch-and-cut algorithm for MAX-SAT and weighted MAX-SAT*, Optim. Methods Softw., to appear.
[32] G. Kunapuli, K. Bennett, J. Hu, and J.S. Pang, *Classification model selection via bilevel programming*, Optim. Methods Softw., to appear.
[33] G. Kunapuli, K. Bennett, J. Hu, and J.S. Pang, *Bilevel Model Selection for Support Vector Machines*, manuscript, Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY, 2007.

[34] L. Lasdon, *Optimization Theory of Large Systems*, Dover, New York, 2002.
[35] S. Leyffer, *Complementarity constraints as nonlinear equations: Theory and numerical experience*, in Optimization and Multivalued Mappings, S. Dempe and V. Kalashnikov, eds., Springer-Verlag, New York, 2006, pp. 169–208.
[36] S. Leyffer, G. Lopéz-Calva, and J. Nocedal, *Interior methods for mathematical programs with complementarity constraints*, SIAM J. Optim., 17 (2006), pp. 52–77.
[37] Z.Q. Luo, J.S. Pang, and D. Ralph, *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press, Cambridge, 1996.
[38] J.V. Outrata, *Optimality conditions for a class of mathematical programs with equilibrium constraints*, Math. Oper. Res., 24 (1999), pp. 627–644.
[39] J. Outrata, M. Koçvara, and J. Zowe, *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints*, Kluwer Academic, Dordrecht, 1998.
[40] J.S. Pang and M. Fukushima, *Complementarity constraint qualifications and simplified B-stationarity conditions for mathematical programs with quilibrium constraints*, Comput. Optim. Appl., 13 (1999), pp. 111–136.
[41] J.S. Pang and S. Leyffer, *On the global minimization of the value-at-risk*, Optim. Methods Softw., 19 (2004), pp. 611–631.
[42] H. Scheel and S. Scholtes, *Mathematical programs with complementarity constraints: Stationarity, optimality and sensitivity*, Math. Oper. Res., 25 (2000), pp. 1–22.
[43] S. Scholtes, *Convergence properties of a regularization scheme for mathematical programs with complementarity constraints*, SIAM J. Optim., 11 (2001), pp. 918–936.
[44] M. Tawarmalani and N.V. Sahinidis, *Convexification and Global Optimization in Continuous and Mixed Integer Nonlinear Programming*, Kluwer Academic, Dordrecht, 2002.
[45] V. Visweswaran, C.A. Floudas, M.G. Ierapetritou, and E.N. Pistikopoulos, *A decomposition based global optimization approach for solving bilevel linear and nonlinear quadratic programs*, in State of the Art in Global Optimization: Computational Methods and Applications, Nonconvex Optim. Appl. 7, C.A. Floudas and P.M. Pardalos, eds., Kluwer Academic, Dordrecht, 1996, pp. 139–162.
[46] J.J. Ye, *Optimality conditions for optimization problems with complementarity constraints*, SIAM J. Optim., 9 (1999), pp. 374–387.
[47] J.J. Ye, *Constraint qualifications and necessary optimality conditions for optimization problems with variational inequality constraints*, SIAM J. Optim., 10 (2000), pp. 943–962.
[48] J.J. Ye, *Necessary and sufficient optimality conditions for mathematical programs with equilibrium constraints*, J. Math. Anal. Appl., 30 (2005), pp. 350–369.
[49] http://www-neos.mcs.anl.gov/neos/solvers/index.html.
[50] http://www.rpi.edu/~mitchj/generators/lpcc/.

SIAM J. OPTIM.
Vol. 19, No. 1, pp. 472–501

EMBEDDED IN THE SHADOW OF THE SEPARATOR*

FRANK GÖRING†, CHRISTOPH HELMBERG†, AND MARKUS WAPPLER†

Abstract. Eigenvectors to the second smallest eigenvalue of the Laplace matrix of a graph, also known as Fiedler vectors, are the basic ingredient in spectral graph partitioning heuristics. Maximizing this second smallest eigenvalue over all nonnegative edge weightings with bounded total weight yields the *absolute algebraic connectivity* introduced by Fiedler, who proved tight connections of this value to the connectivity of the graph. Our objective is to gain a better understanding of the connections between separators and the eigenspace of this eigenvalue by studying the dual semidefinite optimization problem to the absolute algebraic connectivity. By exploiting optimality conditions we show that this problem is equivalent to finding an embedding of the n nodes of the graph in n-space so that their barycenter is the origin, the distance between adjacent nodes is bounded by one, and the nodes are spread as much as possible (the sum of the squared norms is maximized). For connected graphs we prove that, for any separator in the graph, at least one of the two separated node sets is embedded in the shadow (with the origin being the light source) of the convex hull of the separator. Furthermore, we show that there always exists an optimal embedding whose dimension is bounded by the tree width of the graph plus one.

Key words. spectral graph theory, semidefinite programming, eigenvalue optimization, embedding, graph partitioning, tree width

AMS subject classifications. 05C50, 90C22, 90C35, 05C10, 05C78

DOI. 10.1137/050639430

1. Introduction.

Let $G := (N, E)$ be an undirected graph with node set $N := \{1, \dots, n\}$ and edge set $E \subseteq \{\{i, j\} : i, j \in N, i \neq j\}$. Edge $\{i, j\}$ will be abbreviated by ij if there is no danger of confusion. The adjacency matrix $A \in \mathbb{R}^{n\times n}$ of the graph is defined as the (symmetric) matrix having $a_{ij} = 1$ if $ij \in E$ and 0 otherwise. The *Laplace matrix* or *Laplacian* of the graph is the matrix $L := \operatorname{diag}(Ae) - A$, where e denotes the vector of all ones of appropriate dimension and $\operatorname{diag}(v)$ denotes the diagonal matrix having v on its main diagonal. For symmetric matrices $H \in \mathbb{R}^{n\times n}$ we order the eigenvalues by $\lambda_1(H) \leq \lambda_2(H) \leq \cdots \leq \lambda_n(H)$. Because the Laplacian L is positive semidefinite and $Le = 0$, we have $\lambda_1(L) = 0$ with eigenvector e. Fiedler [6, 7] showed that the second smallest eigenvalue $\lambda_2(L)$ is tightly related to edge and vertex connectivity of the graph. In particular, $\lambda_2(G)$ is positive if and only if G is connected. Therefore, Fiedler called $\lambda_2(L)$ the *algebraic connectivity* of the graph. Eigenvectors to $\lambda_2(L)$, often referred to as Fiedler vectors, have been used quite successfully in heuristics for graph partitioning in parallel computing [20, 21, 14], in clustering of geometric objects [1] or hyperlinks in the World Wide Web [12], or even in computer vision [15]. The second smallest eigenvalue allows one to derive various bounds in graph partitioning or bandwidth optimization [13, 11]; further properties of the Laplacian spectrum are presented in [8, 9, 10, 2, 29], and [17, 18] give a survey on the Laplacian spectrum of graphs. See also [3, 19] for related applications of spectral graph theory in combinatorial optimization.

*Received by the editors September 2, 2005; accepted for publication (in revised form) January 6, 2008; published electronically May 9, 2008.
http://www.siam.org/journals/siopt/19-1/63943.html
†Fakultät für Mathematik, Technische Universität Chemnitz, D-09107 Chemnitz, Germany (frank.goering@mathematik.tu-chemnitz.de, helmberg@mathematik.tu-chemnitz.de, markus.wappler@mathematik.tu-chemnitz.de).

By the Courant–Fischer theorem, the eigenvalue $\lambda_2(L)$ and its eigenvectors may be characterized as optimal solutions to the optimization problem

$$\lambda_2(L) = \min_{v \in \mathbb{R}^n, v^T e = 0, \|v\| = 1} v^T L v.$$

The usefulness of λ_2 and its eigenvectors in graph partitioning should relate to this characterization in some way. In order to get a better understanding of these connections, it seems natural to study the eigenspace of λ_2 for weighted matrices on the same support (i.e., arc weighted graphs on the same edge set) that are extremal in the sense that, for their distribution of the weight, λ_2 is maximal. The optimal λ_2 with respect to the support of the graph was introduced by Fiedler [7] under the name "absolute algebraic connectivity." We study the semidefinite dual of this optimization problem. Due to complementarity, the optimal solutions of the dual are restricted to the eigenspace of the optimal λ_2, and so all properties of dual optimal solutions directly provide information on the structure of the eigenspace associated with the absolute algebraic connectivity. It turns out that the dual may be interpreted as an embedding problem of the nodes of G in $\mathbb{R}^n$; see (4). The same optimization problem appears in [24] in connection with finding the fastest mixing Markov process on a graph; this work also mentions interest in low-dimensional solutions of this problem within the field of maximum variance unfolding in machine learning [26, 27].

We show that optimal embeddings of (4) have structural properties tightly connected to the separator structure of the graph (Theorem 3). In particular, if a subset $S \subset N$ of nodes separates the graph into two separated node sets C_1, C_2 that form a partition of $N \setminus S$, then for one of the two sets, say, C_1, all nodes are in the "shadow" of the convex hull of the nodes in S as seen from the origin; i.e., the straight line segment between any node of C_1 and the origin intersects the convex hull of S. Since any nonzero projection of the embedding to a one-dimensional subspace yields an eigenvector to the optimal λ_2 (Remark 2), this offers good geometric insight into the usefulness of Fiedler vectors for graph partitioning.

The embedding may also be interpreted as a variant of vector labelings of graphs as introduced in [16]. On first sight, strong similarities exist with respect to the Colin de Verdière number $\mu(G)$; see the excellent survey [25]. But the strong Arnold property is not required in our context, so no direct connection to $\mu(G)$ should be expected. Yet, similar to maximizing the corank in the Colin de Verdière number, one may ask for an optimal embedding of minimal dimension. Besides general interest in the existence of low-dimensional optimal solutions of semidefinite programs [22], such solutions are also sought in the machine learning applications [26, 27] mentioned above. Even though we are still far from answering this question to our full satisfaction, we are able to exhibit an intriguing bound based on the tree width of the graph. Indeed, we show in the proof of Theorem 5 that there is always an optimal embedding whose dimension is bounded by the cardinality of a "central" node of an arbitrary tree decomposition of G. This bound is tight for some particular graph classes (see Example 8). Nonetheless, the bound seems to be far too pessimistic, e.g., for planar graphs. Therefore it is conceivable that significantly better bounds can be obtained by minor related approaches.

The paper is organized as follows. In section 2 we derive the embedding problem as the dual problem to the eigenvalue optimization problem of determining the absolute algebraic connectivity and present an overview of our main results on this embedding together with some examples. The proof of the first result (the separator-shadow theorem) is given in section 3. Section 4 is devoted to optimality-preserving manipu-

lations of optimal embeddings for reducing the dimension of embeddings. These are rotations and foldings around separators that contain the origin in their convex hull and allow us, in section 5, to design an algorithm that gives rise to the proof of the tree width bound on the minimal dimension of optimal solutions.

We use basic notions and notation from graph theory and semidefinite programming [28]. In particular, for symmetric $H \in \mathbb{R}^{n\times n}$, $H \succeq 0$ is used to denote positive semidefiniteness; for matrices $A, B \in \mathbb{R}^{m\times n}$, $\langle A, B\rangle := \sum_{ij} A_{ij}B_{ij}$ is the canonical inner product; in the case of vectors $a, b \in \mathbb{R}^n$ we will simply use a^Tb; $\|\cdot\|$ refers to the usual Euclidean norm; e denotes the vector of all ones of appropriate size; for a set $\mathcal{S} \subset \mathbb{R}^n$, $\operatorname{conv}\mathcal{S}$ refers to the convex hull of $\mathcal{S}$ and $\operatorname{cone}\mathcal{S} := \{\lambda x : x \in \operatorname{conv}\mathcal{S}, \lambda \geq 0\}$. The projection on a closed convex set C is denoted by $p_C(\cdot)$.

2. Optimal embeddings and main results. In the remainder of the paper we assume that the graph $G = (N, E)$ is connected and $|N| = n \geq 2$. Let

$$\mathcal{C} := \left\{ c \in \mathbb{R}^E_+ : \sum_{ij\in E} c_{ij} = |E| \right\}$$

denote the set of all possible nonnegative edge weightings that sum up to $|E|$. For a particular $c \in \mathcal{C}$, let A_c denote the weighted adjacency matrix, i.e., $A_{ij} = c_{ij}$ for $ij \in E$ and 0 otherwise, and $L_c := \operatorname{diag}(A_ce) - A_c$ the corresponding weighted Laplacian. For $i, j \in N$, $i \neq j$, define $E_{ij} \in \mathbb{R}^{n\times n}$ as the matrix having the two diagonal elements $(E_{ij})_{ii} = (E_{ij})_{jj} = 1$, the two off-diagonal elements $(E_{ij})_{ij} = (E_{ij})_{ji} = -1$, and all other elements equal to zero. Then we may rewrite the Laplacian as

$$L_c = \sum_{ij\in E} c_{ij}E_{ij}.$$

The matrix L_c is positive semidefinite (because E_{ij} is positive semidefinite and $c_{ij} \geq 0$ for all $ij \in E$) and has an eigenvalue zero with eigenvector e (because $E_{ij}e = 0$). Our basic optimization problem is to determine the absolute algebraic connectivity

$$\hat{a}(G) := \max_{c\in\mathcal{C}} \lambda_2(L_c), \tag{1}$$

where $\hat{a}(G)$ denotes the absolute algebraic connectivity introduced in [7]. The maximum is attained, because a continuous function is maximized over a compact set. Since G is assumed to be connected, the result of Fiedler [6] for $c = e$ asserts that $\lambda_2(L_c) = \lambda_2(L) > 0$, so the optimum value is strictly positive. In order to reformulate the optimization problem as a semidefinite program, it will be convenient to shift the smallest eigenvalue 0 to a sufficiently large value. Thus, (1) may be rewritten as the following semidefinite program:

$$\begin{aligned} \hat{a}(G) = \max \quad & \lambda \\ \text{such that (s.t.)} \quad & \sum_{ij\in E} c_{ij}E_{ij} + \rho ee^T - \lambda I \succeq 0, \\ & \sum_{ij\in E} c_{ij} = |E|, \\ & c \geq 0,\ \lambda, \rho \text{ free}. \end{aligned}$$

Because the optimal value is strictly greater than zero by the connectedness of G, we may rescale the problem by $1/\lambda$ and equivalently minimize the sum of the scaled

weights $w_{ij} := c_{ij}/\lambda$ instead. Together with the scaled $\mu := \rho/\lambda$ we obtain

$$\begin{aligned} \frac{|E|}{\hat{a}(G)} = \min \quad & \sum_{ij\in E} w_{ij} \\ \text{s.t.} \quad & \sum_{ij\in E} w_{ij}E_{ij} + \mu ee^T \succeq I, \\ & w \geq 0,\ \mu \text{ free.} \end{aligned} \tag{2}$$

Note that, by the considerations above, choosing $w = \frac{1+\varepsilon}{\lambda_2(L)}e$ and $\mu = 1+\varepsilon$ yields a strictly feasible solution for $\varepsilon > 0$. Therefore the program attains its optimal solution, and semidefinite duality theory together with strict feasibility asserts that the optimal value of its dual semidefinite program is also attained. The dual reads

$$\begin{aligned} \frac{|E|}{\hat{a}(G)} = \max \quad & \langle I, X\rangle \\ \text{s.t.} \quad & \left\langle ee^T, X\right\rangle = 0, \\ & \langle E_{ij}, X\rangle \leq 1 \qquad \text{for } ij \in E, \\ & X \succeq 0. \end{aligned} \tag{3}$$

Now consider a Gram representation of X via a matrix $V \in \mathbb{R}^{n\times n}$ with $X = V^TV$, and denote column i of V by v_i, i.e., $V = [v_1, \ldots, v_n]$. Then

$$X_{ij} = v_i^Tv_j \quad \text{and} \quad \langle E_{ij}, X\rangle = \|v_i\|^2 - 2v_i^Tv_j + \|v_j\|^2 = \|v_i - v_j\|^2.$$

Since $0 = \left\langle ee^T, X\right\rangle = e^TXe = e^TV^TVe$ and $Ve = \sum v_i$, the dual semidefinite program (3) translates directly to

$$\begin{aligned} \frac{|E|}{\hat{a}(G)} = \max \quad & \sum_{i\in N} \|v_i\|^2 \\ \text{s.t.} \quad & \left(\sum_{i\in N} v_i\right)^2 = 0, \\ & \|v_i - v_j\|^2 \leq 1 \qquad \text{for } ij \in E, \\ & v_i \in \mathbb{R}^n \text{ for } i \in N. \end{aligned} \tag{4}$$

Thus, the dual problem to (1) is equivalent to finding an embedding of the nodes of the graph in n-space so that their barycenter is at the origin (we will call this the *equilibrium constraint*), the distances of adjacent nodes are bounded by one (the *distance constraints*), and the sum of their squared norms is maximized.

Remark 1. Together with the KKT conditions (we do not list feasibility constraints again)

$$\begin{aligned} v_j + \sum_{ij\in E} w_{ij}(v_i - v_j) + \mu\sum_{i\in N} v_i = 0 \qquad & \forall j \in N, \\ w_{ij}(1 - \|v_i - v_j\|^2) = 0 \qquad & \forall ij \in E, \\ \mu\left(\sum_{i\in N} v_i\right)^2 = 0, & \end{aligned}$$

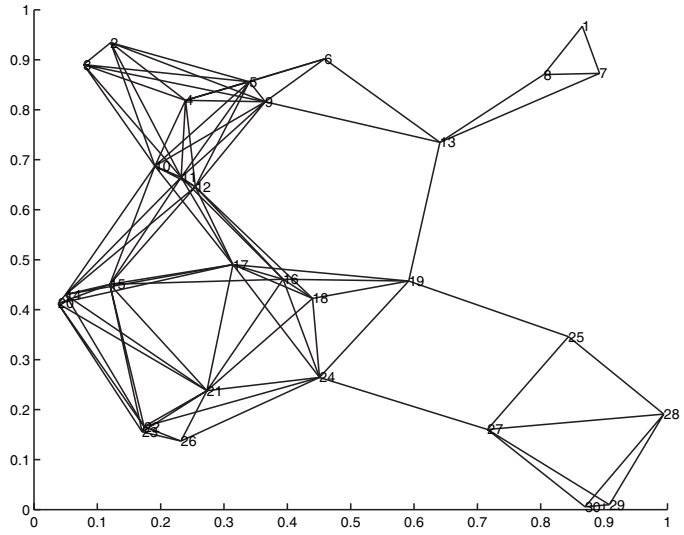


FIG. 1. *Original graph: The* 30 *vertices, picked randomly in* $[0,1]^2$*, are connected by an edge if the Euclidean distance is at most* 0.3.

the embedding problem suggests the following physical interpretation of optimal primal and dual solutions. Consider each node as having a point mass of unit size, and imagine each edge being a mass-free rope of length one that connects the points. Now the optimum solution of (4) corresponds to an equilibrium solution of this net spread within a force field that acts with force v on a point of mass one at position v. The w_{ij} are the forces acting along rope ij. Indeed, all $w_{ij} > 0$ are on the same scale as the force field, because $w_{ij} > 0$ only if $\|v_i - v_j\|^2 = 1$. So the first line of the KKT conditions asserts that these forces are in equilibrium in each point ($\mu \sum v_i = 0$ by feasibility, so this term does not enter). If an optimal two-dimensional embedding exists, such a physical situation is encountered when spreading the net on a disk rotating around its center (the centripetal force is $m\omega^2 r$, where m is the mass, ω the angular frequency, and r the radius). In [24] the same problem (and interpretation) was derived by starting from the problem of determining the fastest mixing Markov chain.

We illustrate this for an example graph on 30 vertices (see Figure 1) that was generated by picking the vertices randomly in the unit square and by connecting two points by an edge if their Euclidean distance is at most 0.3. The edge weights corresponding to an optimal solution of problem (2) are given in gray shades in Figure 2 (white is weight 0, black is maximum weight). The optimal embedding of (4) is displayed in Figure 3. It was computed by using SeDuMi [23] and is in fact two-dimensional in this case. The origin is indicated by the small circle in the center.

Remark 2. The projections of optimal embeddings onto one-dimensional subspaces yield eigenvectors to the algebraic connectivity. To see this, suppose that $V = [v_1, \ldots, v_n]$ is an optimal embedding of (4) and c are the corresponding optimal weights giving rise to the algebraic connectivity $\lambda_2(L_c)$ in (1). For any $p \in \mathbb{R}^n$, the vector $u := V^T p$ is an eigenvector of L_c for the eigenvalue $\lambda_2(L_c)$, i.e., $L_c u = \lambda_2(L_c)u$. Indeed, $X = V^T V$ is then an optimal solution of (3), $w = c/\lambda_2(L_c)$ together with some $\mu = 1$ is an optimal solution of (2), and by complementarity and $\langle ee^T, X \rangle = 0$

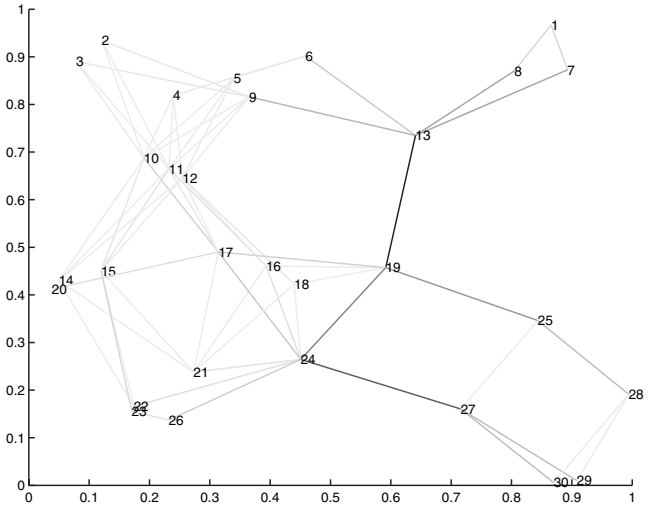


FIG. 2. *Graph with optimal edge weights.*

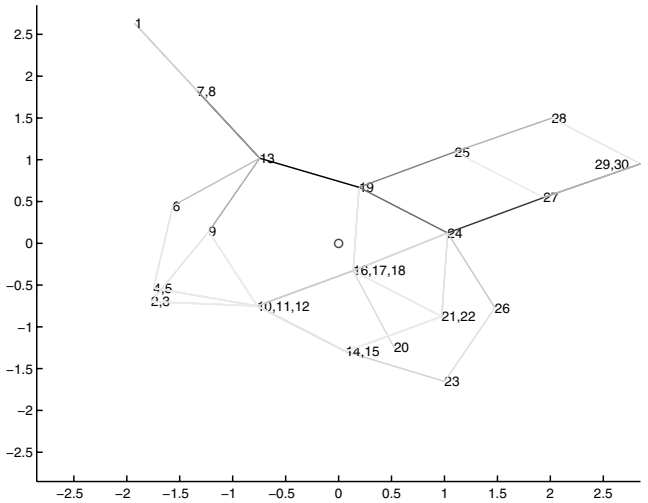


FIG. 3. *Optimal embedding (the central circle indicates the origin).*

we obtain

$$\begin{aligned} 0 &= \left\langle X, \sum_{ij\in E} w_{ij}E_{ij} + \mu ee^T - I \right\rangle \lambda_2(L_c) \\ &= \left\langle X, \sum_{ij\in E} \lambda_2(L_c) w_{ij}E_{ij} - \lambda_2(L_c)I \right\rangle \\ &= \left\langle V^TV, L_c - \lambda_2(L_c)I \right\rangle \\ &= \left\langle I, V(L_c - \lambda_2(L_c)I)V^T \right\rangle. \end{aligned}$$

So each column of V^T (and hence $u = V^Tp$) is in the eigenspace of L_c to eigenvalue $\lambda_2(L_c)$.

Our main results show that structural properties of optimal embeddings v_i, $i \in N$, of (4) are tightly linked to the separator structure of the underlying graph. Here a (node-)*separator* of G is a subset $S \subset N$ of nodes, whose removal disconnects the graph into at least two connected components. Often we will not discern every single component arising this way but simply speak of two or more separated sets of nodes. The first result states that for each separator S all but at most one of its components must be embedded so that any ray emanating from the origin first has to hit $\operatorname{conv}\{v_s : s \in S\}$ before it can reach a node of these components; i.e., by considering the origin as a light source and $\operatorname{conv}\{v_s : s \in S\}$ as a solid object, all but one of the components must be embedded in the shadow of the separator.

THEOREM 3 (separator-shadow). *Let $v_i \in \mathbb{R}^n$ ($i \in N$) be an optimal solution of (4) for a connected graph $G = (N, E)$. Let S be a separator in G giving rise to a partition $N = S \cup C_1 \cup C_2$ where there is no edge between C_1 and C_2. Then for at least one C_j*

$$\operatorname{conv}\{0, v_i\} \cap \operatorname{conv}\{v_s : s \in S\} \neq \emptyset \qquad \forall i \in C_j. \tag{5}$$

In words, the straight line segments $\operatorname{conv}\{0, v_i\}$ of all nodes $i \in C_j$ intersect the convex hull of the points in S.

We encourage the reader to check the separator-shadow property on some of the separators in Figure 3, e.g., for $S = \{13\}$ or $S = \{19, 24\}$.

Condition (5) holds trivially if $\operatorname{conv}\{v_s : s \in S\}$ contains the origin. Considering a separator S with the property $0 \notin \operatorname{conv}\{v_s : s \in S\}$, the separator-shadow theorem ensures that all but one of the components are embedded in the subspace spanned by the separator. Thus, if all minimal separators are small in size, there seems to be hope that there also exists an optimal embedding of small dimension. Our second main result confirms this expectation. In order to state it, we first recall the definitions of tree decomposition and tree width as given in [5].

DEFINITION 4. *For a graph $G = (N, E)$ a* tree decomposition *of G is a tree $(\mathcal{N}, \mathcal{E}) =: T$, with $\mathcal{N} \subseteq 2^N$ and $\mathcal{E} \subseteq \binom{\mathcal{N}}{2}$, satisfying the following requirements:*

(i) $N = \bigcup_{U \in \mathcal{N}} U$.

(ii) *For every $e \in E$ there is a $U \in \mathcal{N}$ with $e \subseteq U$.*

(iii) *If $U_1, U_2, U_3 \in \mathcal{N}$ with U_2 on the T-path from U_1 to U_3, then $U_1 \cap U_3 \subseteq U_2$.*

The width of T is the number $\max\{|U| - 1 : U \in \mathcal{N}\}$. The tree width *$tw(G)$ is the least width of any tree decomposition of G.*

For example, trees have tree width one (each edge forms one set U, so choose $\mathcal{N} = E$, and for $\mathcal{E}$ use the edge set of any spanning tree of the original tree's line graph). In general, it is NP-complete to determine the tree width, but any valid tree decomposition provides an upper bound.

THEOREM 5. *For each connected graph G there exists an optimal embedding of (4) of dimension at most $tw(G) + 1$.*

It is not difficult to devise examples where optimal embeddings of much higher dimensions exist, as well. A simple one, that will also be helpful in the remainder of the paper, is the star $K_{1,n}$.

Example 6. For a star $K_{1,n} := (\{0, 1, \ldots, n\}, \{\{0, i\} : i = 1, \ldots, n\})$, with $n \geq 2$, one optimal solution embeds the center node 0 in the origin and all other nodes in the vertices of a regular $n-1$-dimensional simplex with $\|v_i\| = 1$, $i = 1, \ldots, n$, for an objective value of n (optimality follows from choosing $w_{ij} = 1$ and $\mu = 1$ in (2)). For even $n \geq 2$, a one-dimensional optimal embedding is given by assigning the center node 0 to the origin, half of the outer nodes to $+1$, and the other half to -1. For odd $n \geq 3$, one possibility to find a two-dimensional optimal embedding is to put node 0

into the origin, node 1 to $(1,0)$ even nodes $i \geq 2$ to $(-\frac{1}{n-1}, \sqrt{1-(\frac{1}{n-1})^2})$, and the odd nodes $i \geq 3$ to $(-\frac{1}{n-1}, -\sqrt{1-(\frac{1}{n-1})^2})$.

Solving (3) by interior point methods will, in fact, always generate optimal embeddings of (4) of maximum dimension, because interior point methods generate maximally complementary solutions [4]. So the next question is whether it is difficult to find optimal embeddings satisfying the bound of Theorem 5. For general graphs there is little hope to find a tree decomposition giving the tree width of the graph, but for a given optimal embedding and some tree decomposition of width t our proof of Theorem 5 allows us to transform the embedding algorithmically by a sequence of optimality-preserving rotations and foldings into an optimal embedding of dimension at most $t+1$.

The bound of Theorem 5 on the minimum dimension of optimal embeddings is not tight for all graphs. Already in the example above, any star $K_{1,n}$ with even $n \geq 2$ has an optimal embedding in one dimension. For certain classes of graphs the bound of Theorem 5 on the minimum dimension of optimal embeddings is, in fact, far off (e.g., planar grid graphs have one-dimensional optimal embeddings), but in general the bound cannot be improved as is shown by the second of the following two examples.

Example 7 (complete graphs). For $K_n := (\{1,\ldots,n\}, \{\{i,j\} : 1 \leq i < j \leq n\})$ we show that the unique optimal embedding is the regular $n-1$-dimensional simplex with all points lying on the ball of radius $r_n := \sqrt{\frac{n-1}{2n}}$. The optimal X is given by $X_{ii} = r_n^2 = \frac{n-1}{2n}$ for $1 \leq i \leq n$ and $X_{ij} = X_{ji} = -\frac{r_n^2}{n-1} = -\frac{1}{2n}$ for $1 \leq i < j \leq n$, and the optimal weights are $w_{ij} = \frac{1}{n}$ for $1 \leq i < j \leq n$. By choosing $\mu = \frac{1}{n}$ we compute $L_w + \mu ee^T - I = 0$, so (w,μ) is feasible for (2) with objective $\frac{n-1}{2}$. Likewise, X is feasible for (3) and $\langle I, X\rangle = \frac{n-1}{2}$, so optimality is shown. Furthermore, since $w_{ij} > 0$ for all ij, the constraints $\langle E_{ij}, X\rangle = 1$ hold for all optimal X; i.e., the embedding must have all points pairwise at distance one. So the regular $n-1$-dimensional simplex is the only optimal embedding. Note that the tree width of K_n is $n-1$, and thus the complete graphs are not tight with respect to the bound of Theorem 5.

Example 8 (graphs with tight dimension bound). We append to K_n three independent vertices that are completely linked to K_n resulting in a graph $G(n) := (\{1,\ldots,n+3\}, \{\{i,j\} : 1 \leq i \leq n, i < j \leq n+3\})$. The tree width of $G(n)$ is n, and for $n \geq 4$ the minimal dimension of an optimal embedding of $G(n)$ is $n+1$. In fact, we show that, for $n \geq 4$, the vertices of K_n are again arranged as a centrally symmetric $n-1$-dimensional simplex with all points lying on a ball of radius $r_n := \sqrt{\frac{n-1}{2n}}$, and the three new points are arranged centrally symmetric on a circle orthogonal to this simplex with radius $\bar{r} := \sqrt{\frac{n+1}{2n}}$. The optimum of (3) is obtained by extending the optimum of Example 7 with $X_{ii} = \bar{r}^2 = \frac{n+1}{2n}$ for $n < i \leq n+3$, $X_{ij} = X_{ji} = -\frac{\bar{r}^2}{2} = -\frac{n+1}{4n}$ for $n < i < j \leq n+3$, and $X_{ij} = X_{ji} = 0$ for $1 \leq i \leq n$, $n < j \leq n+3$. The optimal weights are $w_{ij} = \frac{1}{n}$ for $1 \leq i \leq n$, $n < j \leq n+3$, and $w_{ij} = \frac{1}{n} - \frac{3}{n^2}$ for $1 \leq i < j \leq n$ (use Remark 1 and symmetry). By setting $\mu = \frac{1}{n}$ the slack matrix of (2) computes to

$$Z := L_w + \frac{1}{n} ee^T - I = \begin{bmatrix} \frac{3}{n^2} J_n & 0 \\ 0 & \frac{1}{n} J_3 \end{bmatrix} \succeq 0, \tag{6}$$

where J_k denotes the square matrix of all ones of order k. Therefore (w,μ) is feasible for (2), and the objective value is $3n\frac{1}{n} + \frac{n(n-1)}{2}(\frac{1}{n} - \frac{3}{n^2}) = 1 + \frac{n}{2} + \frac{3}{2n}$. Likewise, X is positive semidefinite because it is a Gram matrix. Furthermore, X satisfies all

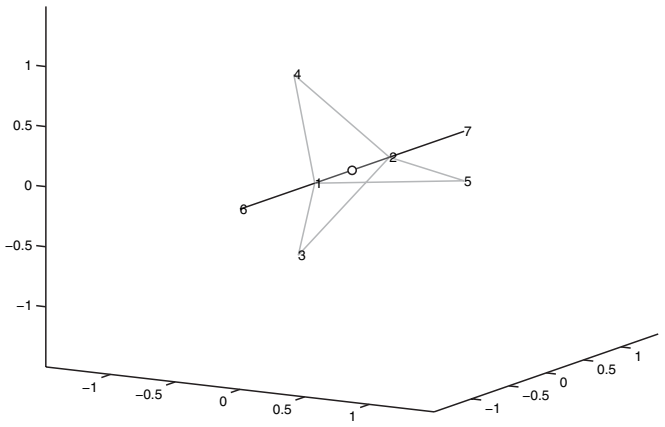


FIG. 4. *A graph with tree width* 2 *and optimal embedding of dimension at least three; see Example* 8 *(the central circle indicates the origin).*

distance constraints and has the same objective value. Hence the primal and the dual solution are optimal.

Now take any optimal embedding v_i, $i = 1, \ldots, n$, so $V = [v_1, \ldots, v_n]$. Since $w > 0$, all optimal embeddings must have all edge lengths equal to one: $\|v_i - v_j\| = 1$ for all $ij \in E(G(n))$. By (6) and semidefinite complementarity it holds that $\langle V^T V, Z \rangle = 0$; thus $\sum_{i=1}^{n} v_i = 0$ and $\sum_{i=n+1}^{n+3} v_i = 0$. So the embedding of K_n must be centrally symmetric like in Example 7, and by the distance constraints each of the three additional vertices must be embedded orthogonal to the embedding of K_n with distance $\bar{r}$ to the origin. As the three vectors have to sum up to zero, this can be done only in two additional dimensions. This completes the proof.

For $n = 1$ the construction yields a star with one central and three exterior nodes, and the bound is also tight. For $n = 2$ the embedding described above is not optimal (it would collapse to the image of the star), for $n = 3$ the embedding is optimal but not of minimal dimension. Without going into details, the cases $n = 2, 3$ can be extended to tight examples by appending to each node of K_n yet another node by a single edge; see Figure 4 for an illustration of the resulting embedding for $n = 2$.

3. The proof of the separator-shadow Theorem 3. Our proof of the separator-shadow theorem will be indirect. Given a feasible embedding that does not satisfy the statement of the theorem, we improve it by folding appropriate components out of the current space in opposite directions (see Figures 6 and 7 below). This requires some preparations. First note that a feasible embedding cannot be full-dimensional, and so there is always space for folding.

OBSERVATION 9. *Given $v_i \in \mathbb{R}^n$ $(i \in N)$ feasible for* (4)*, there is a vector $h \in \mathbb{R}^n$, $\|h\| = 1$, with $v_i \in \mathcal{H} := \{x \in \mathbb{R}^n : h^T x = 0\}$ for $i \in N$.*

Proof. The n vectors v_i satisfy $\sum_{i \in N} v_i = 0$, so they are linearly dependent, and therefore $\dim(\operatorname{span}\{v_1, \ldots, v_n\}) \leq n - 1$. ∎

Given the linear subspace $\mathcal{H} := \{x \in \mathbb{R}^n : h^T x = 0\}$, a normalized $b \in \mathcal{H}$, and some $\beta \in \mathbb{R}$, we next describe the operation of folding the flat half-space $\{x \in \mathcal{H} : b^T x < \beta\}$ along the affine subspace $\mathcal{B} := \{x \in \mathbb{R}^n : h^T x = 0, b^T x = \beta\}$ by rotating it around $\mathcal{B}$ into direction h by an angle γ and show that distances between folded points are not longer than before. The latter fact will help to ensure feasibility with respect to the distance constraints of (4). In stating this operation we make use of

the fact that, due to $\|h\| = \|b\| = 1$ and $h^T b = 0$, the projection of a point $x \in \mathbb{R}^n$ onto $\mathcal{B}$ is computed by

$$p_{\mathcal{B}}(x) = x + (\beta - b^T x)b - h^T x h. \tag{7}$$

Therefore the rotation of $x \in \mathcal{H}$ around $\mathcal{B}$ uses the radius $\|x - p_{\mathcal{B}}(x)\| = |\beta - b^T x|$.

OBSERVATION 10 (folding a flat half-space). *Given $h \in \mathbb{R}^n$, with $\|h\| = 1$, let $\mathcal{H} := \{x \in \mathbb{R}^n : h^T x = 0\}$, and given $b \in \mathcal{H}$, with $\|b\| = 1$, $\beta \in \mathbb{R}$, let $\mathcal{B} := \{x \in \mathcal{H} : b^T x = \beta\}$. Define the continuous map $\varphi : \mathcal{H} \times [-\pi, \pi] \to \mathbb{R}^n$ by*

$$\varphi(x, \gamma) := \begin{cases} p_{\mathcal{B}}(x) - (\beta - b^T x)[b \cos\gamma + h \sin\gamma] & \text{if } b^T x < \beta, \\ x & \text{if } b^T x \geq \beta. \end{cases}$$

For all $\gamma \in [-\pi, \pi]$ and all $x \in \mathcal{H}$,

(i) *$p_{\mathcal{B}}(x) = p_{\mathcal{B}}(\varphi(x, \gamma))$ and $\|x - p_{\mathcal{B}}(x)\| = \|\varphi(x, \gamma) - p_{\mathcal{B}}(x)\|$,*
(ii) *$\|\varphi(x, \gamma) - \varphi(y, \gamma)\| = \|x - y\|$ for all $y \in \mathcal{B}$,*
(iii) *$\|\varphi(x, \gamma) - \varphi(y, \gamma)\| \leq \|x - y\|$ for all $y \in \mathcal{H}$.*

Proof. (i) follows by direct calculation from (7).

If $b^T x \geq \beta$ and $b^T y \geq \beta$, the points are not transformed. If $b^T x < \beta$ and $b^T y \leq \beta$, both points are subject to the same orthogonal transformation which preserves distances. This implies (ii). If, without loss of generality (w.l.o.g.), $b^T x < \beta \leq b^T y$, the intersection of the line segment between x and y and $\mathcal{B}$ determines a unique point $z \in \mathcal{B} \cap \operatorname{conv}\{x, y\}$. The triangle inequality and (ii) yield $\|\varphi(x, \gamma) - y\| \leq \|\varphi(x, \gamma) - z\| + \|z - y\| = \|x - z\| + \|z - y\| = \|x - y\|$, so (iii) holds. ∎

Next, we need to trace the objective value as we fold a subset of nodes. Any such operation can be viewed as a combination of a rotation around the barycenter of the nodes and a uniform translation without rotation. The following two observations show that rotations around the barycenter do not affect the cost function, while the change induced by a translation is easily tracked via the barycenter alone.

OBSERVATION 11 (rotation around the barycenter). *Given $v_i \in \mathbb{R}^n$ ($i \in C \subseteq N$), let $\bar{v} := \frac{1}{|C|} \sum_{i \in C} v_i$, and set $v_i' := Q(v_i - \bar{v}) + \bar{v}$, where Q is an orthogonal matrix. Then*

$$\sum_{i \in C} \|v_i'\|^2 = \sum_{i \in C} \|v_i\|^2 .$$

Proof.

$$\begin{aligned} \sum_{i \in C} \|v_i'\|^2 &= \sum_{i \in C} \|Q(v_i - \bar{v})\|^2 + 2\bar{v}^T Q \underbrace{\sum_{i \in C} (v_i - \bar{v})}_{=0} + |C| \, \|\bar{v}\|^2 \\ &= \sum_{i \in C} \|v_i - \bar{v}\|^2 + 2\bar{v}^T \sum_{i \in C} (v_i - \bar{v}) + |C| \, \|\bar{v}\|^2 \\ &= \sum_{i \in C} \|v_i - \bar{v} + \bar{v}\|^2 \;\; = \;\; \sum_{i \in C} \|v_i\|^2 . \end{aligned}$$ ∎

OBSERVATION 12 (translation). *Given $d \in \mathbb{R}^n$ and $v_i \in \mathbb{R}^n$ ($i \in C \subseteq N$), let $v_i' := v_i + d$ ($i \in C \subseteq N$) and $\bar{v} := \frac{1}{|C|} \sum_{i \in C} v_i$. Then*

$$\sum_{i \in C} \|v_i'\|^2 = \sum_{i \in C} \|v_i\|^2 + |C|(2\bar{v} + d)^T d.$$

Proof. $\sum_{i \in C} \|v_i + d\|^2 = \sum_{i \in C} \|v_i\|^2 + 2|C|\bar{v}^T d + |C| d^T d.$ ∎

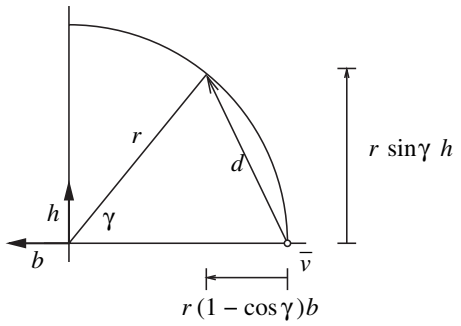


FIG. 5. *Rotation around an affine subspace.*

By putting these together, we now describe the cost change arising in folding a subset of the nodes.

OBSERVATION 13 (the cost of folding). *For h, b, $\beta > 0$, $\mathcal{H}$, $\mathcal{B}$, and φ as in Observation* 10 *and given $v_i \in \{x \in \mathcal{H} : b^T x < \beta\}$ $(i \in C \subseteq N)$, let $\bar{v} := \frac{1}{|C|}\sum_{i\in C} v_i$, $\gamma \in [-\pi, \pi]$, and set for $i \in C$*

$$v_i' := \varphi(v_i, \gamma).$$

Then

$$\sum_{i\in C} \|v_i'\|^2 = \sum_{i\in C} \|v_i\|^2 + 2|C|r(1-\cos\gamma)\beta, \qquad \text{where } r := \beta - b^T\bar{v} > 0.$$

Proof. The rotation around $\mathcal{B}$ may be split into a rotation of the points in C around their barycenter $\bar{v}$ as in Observation 11 and a translation as analyzed in Observation 12. The corresponding displacement for rotating $\bar{v}$ around $\mathcal{B}$ by angle γ is $d := r(\sin\gamma)h + r(1-\cos\gamma)b$, where $r = \beta - b^T\bar{v} > 0$ is the radius (see Figure 5). By $\bar{v}^T h = 0$, $b^T h = 0$, and Observation 12 the cost function changes by

$$\begin{aligned}
|C|(2\bar{v}^T d + d^T d) &= |C|\left(2r(1-\cos\gamma)\bar{v}^T b + r^2[\sin^2\gamma + (1-\cos\gamma)^2]\right)\\
&= |C|\left(2r(1-\cos\gamma)\bar{v}^T b + r^2[2 - 2\cos\gamma]\right)\\
&= 2|C|r(1-\cos\gamma)(\bar{v}^T b + r)\\
&= 2|C|r(1-\cos\gamma)\beta. \qquad \square
\end{aligned}$$

Proof of Theorem 3. Let $h \in \mathbb{R}^n$, with $\|h\| = 1$, satisfy $h^T v_i = 0$ for all $i \in N$ as in Observation 9, and let $\mathcal{S} := \operatorname{conv}\{v_s : s \in S\}$. Assume, for contradiction, that the theorem is not true. Then there is a node in C_1 (call it node 1) and a node in C_2 (call it node 2) embedded in v_1 and v_2, respectively, that satisfy $\operatorname{conv}\{0, v_1\} \cap \mathcal{S} = \operatorname{conv}\{0, v_2\} \cap \mathcal{S} = \emptyset$. By convex separation each set $\operatorname{conv}\{0, v_j\}$ can be separated from $\mathcal{S}$ by a separating hyperplane within the subspace $\operatorname{span}\{v_i : i \in N\}$. So for $j \in \{1, 2\}$ there are vectors $b_j \in \operatorname{span}\{v_i : i \in N\}$ (these satisfy $b_j^T h = 0$) and scalars $\beta_j > 0$ so that $b_j^T x \geq \beta_j$ for all $x \in \mathcal{S}$ and $b_j^T x < \beta_j$ for all $x \in \operatorname{conv}\{0, v_j\}$.

Next we show that we can find a convex combination of these two inequalities by choosing an appropriate $\alpha \in [0, 1]$ so that, for $b(\alpha) := (1-\alpha)b_1 + \alpha b_2$ and $\beta(\alpha) := (1-\alpha)\beta_1 + \alpha\beta_2$, the open half-space $\{x : b(\alpha)^T x < \beta(\alpha)\}$ contains points of both C_1 and C_2 (illustrated in Figure 6). Indeed, for $\alpha = 0$ the half-space contains v_1 and so a point of C_1, for $\alpha = 1$ it contains v_2 which belongs to C_2, and it contains the

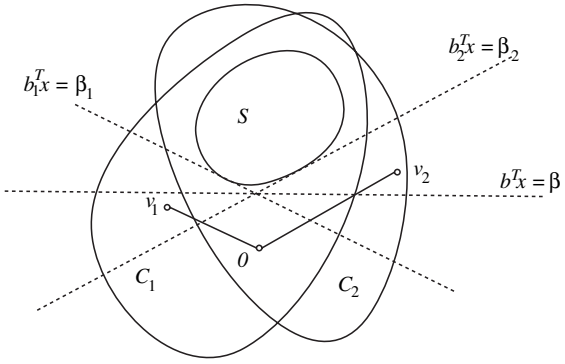


FIG. 6. *Initial setting in case* (i) *of the separator-shadow proof.*

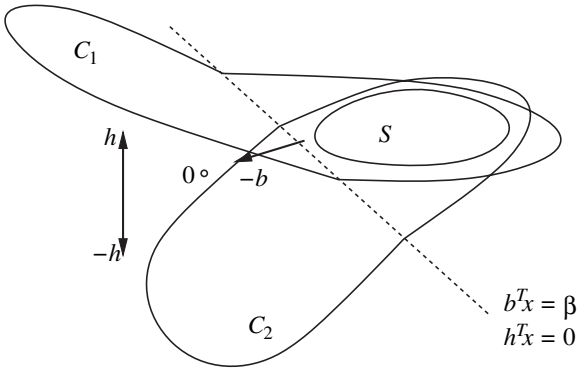


FIG. 7. *Improving movement in case* (i) *of the separator-shadow proof.*

origin for all $\alpha \in [0,1]$. Suppose, for contradiction, that in sweeping α through $[0,1]$ the half-space loses the last point of C_1 before it encounters the first point of C_2 at some particular $\bar{\alpha}$. Then the corresponding hyperplane defined by $b(\bar{\alpha})^T x = \beta(\bar{\alpha}) > 0$ would separate 0 strictly from $\text{conv}\{v_i : i \in N\}$, but this contradicts the feasibility of the v_i as the origin is a convex combination of the v_i by the equilibrium constraint $\frac{1}{n}\sum_{i\in N} v_i = 0$.

Thus we have found $b := b(\alpha)$ and $\beta := \beta(\alpha) > 0$ such that the open half-space $\{x : b^T x < \beta\}$ contains points from C_1 and C_2. Note that $b^T h = 0$ holds, and by scaling b and β we may assume w.l.o.g. that $\|b\| = 1$. Let, for $j \in \{1,2\}$, $M_j := \{i \in C_j : b^T v_i < \beta\}$, $m_j := |M_j| > 0$, and $\bar{v}_j := \frac{1}{m_j}\sum_{i\in M_j} v_i$. Next, consider rotating independently for each j the points in M_j around the affine subspace $\mathcal{B} = \{x \in \mathbb{R}^n : h^T x = 0, b^T x = \beta\}$ as specified in Observation 13. Because the points in M_1 and M_2 are not adjacent and distances to the remaining points are not increased by Observation 10(iii), the edge constraints in (4) remain satisfied. We show that rotating the points in M_1 in direction h and the points in M_2 against direction h by sufficiently small angles γ_1 and γ_2 improves the solution (see Figure 7). As in the proof of Observation 13, denote, for $j \in \{1,2\}$, the radius and displacement of $\bar{v}_j$ by

$$r_j := \beta - b^T \bar{v}_j > 0 \qquad \text{and} \qquad d_j := r_j[(\sin\gamma_j)h + (1-\cos\gamma_j)b],$$

respectively, yielding the improvement $2m_j r_j(1-\cos\gamma_j)\beta$. Rotation j adds $m_j d_j$ to

the barycenter of all points and has to be compensated in order to maintain feasibility with respect to the equilibrium constraint. Shifts of the global barycenter in the direction of h can be avoided by requiring $m_1 d_1^T h = -m_2 d_2^T h$; i.e., given γ_1, choose γ_2 in dependence of γ_1 so that $m_1 r_1 \sin\gamma_1 = -m_2 r_2 \sin\gamma_2$. After carrying out these rotations it therefore remains to shift all points by

$$d := -(m_1 d_1^T b + m_2 d_2^T b)b/n = -[m_1 r_1(1-\cos\gamma_1) + m_2 r_2(1-\cos\gamma_2)]b/n$$

for feasibility in (4). By using Observation 12, the total objective improvement is

$$\begin{aligned}&\sum_{j\in\{1,2\}} 2m_j r_j(1-\cos\gamma_j)\beta - nd^T d\\ &= \sum_{j\in\{1,2\}} 2m_j r_j(1-\cos\gamma_j)\beta - \frac{1}{n}[m_1 r_1(1-\cos\gamma_1) + m_2 r_2(1-\cos\gamma_2)]^2.\end{aligned}$$

This is positive for γ_1 and $\gamma_2(\gamma_1)$ close enough to zero, yielding a contradiction to the optimality of the embedding. ∎

4. Separators containing the origin. The freedom for squeezing optimal embeddings into lower dimensions that will be needed for the proof of Theorem 5 in section 5 is offered by separators that contain the origin in the convex hull of their embedded nodes. Example 6 of the star $K_{1,n}$ may help to illustrate the main idea of the transformations we employ. By alluding to the physical interpretation, we will proceed in two steps:

Step one: We rearrange the cumulated force vectors of the separated node sets so that they are balanced in just one or two additional dimensions with respect to this central separator.

Step two: We show how to combine this with reducing the dimension of each component.

The result will be that either we find a particularly large component that governs the dimension of the entire embedding or no such component exists and we succeed in flattening the embedding to a space exceeding the dimension of the separator by at most two.

We start with an optimal embedding v_i ($i \in N$) of G and, by using Observation 9, fix some $h \in \mathbb{R}^n$, $\|h\| = 1$, so that

$$\{v_i : i \in N\} \subset \mathcal{H} := \{x \in \mathbb{R}^n : h^T x = 0\}.$$

Throughout this section we assume that $S \subset N$ is a separator in G satisfying

$$0 \in \mathcal{S} := \operatorname{conv}\{v_i : i \in S\}$$

and separating G into m sets

$$C_j \subset N, \quad j \in M := \{1, \dots, m\}.$$

Together with S, they form a partition of N, and each edge of G is incident to at most one of the sets C_j. Sometimes we also use sets C_j that contain more than one of the separated connected components. For each $j \in M$, the *cumulated vector* is denoted by

$$\bar{v}_j := \sum_{i\in C_j} v_i.$$

We will not modify the embedding on the linear subspace

$$\mathcal{L} := \operatorname{span} \mathcal{S}$$

spanned by the vectors of the separator. Modifications will be restricted to its orthogonal complement $\mathcal{L}^\perp$, so mostly our illustrations are given with respect to the embedding obtained by projecting the v_i onto $\mathcal{L}^\perp$. In the projected embedding $p_{\mathcal{L}^\perp}(v_i)$ $(i \in N)$, all nodes $i \in S$ are embedded in the origin, and, like in the case of the star, the projected cumulated vectors $p_{\mathcal{L}^\perp}(\bar{v}_j)$, $j \in M$, pointing out of the origin in various directions, are in equilibrium, i.e., $\sum_{j\in M} p_{\mathcal{L}^\perp}(\bar{v}_j) = 0$ by feasibility. We note for later use that, in any such configuration, none of the vectors may be longer than the sum of the others. Indeed, set

$$\bar{\delta}_j := \|p_{\mathcal{L}^\perp}(\bar{v}_j)\| \quad \text{for } j \in M,$$

and then $\sum_{j\in M} p_{\mathcal{L}^\perp}(\bar{v}_j) = 0$ implies that

$$\sum_{j\in M\setminus\{\hat{\jmath}\}} \bar{\delta}_j \geq \bar{\delta}_{\hat{\jmath}} \quad \text{for all } \hat{\jmath} \in M. \tag{8}$$

The following fundamental fact will be used repeatedly (the equilibrium constraint may get violated initially, but this will be taken care of later). For each $j \in M$ the vector $p_{\mathcal{L}^\perp}(\bar{v}_j)$ can be rotated around the origin freely within $\mathcal{L}^\perp$ while preserving all distances between nodes in $C_j \cup S$ by applying to all $p_{\mathcal{L}^\perp}(v_i)$, $i \in C_j$, an orthogonal transformation Q_j with $\mathcal{L}$ contained in its invariant subspace (i.e., Q_j restricted to $\mathcal{L}$ is the identity). Furthermore, such transformations do not influence the objective value, as distances to $0 \in \mathcal{L}$ are preserved. We complete step one by showing that the vectors $p_{\mathcal{L}^\perp}(\bar{v}_j)$ with their lengths $\bar{\delta}_j$ can always be rotated into at most three normalized directions d_1, d_2, d_3 so that the equilibrium constraint holds again in $\mathcal{L}^\perp$ (by definition, the equilibrium constraint stays valid within $\mathcal{L}$).

OBSERVATION 14. *Given scalars $\bar{\delta}_j \geq 0$ for $j \in M = \{1, \dots, m\}$, $m \geq 2$, so that, for each $\hat{\jmath} \in M$, $\sum_{j\in M\setminus\{\hat{\jmath}\}} \bar{\delta}_j \geq \bar{\delta}_{\hat{\jmath}}$. There exist vectors $d_1, d_2, d_3 \in \mathbb{R}^2$ with $\|d_1\| = \|d_2\| = \|d_3\| = 1$ and an assignment $\kappa : M \to \{1, 2, 3\}$ so that $\sum_{j\in M} \bar{\delta}_j d_{\kappa(j)} = 0$. This also holds if in addition $|\{j \in M : \kappa(j) = 1\}| = 1$ is required.*

Proof. If $|M| = 2$, then $\bar{\delta}_1 = \bar{\delta}_2$, and the claim holds for $d_1 = -d_2$ and κ correspondingly. Otherwise let $\hat{\jmath} \in M$ be the smallest number so that $\sum_{j=1}^{\hat{\jmath}-1} \bar{\delta}_j < \frac{1}{2}\sum_{j\in M} \bar{\delta}_j \leq \sum_{j=1}^{\hat{\jmath}} \bar{\delta}_j$, and set $\kappa(\hat{\jmath}) = 1$, $\kappa(j) = 2$ for $\hat{\jmath} > j \in M$ and $\kappa(j) = 3$ for $\hat{\jmath} < j \in M$. Set $\check{\delta}_h := \sum_{j\in M,\kappa(j)=h} \bar{\delta}_j$, $h \in \{1, 2, 3\}$. Note that $\check{\delta}_1 \leq \check{\delta}_2 + \check{\delta}_3$, $\check{\delta}_2 \leq \check{\delta}_1 + \check{\delta}_3$, $\check{\delta}_3 \leq \check{\delta}_1 + \check{\delta}_2$. Assume, w.l.o.g., that $\check{\delta}_1 \leq \check{\delta}_2 \leq \check{\delta}_3$. Set $d_1(\alpha) := (\cos\alpha, -\sin\alpha)^T$ for $0 \leq \alpha \leq \pi$, $d_2(\alpha) := (\cos\gamma(\alpha), \sin\gamma(\alpha))^T$, where $\gamma(\alpha)$ is defined implicitly by $\check{\delta}_2 \sin\gamma(\alpha) = \check{\delta}_1 \sin\alpha$, and $d_3 = (-1, 0)^T$. Then $b(\alpha) := \check{\delta}_1 d_1(\alpha) + \check{\delta}_2 d_2(\alpha) + \check{\delta}_3 d_3$ satisfies $[b(\alpha)]_2 = 0$ for all $0 \leq \alpha \leq \pi$, $[b(0)]_1 \geq 0$, and $[b(\pi)]_1 \leq 0$, so by continuity of $b(\alpha)$ there is an $\hat{\alpha} \in [0, \pi]$ with $b(\hat{\alpha}) = 0$. ∎

Let us now turn towards step two and try to reduce the dimension of the node sets. If $\operatorname{span}\{v_i : i \in C_j\} \subseteq \mathcal{L}$ for $j \in M$, then the embedding is good enough for our purposes. Assume therefore that there is some $j \in M$ with $\operatorname{span}\{v_i : i \in C_j\} \not\subseteq \mathcal{L}$. In manipulating the embedding of C_j in Observations 15–19 we will again apply only orthogonal transformations (sometimes we will simultaneously use separate ones for each point in C_j) that contain $\mathcal{L}$ in their invariant subspace. Therefore all distances of points in C_j will preserve their distance to the origin and to the embedding of S. In

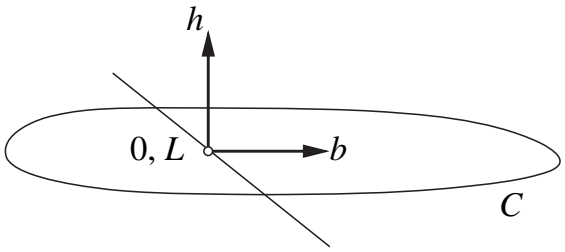


FIG. 8. *Initial setting before the transformation of C in Observations* 15–19.

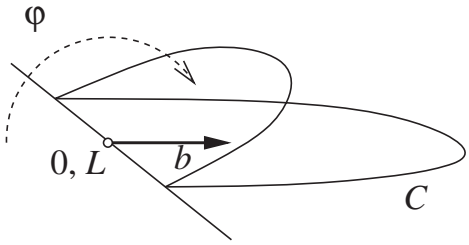


FIG. 9. *φ folds C into the half-space specified by b (Observation* 15*).*

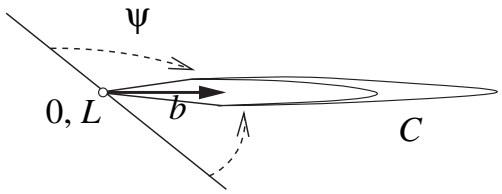


FIG. 10. *ψ collapses C into the flat half-space spanned by $\mathcal{L}$ and direction b (Observation* 17*).*

consequence, optimality is guaranteed if feasibility can be maintained. In particular, feasibility of the distance constraints is ensured whenever distances within C_j are not increased. Our manipulations may, however, increase the length of $p_{\mathcal{L}^\perp}(\bar{v}_j)$ and thus $\bar{\delta}_j$. But by Observation 14 it suffices that condition (8) is satisfied at the end in order to restore the equilibrium constraint, as well.

Before giving the transformations in detail, we outline the main idea by sketching the underlying geometric intuition. The goal is to squeeze the entire embedding of component C_j into the flat half-space spanned by $\mathcal{L}$ and one additional direction $b_j \in \mathcal{H} \cap \mathcal{L}^\perp$, with $\|b_j\| = 1$. This works as follows. The transformation of Observation 15 folds all nodes into the flat half-space $\{x \in \mathcal{H} : b^T x \geq 0\}$ via Observation 10 (put $b = b_j$ and $\beta = 0$)—see Figures 8 and 9; this leaves $\mathcal{L} \subseteq \mathcal{B}$ untouched as required. Then the transformation of Observation 17 collapses this flat half-space into the even flatter half-space cone$(\mathcal{L} \cup \{b_j\})$ as if collapsing an umbrella by rotating the ribs towards its handle (see Figure 10). In Observation 19 these two operations are concatenated to a continuous transformation $u_i(t)$ of the embedding for $t \in [0, 1]$, and we will see via Observations 16 and 18 that the norm $\|p_{\mathcal{L}^\perp}(\bar{u}_j(t))\|$ of the cumulated vector $\bar{u}_j(t) = \sum_{i \in C_j} u_i(t)$ is nondecreasing throughout, so that we can easily stop the transformation at an appropriate t to ensure condition (8). We start with the folding operation.

OBSERVATION 15 (transformation part 1, folding). *Given $j \in M$ and $b_j \in \mathcal{H} \cap \mathcal{L}^\perp$, with $\|b_j\| = 1$, define $\varphi_i : [0,1] \to \mathbb{R}^n$ for $i \in C_j$ by*

$$\varphi_i(t) := \begin{cases} v_i - (v_i^T b_j) b_j + (v_i^T b_j)[b_j \cos t\pi + h \sin t\pi] & \text{if } v_i^T b_j < 0, \\ v_i & \text{if } v_i^T b_j \geq 0. \end{cases}$$

Then, for $i \in C_j$,

(i) $\varphi_i(0) = v_i$,

(ii) $\varphi_i(1) \in \{x \in \mathcal{H} : b_j^T x \geq 0\}$,

and for all $t \in [0,1]$ it holds that

(iii) $p_{\mathcal{L}}(v_i) = p_{\mathcal{L}}(\varphi_i(t))$ *and* $\|p_{\mathcal{L}^\perp}(v_i)\| = \|p_{\mathcal{L}^\perp}(\varphi_i(t))\|$,

(iv) $\|\varphi_i(t) - v\| = \|v_i - v\|$ *for* $v \in \mathcal{L} \supseteq \{v_s : s \in S\}$,

(v) $\|\varphi_i(t) - \varphi_k(t)\| \leq \|v_i - v_k\|$ *for* $k \in C_j$.

Proof. (i) and (ii) follow from direct calculation and $v_i \in \mathcal{H}$; (iii)–(v) follow from Observation 10(i)–(iii) by using $b = b_j$, $\beta = 0$, $\varphi_i(t) = \varphi(v_i, t\pi)$, and the fact that $\mathcal{S} \subseteq \mathcal{L} \subseteq \mathcal{B}$. ∎

Next, we show that throughout this first transformation the length of the projected cumulated vector increases.

OBSERVATION 16. *For φ_i ($i \in C_j$) as defined in Observation 15, define $\bar{\varphi}_j : [0,1] \to \mathbb{R}^n$ by*

$$\bar{\varphi}_j(t) := \sum_{i \in C_j} \varphi_i(t).$$

The length $\|p_{\mathcal{L}^\perp}(\bar{\varphi}_j(t))\|$ is nondecreasing in $t \in [0,1]$.

Proof. The choice of b_j ensures that $\mathcal{B} := \{x \in \mathcal{H} : b_j^T x = 0\} \supseteq \mathcal{L}$ and $\mathcal{B}^\perp = \operatorname{span}\{h, b_j\}$. By definition of the φ_i in Observation 15 we obtain

$$\begin{aligned} \|p_{\mathcal{L}^\perp}(\bar{\varphi}_j(t))\|^2 &= \|p_{\mathcal{L}^\perp}(p_{\mathcal{B}}(\bar{\varphi}_j(0)))\|^2 + \|p_{\mathcal{B}^\perp}(\bar{\varphi}_j(t))\|^2 \\ &= \|p_{\mathcal{L}^\perp}(p_{\mathcal{B}}(\bar{\varphi}_j(0)))\|^2 \\ &\quad + \left\| \sum_{i \in C_j, v_i^T b_j < 0} (v_i^T b_j)[b_j \cos t\pi + h \sin t\pi] + \sum_{i \in C_j, v_i^T b_j \geq 0} (v_i^T b_j) b_j \right\|^2. \end{aligned}$$

As b_j and h are orthogonal it remains to study the monotonicity of

$$\begin{aligned} &\left[\sum_{i \in C_j, v_i^T b_j < 0} v_i^T b_j \cos t\pi + \sum_{i \in C_j, v_i^T b_j \geq 0} v_i^T b_j \right]^2 + \left[\sum_{i \in C_j, v_i^T b_j < 0} v_i^T b_j \sin t\pi \right]^2 \\ &= \left[\sum_{i \in C_j, v_i^T b_j < 0} v_i^T b_j \right]^2 (\cos^2 t\pi + \sin^2 t\pi) + \left[\sum_{i \in C_j, v_i^T b_j \geq 0} v_i^T b_j \right]^2 \\ &\quad + 2 \underbrace{\left[\sum_{i \in C_j, v_i^T b_j < 0} v_i^T b_j \right] \left[\sum_{i \in C_j, v_i^T b_j \geq 0} v_i^T b_j \right]}_{\leq 0} \cos t\pi. \end{aligned}$$

The last term is clearly nondecreasing. ∎

The collapsing transformation starts from the points $\varphi_i(1)$ and runs as follows.

OBSERVATION 17 (transformation part 2, collapsing). *Given the setting of Observation* 15, *for* $i \in C_j$, *set* $\delta_i := \|p_{\mathcal{L}^\perp}(\varphi_i(1))\|$, *determine* $0 \le \gamma_i \le \frac{\pi}{2}$ *and* $q_i \in \mathcal{L}^\perp$, $q_i^T b_j = 0$, $\|q_i\| = 1$ *so that* $p_{\mathcal{L}^\perp}(\varphi_i(1)) = \delta_i(q_i \cos\gamma_i + b_j \sin\gamma_i)$, *and define* $\psi_i : [0,1] \to \mathbb{R}^n$ *by*

$$\psi_i(t) := p_{\mathcal{L}}(\varphi_i(1)) + \delta_i \left[q_i \cos\left(\gamma_i + t\left[\frac{\pi}{2} - \gamma_i\right]\right) + b_j \sin\left(\gamma_i + t\left[\frac{\pi}{2} - \gamma_i\right]\right)\right].$$

Then, for $i \in C_j$,

(i) $\psi_i(0) = \varphi_i(1)$,

(ii) $\psi_i(1) = p_{\mathcal{L}}(v_i) + \|p_{\mathcal{L}^\perp}(v_i)\| b_j \in \mathcal{L} + \{\beta b_j : \beta \ge 0\}$,

and for all $t \in [0,1]$ *it holds that*

(iii) $p_{\mathcal{L}}(v_i) = p_{\mathcal{L}}(\psi_i(t))$ *and* $\|p_{\mathcal{L}^\perp}(v_i)\| = \|p_{\mathcal{L}^\perp}(\psi_i(t))\|$,

(iv) $\|\psi_i(t) - v\| = \|v_i - v\|$ *for* $v \in \mathcal{L} \supseteq \{v_s : s \in S\}$,

(v) $\|\psi_i(t) - \psi_k(t)\| \le \|v_i - v_k\|$ *for* $k \in C_j$.

Proof. First note that Observation 15(iii) implies that $p_{\mathcal{L}}(v_i) = p_{\mathcal{L}}(\varphi_i(1))$ and $\delta_i = \|p_{\mathcal{L}^\perp}(v_i)\|$. Now (i) and (ii) follow from direct calculation, and (iii) and (iv) are proved in the same way as (i) and (ii) of Observation 10. It remains to prove (v).

Because of Observation 15(v) it suffices to prove that $\|\psi_i(t) - \psi_k(t)\|^2 \le \|\varphi_i(1) - \varphi_k(1)\|^2$ for $i, k \in C_j$. For this we need to show that $\psi_i(t)^T \psi_k(t) \ge \varphi_i(1)^T \varphi_k(1)$, which leads to the condition

$$\begin{aligned} f_{ik}(t) &:= (q_i^T q_k)\left[\cos\left(\gamma_i + t\left[\frac{\pi}{2} - \gamma_i\right]\right)\cos\left(\gamma_k + t\left[\frac{\pi}{2} - \gamma_k\right]\right)\right] \\ &\quad + \sin\left(\gamma_i + t\left[\frac{\pi}{2} - \gamma_i\right]\right)\sin\left(\gamma_k + t\left[\frac{\pi}{2} - \gamma_k\right]\right) \\ &\ge (q_i^T q_k)[\cos\gamma_i \cos\gamma_k] + \sin\gamma_i \sin\gamma_k = f_{ik}(0). \end{aligned} \tag{9}$$

We prove that $f_{ik}(t)$ is nondecreasing in $t \in [0,1]$. In the case $q_i^T q_k < 0$, both cosine terms in $f_{ik}(t)$ are nonincreasing, and the sine terms are nondecreasing. In the remaining case we use the angle addition formulas to find

$$\begin{aligned} f_{ik}(t) &= q_i^T q_k \cos((1-t)[\gamma_i - \gamma_k]) \\ &\quad + (1 - q_i^T q_k)\sin\left(\gamma_i + t\left[\frac{\pi}{2} - \gamma_i\right]\right)\sin\left(\gamma_k + t\left[\frac{\pi}{2} - \gamma_k\right]\right). \end{aligned}$$

But $0 \le q_i^T q_k \le 1$, and so the cosine and sine terms are nondecreasing. ∎

Again, we continue with showing that during this transformation the length of the projected cumulated vector is nondecreasing.

OBSERVATION 18. *For* ψ_i $(i \in C_j)$ *as defined in Observation* 17, *define* $\bar{\psi}_j : [0,1] \to \mathbb{R}^n$ *by*

$$\bar{\psi}_j(t) := \sum_{i \in C_j} \psi_i(t).$$

The length $\|p_{\mathcal{L}^\perp}(\bar{\psi}_j(t))\|$ *is nondecreasing in* $t \in [0,1]$.

Proof. By using the functions f_{ik} introduced in (9) we may write

$$\begin{aligned} \left\| \sum_{i \in C_j} p_{\mathcal{L}^\perp}(\psi_i(t)) \right\|^2 &= \sum_{i \in C_j} \|p_{\mathcal{L}^\perp}(\psi_i(t))\|^2 + \sum_{i,k \in C_j, i<k} 2(p_{\mathcal{L}^\perp}(\psi_i(t)))^T (p_{\mathcal{L}^\perp}(\psi_k(t))) \\ &= \sum_{i \in C_j} \|p_{\mathcal{L}^\perp}(\psi_i(t))\|^2 + \sum_{i,k \in C_j, i<k} \delta_i \delta_k f_{ik}(t), \end{aligned}$$

and we have shown in the proof of Observation 17 that each $f_{ik}(t)$ is nondecreasing in $t \in [0,1]$. ∎

We concatenate both transformations into one and summarize our findings on this collapsing transformation.

OBSERVATION 19 (collapsing transformation). *Given $j \in M$ and $b_j \in \mathcal{H} \cap \mathcal{L}^\perp$, with $\|b_j\| = 1$, define $u_i : [0,1] \to \mathbb{R}^n$ for $i \in C_j$ by*

$$u_i(t) := \begin{cases} \varphi_i(2t) & \text{for } t \in [0, \frac{1}{2}], \\ \psi_i(2[t - \frac{1}{2}]) & \text{for } t \in (\frac{1}{2}, 1], \end{cases} \tag{10}$$

with φ_i and ψ_i as given in Observations 15 and 17. Then, for $i \in C_j$,

(i) $u_i(0) = v_i$,

(ii) $u_i(1) = p_{\mathcal{L}}(v_i) + \|p_{\mathcal{L}^\perp}(v_i)\| b_j \in \mathcal{L} + \{\beta b_j : \beta \geq 0\}$,

and for all $t \in [0,1]$ it holds that

(iii) $p_{\mathcal{L}}(v_i) = p_{\mathcal{L}}(u_i(t))$ *and* $\|p_{\mathcal{L}^\perp}(v_i)\| = \|p_{\mathcal{L}^\perp}(u_i(t))\|$,

(iv) $\|u_i(t) - v\| = \|v_i - v\|$ *for* $v \in \mathcal{L} \supseteq \{v_s : s \in S\}$,

(v) $\|u_i(t) - u_k(t)\| \leq \|v_i - v_k\|$ *for* $k \in C_j$.

Furthermore, for

$$\bar{u}_j(t) := \sum_{i \in C_j} u_i(t)$$

the length $\|p_{\mathcal{L}^\perp}(\bar{u}_j(t))\|$ is nondecreasing in $t \in [0,1]$ and

$$\|p_{\mathcal{L}^\perp}(\bar{u}_j(1))\| = \sum_{i \in C_j} \|p_{\mathcal{L}^\perp}(v_i)\|.$$

Proof. The result follows from Observations 15, 17, 16, and 18. ∎

Suppose that the lengths

$$\tilde{\delta}_j := \sum_{i \in C_j} \|p_{\mathcal{L}^\perp}(v_i)\| \qquad (j \in M)$$

of the collapsed sets satisfy the condition corresponding to (8). Then, in order to obtain an embedding that is also in equilibrium with respect to the subspace $\mathcal{L}^\perp$, we have to choose only the collapsing direction b_j of each component C_j according to the vectors d_k (embedded in $\mathcal{L}^\perp$) with the assignment κ of Observation 14, $b_j = d_{\kappa(j)}$. This will yield an optimal embedding of dimension at most $\dim \mathcal{L} + 2$ as described in the following lemma.

LEMMA 20. *Let $v_i \in \mathbb{R}^n$ for $i \in N$ be an optimal solution of (4) for a connected graph $G = (N, E)$, and let $S \subset N$, with $0 \in \mathcal{S} := \operatorname{conv}\{v_s : s \in S\}$, be a separator in G giving rise to separated sets $C_j \subset N$, $j \in M := \{1, \ldots, m\}$. Put $\mathcal{L} := \operatorname{span} \mathcal{S}$ and, for $j \in M$, $\tilde{\delta}_j := \sum_{i \in C_j} \|p_{\mathcal{L}^\perp}(v_i)\|$.*

If $\tilde{\delta}_{\hat{\jmath}} \leq \sum_{j \in M \setminus \{\hat{\jmath}\}} \tilde{\delta}_j$ for all $\hat{\jmath} \in M$, then there exist vectors $d_1, d_2, d_3 \in \mathcal{L}^\perp$, $\|d_1\| = \|d_2\| = \|d_3\| = 1$, with $\dim \operatorname{span}\{d_1, d_2, d_3\} \leq 2$, $b_j \in \{d_1, d_2, d_3\}$, $j \in M$, so that the embedding v'_i, $i \in N$, with

$$v'_i := \begin{cases} v_i & \text{for } \ i \in S, \\ p_{\mathcal{L}}(v_i) + \|p_{\mathcal{L}^\perp}(v_i)\| b_j & \text{for } \ i \in C_j, \end{cases}$$

is also an optimal embedding of (4). Furthermore, such an embedding exists with $b_j = d_1$ for at most one $j \in M$ and satisfies $\dim \operatorname{span}\{v'_i : i \in N\} \leq \dim \mathcal{L} + 2 \leq |S| + 1$.

Proof. Observe that $\dim \mathcal{L} \le |S| - 1$ because by $0 \in \mathcal{S}$ the v_i $(i \in S)$ are linearly dependent. Choose h, and define $\mathcal{H}$ as specified in Observation 9. If $\tilde{\delta}_j = 0$ for all $j \in M$, then the statement holds for $d_1 = d_2 = d_3 = h$ because $v'_i = v_i \in \mathcal{L}$ for $i \in N$. So we may assume that $\tilde{\delta}_j > 0$ for at least two $j \in M$. In the case $\dim(\mathcal{H} \cap \mathcal{L}^\perp) = 1$ we must have $|S| = n - 2$, $m = 2$, and $|C_1| = |C_2| = 1$, so $b_1 = d_1 = -b_2 = -d_2 = -d_3$, with $d_1 = p_{\mathcal{L}^\perp}(v_i)/\|p_{\mathcal{L}^\perp}(v_i)\|$, satisfies all requirements. It remains to consider the case $\dim(\mathcal{H} \cap \mathcal{L}^\perp) \ge 2$.

By Observation 14 we find three vectors $d_1, d_2, d_3 \in \mathcal{H} \cap \mathcal{L}^\perp$ of norm one and an assignment $\kappa : M \to \{1, 2, 3\}$ satisfying $\sum_{j \in M} \tilde{\delta}_j d_{\kappa(j)} = 0$ and $\{j \in M : \kappa(j) = 1\} = 1$. For $j \in M$ set $b_j = d_{\kappa(j)}$, and let $u_i(t)$, $i \in C_j$, be the transformations of Observation 19 for the respective b_j. Then $v'_i = u_i(1)$ for $i \in C_j$, $j \in M$ by Observation 19(ii). The distance constraints are satisfied for the new embedding because for $\{i, k\} \in E$

$$\begin{array}{ll} i, k \in S: & \|v'_i - v'_k\| = \|v_i - v_k\| \text{ by definition,} \\ i \in C_j \text{ for some } j \in M, k \in S: & \|v'_i - v'_k\| = \|v_i - v_k\| \text{ by Observation 19(iv),} \\ i, k \in C_j \text{ for some } j \in M: & \|v'_i - v'_k\| \le \|v_i - v_k\| \text{ by Observation 19(v).} \end{array}$$

The equilibrium constraint is satisfied on $\mathcal{L}$, because $p_{\mathcal{L}}(v_i) = p_{\mathcal{L}}(v'_i)$ for all $i \in N$ (by definition for $i \in S$ and by Observation 19(iii) otherwise). It is also satisfied on $\mathcal{L}^\perp$, because

$$\begin{aligned} \sum_{i \in N} p_{\mathcal{L}^\perp}(v'_i) &= \sum_{i \in S} \underbrace{p_{\mathcal{L}^\perp}(v'_i)}_{=0} + \sum_{j \in M} \sum_{i \in C_j} p_{\mathcal{L}^\perp}(v'_i) \\ &= \sum_{j \in M} \sum_{i \in C_j} \|p_{\mathcal{L}^\perp}(v_i)\| b_j = \sum_{j \in M} \tilde{\delta}_j d_{\kappa(j)} = 0 \end{aligned}$$

by construction of the d_j. Finally, the objective value has not changed because $\|v_i\| = \|v'_i\|$ for all $i \in N$ (by definition for $i \in S$ and by Observation 19(iii) otherwise). ∎

If one set $\hat{\jmath} \in M$ is "heavier" than the other sets, $\tilde{\delta}_{\hat{\jmath}} > \sum_{j \in M \setminus \{\hat{\jmath}\}} \tilde{\delta}_j$, the need to recover feasibility in the equilibrium constraint will not allow us to collapse $\hat{\jmath}$ in full. We can, however, collapse all other sets and compensate this by carrying through the transformation in $\hat{\jmath}$ up to the $t_{\hat{\jmath}} \in [0, 1]$ when $\|p_{\mathcal{L}^\perp}(\bar{u}_{\hat{\jmath}}(t_{\hat{\jmath}}))\| = \sum_{j \in M \setminus \{\hat{\jmath}\}} \tilde{\delta}_j$. Even though this may lead to a slight increase in the overall dimension if $t_{\hat{\jmath}} < \frac{1}{2}$, it will help later to reduce the number of components about which we have to worry.

LEMMA 21. *Given the setting of Lemma 20 assume that there is a $\hat{\jmath} \in M$ with $\tilde{\delta}_{\hat{\jmath}} > \sum_{j \in M \setminus \{\hat{\jmath}\}} \tilde{\delta}_j$. There exists an $h \in \operatorname{span}\{v_i : i \in N\}^\perp$ and an optimal embedding v'_i $(i \in N)$ of (4), with*

$$\begin{array}{ll} v'_i \in \operatorname{span}\{h, v_i : i \in C_{\hat{\jmath}}\} & \text{for } i \in C_{\hat{\jmath}}, \\ v'_i = v_i & \text{for } i \in S, \\ v'_i = p_{\mathcal{L}}(v_i) + \|p_{\mathcal{L}^\perp}(v_i)\| \bar{b} & \text{for } i \in C_j, \text{ with } j \in M \setminus \{\hat{\jmath}\}, \end{array}$$

where $\bar{b} := -\frac{p_{\mathcal{L}^\perp}(\bar{v}'_{\hat{\jmath}})}{\|p_{\mathcal{L}^\perp}(\bar{v}'_{\hat{\jmath}})\|}$ if $\bar{v}'_{\hat{\jmath}} = \sum_{i \in C_{\hat{\jmath}}} v'_i \notin \mathcal{L}$ and $\bar{b} := 0$ otherwise.

Furthermore, if there is some direction $\hat{b} \in \operatorname{span}\{v_i : i \in C_{\hat{\jmath}}\} \cap \mathcal{L}^\perp \setminus \{0\}$ with $\hat{b}^T v_i \ge 0$ for $i \in C_{\hat{\jmath}}$, then such an embedding exists with $v'_i \in \operatorname{span}\{v_i : i \in C_{\hat{\jmath}}\}$ for $i \in C_{\hat{\jmath}}$.

Proof. If $\sum_{j \in M \setminus \{\hat{\jmath}\}} \tilde{\delta}_j = 0$, then we may choose $h = \bar{b} = 0$ and not transform the embedding at all to obtain the result. Therefore assume that $\sum_{j \in M \setminus \{\hat{\jmath}\}} \tilde{\delta}_j \ne 0$.

Choose h, and define $\mathcal{H}$ as specified in Observation 9. Since $\tilde{\delta}_{\hat{\jmath}} > 0$, we can find a $b_{\hat{\jmath}} \in \mathcal{L}^{\perp} \cap \operatorname{span}\{v_i : i \in C_{\hat{\jmath}}\}$ with $\|b_{\hat{\jmath}}\| = 1$. Let $u_i(t)$ $(i \in C_{\hat{\jmath}})$ denote the transformations of Observation 19 for this $b_{\hat{\jmath}}$, and set $\bar{u}_{\hat{\jmath}}(t) := \sum_{i \in C_{\hat{\jmath}}} u_i(t)$. By Observation 19, the function $\|p_{\mathcal{L}^{\perp}}(\bar{u}_{\hat{\jmath}}(t))\|$ is continuous and nondecreasing. As the equilibrium constraint is satisfied for the v_i $(i \in N)$, we have

$$\begin{aligned}
\|p_{\mathcal{L}^{\perp}}(\bar{u}_{\hat{\jmath}}(0))\| &\overset{\text{Observation19(i)}}{=} \left\| \sum_{i \in C_{\hat{\jmath}}} p_{\mathcal{L}^{\perp}}(v_i) \right\| \overset{\text{equilib.}}{=} \left\| \sum_{i \in N \setminus C_{\hat{\jmath}}} p_{\mathcal{L}^{\perp}}(v_i) \right\| \\
&\overset{p_{\mathcal{L}^{\perp}}(v_i)=0 \ (i \in S)}{=} \left\| \sum_{j \in M \setminus \{\hat{\jmath}\}} \sum_{i \in C_j} p_{\mathcal{L}^{\perp}}(v_i) \right\| \\
&\leq \sum_{j \in M \setminus \{\hat{\jmath}\}} \left\| \sum_{i \in C_j} p_{\mathcal{L}^{\perp}}(v_i) \right\| \overset{\text{(by def.)}}{=} \sum_{j \in M \setminus \{\hat{\jmath}\}} \tilde{\delta}_j
\end{aligned}$$

and, by assumption, $\|p_{\mathcal{L}^{\perp}}(\bar{u}_{\hat{\jmath}}(1))\| = \tilde{\delta}_{\hat{\jmath}} > \sum_{j \in M \setminus \{\hat{\jmath}\}} \tilde{\delta}_j$. So there is a $t_{\hat{\jmath}} \in [0,1]$ with

$$\|p_{\mathcal{L}^{\perp}}(\bar{u}_{\hat{\jmath}}(t_{\hat{\jmath}}))\| = \sum_{j \in M \setminus \{\hat{\jmath}\}} \tilde{\delta}_j. \tag{11}$$

Choose $v'_i = u_i(t_{\hat{\jmath}})$ for $i \in C_{\hat{\jmath}}$, and put

$$\bar{v}'_{\hat{\jmath}} = \sum_{i \in C_{\hat{\jmath}}} v'_i = \bar{u}_{\hat{\jmath}}(t_{\hat{\jmath}}) \quad \text{and} \quad \bar{b} = -p_{\mathcal{L}^{\perp}}(\bar{v}'_{\hat{\jmath}}) / \|p_{\mathcal{L}^{\perp}}(\bar{v}'_{\hat{\jmath}})\|. \tag{12}$$

For $j \in M \setminus \{\hat{\jmath}\}$ choose $b_j = \bar{b}$, and let $u_i(t)$, $i \in C_j$, be the transformations of Observation 19 for the respective b_j. Then, by Observation 19(ii), $v'_i = u_i(1)$ for $i \in C_j$, with $j \in M \setminus \{\hat{\jmath}\}$, satisfies the requirements of the lemma. The equilibrium constraint is satisfied for the embedding v'_i, $i \in N$, because it holds on $\mathcal{L}$ due to $p_{\mathcal{L}}(v_i) = p_{\mathcal{L}}(v'_i)$ for $i \in N$ by Observation 19(iii) and it holds on $\mathcal{L}^{\perp}$, because

$$\begin{aligned}
\sum_{i \in N} p_{\mathcal{L}^{\perp}}(v'_i) &= \sum_{i \in S} \underbrace{p_{\mathcal{L}^{\perp}}(v'_i)}_{=0} + \sum_{i \in C_{\hat{\jmath}}} p_{\mathcal{L}^{\perp}}(v'_i) + \sum_{j \in M \setminus \{\hat{\jmath}\}} \sum_{i \in C_j} p_{\mathcal{L}^{\perp}}(v'_i) \\
&\overset{\text{(by def.)}}{=} p_{\mathcal{L}^{\perp}}(\bar{v}'_{\hat{\jmath}}) + \sum_{j \in M \setminus \{\hat{\jmath}\}} \sum_{i \in C_j} \|p_{\mathcal{L}^{\perp}}(v_i)\| \bar{b} \\
&\overset{(12)}{=} \left(\|p_{\mathcal{L}^{\perp}}(\bar{v}'_{\hat{\jmath}})\| - \sum_{j \in M \setminus \{\hat{\jmath}\}} \tilde{\delta}_j \right) \frac{p_{\mathcal{L}^{\perp}}(\bar{v}'_{\hat{\jmath}})}{\|p_{\mathcal{L}^{\perp}}(\bar{v}'_{\hat{\jmath}})\|} \overset{(11)}{=} 0.
\end{aligned}$$

The feasibility of v'_i, $i \in N$, with respect to the distance constraints and optimality follows from Observation 19(iv) and (v) as in the proof of Lemma 20.

Finally, suppose that $\hat{b}$ exists as described in the statement of the lemma. Then we may choose $b_{\hat{\jmath}} = \frac{\hat{b}}{\|\hat{b}\|}$ and, by construction (10) of the u_i $(i \in C_{\hat{\jmath}})$, $u_i(t) = v_i$ for $t \in [0, \frac{1}{2}]$ (see Observation 15 for φ_i) and $u_i(t) \in \mathcal{L} + \operatorname{span}\{\hat{b}, v_i\}$ for $t \in [\frac{1}{2}, 1]$ (see Observation 17 for ψ_i). This completes the proof. □

Remark 22. A solution corresponding to the modified solution of this lemma is not necessarily an optimal embedding of minimal dimension. Consider, e.g., the graph

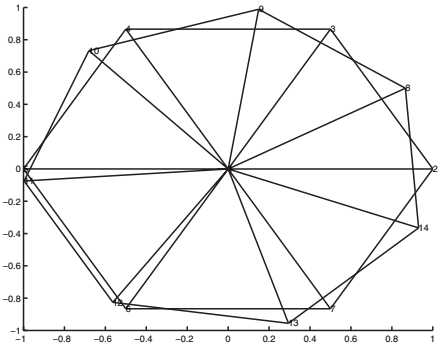


FIG. 11. *Optimal two-dimensional embedding of two wheels with identical hub; see Remark* 22. *The construction of the proof of Lemma* 21 *would yield a three-dimensional embedding.*

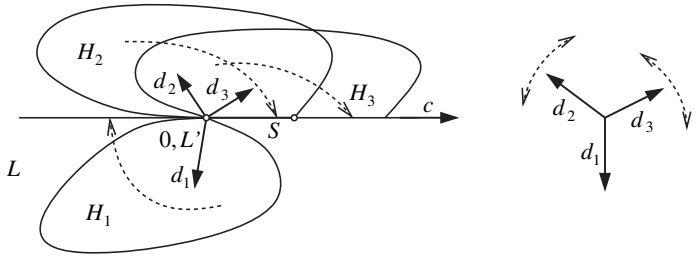


FIG. 12. *Transformation in the proof of Lemma* 25.

consisting of two wheels with identical hub, and rims of k and $k+1$ nodes with $k \geq 6$; see Figure 11.

Lemma 21 does not provide a bound on the dimension of the embedding but will tell us which component we have to think about next in order to get to such a bound. In order to arrive at the result of Theorem 5, we will need a further refinement of Lemma 20 in the case $\dim \mathcal{L} = |S| - 1$. In order to set up the scene, assume that the v_i $(i \in N)$ are already embedded in dimension $|S|+1$ as described in Lemma 20 with each node set C_j collapsed to some flat half-space $\mathcal{L} + \{\delta d_{\kappa(j)} : \delta \geq 0\}$, and denote by H_1 the set C_j that is the only one assigned to direction d_1. We are interested in the case that H_1 is not connected to some $\hat{s} \in S$, so $S' := S \backslash \{\hat{s}\}$ is a separator for H_1 in G. By Theorem 3 we must have $0 \in \mathcal{S}' := \operatorname{conv}\{v_i : i \in S'\}$, with $\mathcal{L}' := \operatorname{span} \mathcal{S}'$ a linear subspace of dimension $\dim(\mathcal{L}') = \dim(\mathcal{L}) - 1$ and $v_{\hat{s}} \neq 0$ (otherwise the dimension of the embedding would be $|S|$ already). Figure 12 depicts the situation when projected onto $\mathcal{L}'^{\perp}$ with $H_i := \bigcup_{j \in M:\kappa(j)=i} C_j$ the set of nodes which are embedded in direction d_i, $i = 2, 3$. It will turn out that the transformation indicated in this illustration will yield an optimal embedding of dimension at most $|S|$, so we can get rid of one more dimension. For this purpose we introduce yet another transformation comparable to closing a fan; see Figure 13. In the following observation think of vector g as spanning the missing direction $\pm p_{\mathcal{L}'^{\perp}}(v_{\hat{s}}) / \|p_{\mathcal{L}'^{\perp}}(v_{\hat{s}})\|$ in $\mathcal{L}$ and d as the additional direction d_j that spans the embedding of node set H_j.

OBSERVATION 23. *Given a linear subspace* $\mathcal{L}' \subset \mathbb{R}^n$, *vectors* $d, g \in \mathcal{L}'^{\perp}$, *with* $\|d\| = \|g\| = 1$, $d^T g = 0$, *and* $v_i \in \{x \in \mathcal{L}' + \operatorname{span}\{d, g\} : d^T x \geq 0\}$ *(*$i \in C \subseteq N$*). For* $i \in C$, *set* $\delta_i' := \|p_{\mathcal{L}'^{\perp}}(v_i)\|$, *determine* $\gamma_i \in [0, \pi]$ *so that* $p_{\mathcal{L}'^{\perp}}(v_i) = \delta_i'(g \cos \gamma_i +$

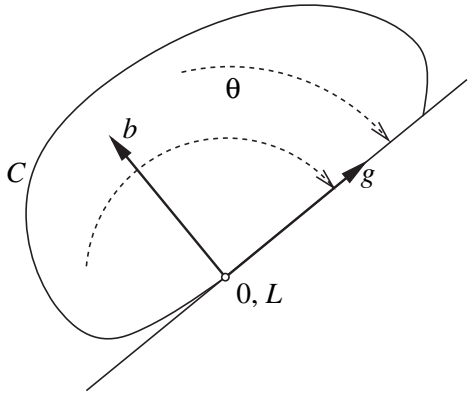


FIG. 13. *θ squeezes C spanned by $\mathcal{L}' + \operatorname{span}\{g\}$ and nonnegative d into the boundary half-space spanned by $\mathcal{L}$ and nonnegative g (Observation 23).*

$d \sin\gamma_i$), and define continuous maps $\theta_i : [0,1] \to \mathbb{R}^n$:

$$\theta_i(t) := p_{\mathcal{L}'}(v_i) + \delta_i' \left[g\cos(\gamma_i - t\gamma_i) + d\sin(\gamma_i - t\gamma_i) \right].$$

Then, for $i \in C$,

(i) $\theta_i(0) = v_i$,

(ii) $\theta_i(1) = p_{\mathcal{L}'}(v_i) + \|p_{\mathcal{L}'^{\perp}}(v_i)\| g$,

and for all $t \in [0,1]$ it holds that

(iii) $p_{\mathcal{L}'}(v_i) = p_{\mathcal{L}'}(\theta_i(t))$ *and* $\|p_{\mathcal{L}'^{\perp}}(v_i)\| = \|p_{\mathcal{L}'^{\perp}}(\theta_i(t))\|$,

(iv) $\|\theta_i(t) - v\| \le \|v_i - v\|$ *for* $v \in \mathcal{L}' + \{\beta g : \beta > 0\}$,

(v) $\|\theta_i(t) - \theta_k(t)\| \le \|v_i - v_k\|$ *for* $k \in C$.

Furthermore, for $\bar{\theta}_C(t) := \sum_{i \in C} \theta_i(t)$,

(vi) $p_{\mathcal{L}'^{\perp}}(\bar{\theta}_C(t)) \in \operatorname{span}\{g\} + \{\beta d : \beta \ge 0\}$ *for* $t \in [0,1]$,

(vii) $g^T \bar{\theta}_C(t)$ *is strictly increasing in* $t \in [0,1]$ *if* $\gamma_i \in (0,\pi]$ *and* $\delta_i' > 0$ *for some* $i \in C$,

(viii) $\bar{\theta}_C(1) = \sum_{i \in C} p_{\mathcal{L}'}(v_i) + g \sum_{i \in C} \delta_i'$.

Proof. (i)–(iii) follow from direct calculation and by exploiting the fact that $g, d \in \mathcal{L}'^{\perp}$ are orthonormal vectors. In order to prove (iv), apply the same arguments used in the proofs of Observation 10(ii) and (iii) to $v \in \mathcal{L}'$ and to $v \in \mathcal{L}' + \{\beta g : \beta \ge 0\}$, respectively.

For proving (v), i.e., $\|\theta_i(t) - \theta_k(t)\|^2 \le \|v_i - v_k\|^2$ for $i, k \in C$, it suffices to show that

$$f_{ik}(t) := \theta_i(t)^T \theta_k(t) \ge v_i^T v_k \overset{\text{(i)}}{=} \theta_i(0)^T \theta_k(0) = f_{ik}(0)$$

or, as $g, d \in \mathcal{L}'^{\perp}$ are orthonormal vectors, that the function

$$f_{ik}(t) = p_{\mathcal{L}'}(v_i)^T p_{\mathcal{L}'}(v_k) + \delta_i' \delta_k' [\cos(\gamma_i - t\gamma_i)\cos(\gamma_k - t\gamma_k) + \sin(\gamma_i - t\gamma_i)\sin(\gamma_k - t\gamma_k)]$$

is nondecreasing in $t \in [0,1]$. By the angle addition formulas and since the cosine is an even function,

$$\cos(\gamma_i - t\gamma_i)\cos(\gamma_k - t\gamma_k) + \sin(\gamma_i - t\gamma_i)\sin(\gamma_k - t\gamma_k) = \cos((1-t)|\gamma_i - \gamma_k|).$$

The right-hand side is nondecreasing, and, thus, f_{ik} is nondecreasing.

(vi) and (viii) follow from direct computation, and for (vii) it suffices to observe that

$$g^T \bar{\theta}_C(t) = \sum_{i \in C} \delta'_i \cos(\gamma_i - t\gamma_i)$$

is strictly increasing because $\delta'_i \geq 0$ for all $i \in C$ and $\cos(\gamma_i - t\gamma_i)$ is strictly increasing in $t \in [0,1]$ whenever $\gamma_i \in (0, \pi]$. ∎

The next observation will serve to find the correct balancing of the parameters for each H_j in order to guarantee the equilibrium constraint on the subspace spanned by g and appropriately chosen d_j.

OBSERVATION 24. *Given continuous functions $\lambda_j : [0,1] \to \mathbb{R}^2$ ($j \in \{1,2,3\}$) and $\sigma \in \mathbb{R}$, with*

(i) $[\lambda_1(t)]_1$ *is strictly decreasing,* $[\lambda_2(t)]_1$ *and* $[\lambda_3(t)]_1$ *are strictly increasing,*
(ii) $[\lambda_i(t)]_2 \geq 0$ *for* $t \in [0,1]$ *and* $i = 1,2,3$,
(iii) $[\lambda_1(0)]_1 + [\lambda_2(0)]_1 + [\lambda_3(0)]_1 + \sigma = 0$,
(iv) $[\lambda_i(0)]_2 < [\lambda_j(0)]_2 + [\lambda_k(0)]_2$ *for pairwise distinct* $i,j,k \in \{1,2,3\}$,
(v) $[\lambda_1(1)]_2 = [\lambda_2(1)]_2 = [\lambda_3(1)]_2 = 0$,

there exist $t_1, t_2, t_3 \in [0,1]$ *and pairwise distinct* $\hat{\imath}, \hat{\jmath}, \hat{k} \in \{1,2,3\}$ *satisfying*

(vi) $[\lambda_1(t_1)]_1 + [\lambda_2(t_2)]_1 + [\lambda_3(t_3)]_1 + \sigma = 0$,
(vii) $[\lambda_{\hat{\imath}}(t_{\hat{\imath}})]_2 = [\lambda_{\hat{\jmath}}(t_{\hat{\jmath}})]_2 + [\lambda_{\hat{k}}(t_{\hat{k}})]_2$.

Proof. Due to continuity, the monotonicity property (i), and the initial condition (iii), there exists a continuous nondecreasing function $\tau : [0, \bar{\tau}] \to [0,1]$ defined implicitly via

$$[\lambda_1(t)]_1 + [\lambda_2(\tau(t))]_1 + [\lambda_3(\tau(t))]_1 + \sigma = 0,$$

where

$$\bar{\tau} := \max\{t \in [0,1] : [\lambda_1(t)]_1 + [\lambda_2(t')]_1 + [\lambda_3(t')]_1 + \sigma = 0 \text{ for some } t' \in [0,1]\}.$$

By definition, $(t'_1, t'_2, t'_3) = (\bar{\tau}, \tau(\bar{\tau}), \tau(\bar{\tau}))$ satisfies (vi), and by monotonicity at least one of t'_1, t'_2, t'_3 is equal to one. Then (v) and (ii) imply that there are pairwise distinct $i,j,k \in \{1,2,3\}$ with $[\lambda_i(t_i)]_2 \geq [\lambda_j(t_j)]_2 + [\lambda_k(t_k)]_2$. By the initial condition (iv) and the continuity of the λ_j and τ, there must be a smallest $t_1 \in (0,1]$ so that $t_2 = t_3 = \tau(t_1)$ satisfy (vi) and (vii). ∎

LEMMA 25. *Given the setting of Lemma 20, assume that $\tilde{\delta}_{\hat{\jmath}} \leq \sum_{j \in M \setminus \{\hat{\jmath}\}} \tilde{\delta}_j$ holds for all $\hat{\jmath} \in M$, and let $\bar{\jmath} \in M$ be the only index with $b_{\bar{\jmath}} = d_1$ within the new embedding of Lemma 20. If at most $|S|-1$ nodes of S are adjacent to nodes in $C_{\bar{\jmath}}$, then there is an optimal embedding of dimension at most $|S|$.*

Proof. Let v_i, $i \in N$, be the optimal embedding resulting from Lemma 20 with normalized vectors $d_1, d_2, d_3 \in \mathcal{L}^\perp$ satisfying $\dim \operatorname{span}\{d_1, d_2, d_3\} \leq 2$ and an assignment $\kappa : M \to \{1,2,3\}$ with $b_j = d_{\kappa(j)}$ for $j \in M$. Choose $H_k := \bigcup_{j \in M : \kappa(j) = k} C_j$ for $k \in \{1,2,3\}$. Then

$$v_i \in \mathcal{L} + \{\beta d_j : \beta \geq 0\} \quad \text{for } i \in H_j,\ j \in \{1,2,3\}. \tag{13}$$

Together with $\mathcal{L} = \operatorname{span} \mathcal{S}$ and $0 \in \mathcal{S} = \operatorname{conv}\{v_s : s \in S\}$, the dimension of this embedding is bounded by $\dim \mathcal{L} + \dim \operatorname{span}\{d_1, d_2, d_3\}$ and $\dim \mathcal{L} \leq |S| - 1$. If $\dim \mathcal{L} < |S| - 1$ or $\dim \operatorname{span}\{d_1, d_2, d_3\} < 2$, then the statement holds, so we may assume that $\dim \mathcal{L} = |S| - 1$ and $\dim \operatorname{span}\{d_1, d_2, d_3\} = 2$. Next suppose that there

is a $j \in \{1,2,3\}$ with $v_i^T d_j = 0$ for all $i \in H_j$; w.l.o.g. assume this to hold for $j = 1$. Then the equilibrium constraint on $\mathcal{L}^\perp$ simplifies to $\sum_{i\in H_2} \|p_{\mathcal{L}^\perp}(v_i)\| d_2 = \sum_{i\in H_3} \|p_{\mathcal{L}^\perp}(v_i)\| d_3$. Thus, the embedding on $\mathcal{L}^\perp$ is restricted to a one-dimensional subspace, and the dimension of the embedding is again bounded by $|S|$. So it remains to consider the case

$$\text{for each } j \in \{1,2,3\}, \quad v_i^T d_j > 0 \text{ for some } i \in H_j. \tag{14}$$

By assumption there is a node $\hat{s} \in S$ not adjacent to any node in $H_1 = C_{\bar{\jmath}}$. Put $S' := S \setminus \{\hat{s}\}$. This set S' separates H_1 from G. We have $0 \in \mathcal{S}' := \operatorname{conv}\{v_s : s \in S'\}$, because otherwise the separator-shadow Theorem 3 would imply that $v_i \in \mathcal{L}' := \operatorname{span} \mathcal{S}'$ for $i \in H_1$, in contradiction to (14). Now $0 \in \mathcal{S}'$ yields $\dim \mathcal{L}' = |S'| - 1$, and as $\dim \mathcal{L} = |S| - 1$ we find a vector $\hat{g}$ with

$$0 \neq \hat{g} = \frac{p_{\mathcal{L}'^\perp}(v_{\hat{s}})}{\|p_{\mathcal{L}'^\perp}(v_{\hat{s}})\|} \in \mathcal{L} \cap \mathcal{L}'^\perp \quad \text{and} \quad \hat{g}^T v_s = 0 \text{ for } s \in S'. \tag{15}$$

Set $g_1 := -\hat{g}$ and $g_2 := g_3 := \hat{g}$, and then by (13)

$$\text{for each } j \in \{1,2,3\}, \quad v_i \in \{x \in \mathcal{L}' + \operatorname{span}\{d_j, g_j\} : d_j^T x \geq 0\} \text{ for all } i \in H_j.$$

Therefore we may use Observation 23 for $j \in \{1,2,3\}$ with $C = H_j$, $d = d_j$, $g = g_j$ to define transformations $\theta_i(t)$ for $i \in H_j$ and $\bar{\theta}_j(t) = \bar{\theta}_{H_j}(t)$. Observe that $\mathcal{S}' \subset \mathcal{L}'$ and $\mathcal{S} \subset \mathcal{L}' + \{\beta g_j : \beta \geq 0\}$ for $j \in \{2,3\}$, so Observation 23(iv) and (v) establish that for $j \in \{1,2,3\}$ and $t_j \in [0,1]$ the distance constraints of edges incident to nodes $i \in H_j$ remain satisfied for embedding $\theta_i(t_j)$, and the objective value remains unchanged due to Observation 23(iii) by $0 \in \mathcal{L}'$. Also note that replacing d_j by some other normalized $d_j' \in \mathcal{L}^\perp$ will not affect distance constraints but only the equilibrium constraint. So it remains to find appropriate $t_j \in [0,1]$ and normalized $d_j' \in \mathcal{L}^\perp$ so that the equilibrium constraint holds while the dimension of the embedding is reduced by at least one. For this purpose, define for $j \in \{1,2,3\}$ the function $\lambda_j : [0,1] \to \mathbb{R}^2$ by

$$\lambda_j(t) := \begin{pmatrix} \hat{g}^T \bar{\theta}_j(t) \\ d_j^T \bar{\theta}_j(t) \end{pmatrix} \qquad \text{for } t \in [0,1].$$

We show that the λ_j and $\sigma := \hat{g}^T v_{\hat{s}}$ satisfy the requirements of Observation 24. Observation 24(i) holds because of Observation 23(vii) and (14). Observation 24(ii) follows from Observation 23(vi). Observation 24(iii) is implied by the feasibility of equilibrium constraint on the linear subspace spanned by $\hat{g}$ for the embedding v_i, $i \in N$; for this, use Observation 23(i), (15), and the definition of σ. Suppose that Observation 24(iv) does not hold, and assume, w.l.o.g., that $\lambda_1(0) \geq \lambda_2(0) + \lambda_3(0)$; then by (13) and Observation 23(i) this is equivalent to

$$\sum_{i\in H_1} \|p_{\mathcal{L}^\perp}(v_i)\| \geq \sum_{i\in H_2 \cup H_3} \|p_{\mathcal{L}^\perp}(v_i)\|,$$

and together with the equilibrium constraint

$$\sum_{i\in H_1} \|p_{\mathcal{L}^\perp}(v_i)\| d_1 + \sum_{i\in H_2} \|p_{\mathcal{L}^\perp}(v_i)\| d_2 + \sum_{i\in H_2} \|p_{\mathcal{L}^\perp}(v_i)\| d_3 = 0$$

this implies that $d_1 = -d_2 = -d_3$ in contradiction to $\dim \operatorname{span}\{d_1, d_2, d_3\} = 2$. Thus, Observation 24(iv) holds. Finally, Observation 24(v) follows from Observation 23(viii). Hence, there exist $t_1, t_2, t_3 \in [0,1]$ and pairwise distinct $\hat{\imath}, \hat{\jmath}, \hat{k} \in \{1,2,3\}$ so that Observation 24(vi) and (vii) hold. Now

$$\text{choose } \hat{d} \in \mathcal{L}^\perp, \|\hat{d}\| = 1, \quad \text{set } d'_{\hat{\imath}} := -d'_{\hat{\jmath}} := -d'_{\hat{k}} := \hat{d}, \tag{16}$$

and

$$v'_i := \begin{cases} v_i & i \in S, \\ p_{\mathcal{L}'}(v_i) + \delta_i[g_j \cos(\gamma_i - t_j\gamma_i) + d'_j \sin(\gamma_i - t_j\gamma_i)] & i \in H_j, j \in \{1,2,3\}. \end{cases}$$

Since only the d_j have been replaced by d'_j, $j \in \{1,2,3\}$, the distance constraints are still valid for the new embedding v'_i, $i \in N$, and the objective value is unchanged. Furthermore, by setting

$$\bar{\theta}'_j(t) := \sum_{i \in H_j} p_{\mathcal{L}'}(v_i) + \delta_i[g_j \cos(\gamma_i - t\gamma_i) + d'_j \sin(\gamma_i - t\gamma_i)] \quad \text{for } j \in \{1,2,3\},$$

we see that the functions λ_j, $j \in \{1,2,3\}$, also satisfy

$$\lambda_j(t) = \begin{pmatrix} \hat{g}^T \bar{\theta}'_j(t) \\ d'^T_j \bar{\theta}'_j(t) \end{pmatrix} \qquad \text{for } t \in [0,1].$$

Therefore Observation 24(vi) and (vii) still hold for t_1, t_2, t_3 and $\hat{\imath}, \hat{\jmath}, \hat{k}$ yielding

$$0 = \sigma + \hat{g}^T(\bar{\theta}'_1(t_1) + \bar{\theta}'_2(t_2) + \bar{\theta}'_2(t_2)) = \hat{g}^T \left(v_{\hat{s}} + \sum_{j \in \{1,2,3\}} \sum_{i \in H_j} v'_i \right) \overset{(15)}{=} \hat{g}^T \left(\sum_{i \in N} v'_i \right),$$

$$0 = d'^T_{\hat{\imath}} \bar{\theta}'_{\hat{\imath}}(t_{\hat{\imath}}) - d'^T_{\hat{\jmath}} \bar{\theta}'_{\hat{\jmath}}(t_{\hat{\jmath}}) - d'^T_{\hat{k}} \bar{\theta}'_{\hat{k}}(t_{\hat{k}}) \overset{(16)}{=} \hat{d}^T \sum_{j \in \{1,2,3\}} \sum_{i \in H_j} v'_i \overset{\hat{d} \in \mathcal{L}^\perp}{=} \hat{d}^T \left(\sum_{i \in N} v'_i \right).$$

So the equilibrium constraint holds on the linear subspaces spanned by $\hat{g}$ and $\hat{d}$. It also holds on $\mathcal{L}'$ because $p_{\mathcal{L}'}(v_i) = p_{\mathcal{L}'}(v'_i)$ for $i \in N$ and the embedding v_i was feasible. Since $v'_i \in \mathcal{L}' + \operatorname{span}\{\hat{g}, \hat{d}\} = \mathcal{L} + \operatorname{span}\{\hat{d}\}$ for $i \in N$, the new embedding satisfies the equilibrium constraint on the entire space. Therefore it is an optimal embedding of dimension at most $\dim \mathcal{L} + 1 = |S|$. ∎

5. The proof of the tree-width Theorem 5. We will show that for any tree decomposition $T = (\mathcal{N}, \mathcal{E})$ of G (see Definition 4) there is always an optimal embedding of dimension at most $\max\{|U| : U \in \mathcal{N}\}$. As this also holds for a tree decomposition giving the tree width of G, this will prove the theorem.

Note that in a tree decomposition any $U \in \mathcal{N}$ and any $U \cap U'$ with $\{U, U'\} \in \mathcal{E}$ is a separator of G (see, e.g., Lemma 12.3.1 in [5]). In the proof we will show that by successively transforming the optimal embedding v_i, $i \in N$, we can find a separator of the form $U \in \mathcal{N}$ or $U \cap U'$ for some $\{U, U'\} \in \mathcal{E}$ containing 0 in the convex hull of its points so that either Lemma 20 or Lemma 25 yields an optimal embedding of appropriate dimension.

The first step asserts that for any optimal embedding any tree decomposition has "zero-nodes" containing the origin in their convex hull.

LEMMA 26. *Consider a tree decomposition $T = (\mathcal{N}, \mathcal{E})$ of a connected graph $G = (N, E)$ and an optimal embedding $v_i \in \mathbb{R}^n$ $(i \in N)$ of* (4). *There is a $U \in \mathcal{N}$ with $0 \in \operatorname{conv}\{v_u : u \in U\}$.*

Proof. Consider a subtree $(\mathcal{N}', \mathcal{E}') =: T'$ of T, with $|\mathcal{N}'|$ minimal so that $0 \in \operatorname{conv}\{v_i : i \in \bigcup_{U \in \mathcal{N}'} U\}$. Such a tree exists since the condition holds for $T' = T$ by the equilibrium constraint. Let the convex combination giving the origin be described by $C := \bigcup_{U \in \mathcal{N}'} U$ and $\alpha \in \mathbb{R}_+^C$, with $\alpha^T e = 1$, so that $\sum_{i \in C} \alpha_i v_i = 0$. If $|\mathcal{N}'| = 1$, we are done.

Assume, for contradiction, that $|\mathcal{N}'| > 1$. Then there is an edge $\{U, U'\} \in \mathcal{E}'$, and $S' := U \cap U'$ is a separator of G. Deleting edge $\{U, U'\}$ from T' splits T' into two nonempty subtrees $(\mathcal{N}'_j, \mathcal{E}'_j) =: T'_j$ for $j \in \{1, 2\}$ with $0 \notin \operatorname{conv}\{v_i : i \in \bigcup_{U \in \mathcal{N}'_j} U\}$ by assumption. Set $N'_j := \bigcup_{U \in \mathcal{N}'_j} U$. Because $S' \subseteq N'_j$ for $j \in \{1, 2\}$ we obtain $0 \notin \mathcal{S}' := \operatorname{conv}\{v_i : i \in S'\}$. Applying the separator-shadow Theorem 3 with respect to $S = S'$ and $C_j = N'_j \setminus S'$ $(j \in \{1, 2\})$ yields, w.l.o.g., $\operatorname{conv}\{v_i, 0\} \cap \mathcal{S}' \neq \emptyset$ for all $i \in C_1 = N'_1 \setminus S'$. But then the origin must be contained in the convex hull of subtree T'_2 as we show next. Put $C'_1 := N'_1 \setminus S'$, $C'_2 := N'_2$ and set, for $j \in \{1, 2\}$, $\bar{\alpha}_j := \sum_{i \in C'_j} \alpha_i$ and $\bar{v}_j := \frac{1}{\bar{\alpha}_j} \sum_{i \in C'_j} \alpha_i v_i \in \operatorname{conv}\{v_i : i \in N'_j\}$. Then $0 = \bar{\alpha}_1 \bar{v}_1 + \bar{\alpha}_2 \bar{v}_2 \in \operatorname{conv}\{\bar{v}_1, \bar{v}_2\}$ (by definition of the α_i) and $\emptyset \neq \mathcal{S}' \cap \operatorname{conv}\{\bar{v}_1, 0\} \subset \operatorname{conv}\{\bar{v}_1, \bar{v}_2\}$ (as the separator-shadow property holds for C'_1), so there is a $p \in \mathcal{S}' \subset \operatorname{conv}\{v_i : i \in N'_2\}$ with $0 \in \operatorname{conv}\{p, \bar{v}_2\} \subset \operatorname{conv}\{v_i : i \in N'_2\}$, a contradiction to the minimality of $|\mathcal{N}'|$. Hence, T' consists of only one node. ∎

We will call a node $U \in \mathcal{N}$ a *zero-node* (with respect to the embedding v_i, $i \in N$) if $0 \in \operatorname{conv}\{v_i : i \in U\}$ and an edge $\{U, U'\} \in \mathcal{E}$ a *zero-edge* (with respect to the embedding v_i, $i \in N$) if $0 \in \operatorname{conv}\{v_i : i \in U \cap U'\}$. Note that for a zero-edge both end points are zero-nodes.

OBSERVATION 27. *The subgraph $(\mathcal{N}', \mathcal{E}') =: T'$ of $T = (\mathcal{N}, \mathcal{E})$ induced by the zero-nodes of an optimal embedding v_i $(i \in N)$ of* (4) *is a tree, and $\mathcal{E}'$ is the set of zero-edges.*

Proof. Suppose that there are two zero-nodes U_1 and U_2 that are not connected in T' or that are connected in T' by an edge that is not a zero-edge. In both cases there is an edge $\{\bar{U}, \bar{U}'\} \in \mathcal{E}$ inducing a separator $S := \bar{U} \cap \bar{U}'$ in G with $0 \notin \mathcal{S} := \operatorname{conv}\{v_i : i \in S\}$ on the path connecting U_1 and U_2 in T. Deleting edge $\{\bar{U}, \bar{U}'\}$ from T splits T into two nonempty subtrees $(\mathcal{N}_j, \mathcal{E}_j) =: T_j$, with $U_j \in \mathcal{N}_j$ for $j \in \{1, 2\}$, so that the node sets $C_j := \bigcup_{U \in \mathcal{N}_j} U \setminus S$ have no common incident edge in G. For S, C_1, and C_2 the separator-shadow Theorem 3 implies, w.l.o.g., that $\operatorname{conv}\{v_i, 0\} \cap \mathcal{S} \neq \emptyset$ for all $i \in C_1$. Because $U_1 \subseteq C_1 \cup S$ we obtain $0 \notin \operatorname{conv}\{v_i : i \in U_1\}$, which contradicts the assumption that U_1 is a zero-node. ∎

Hence, for a given tree decomposition any optimal embedding induces a *zero-tree* (with respect to the embedding v_i, $i \in N$) consisting of the zero-nodes and zero-edges.

The algorithmic idea is to pick a zero-node U, transform the embedding for $S = U$ as suggested in Lemmas 20, 21, and 25, and check whether the resulting dimension is at most $|U|$. If it is not, it will turn out that, in the zero-tree of the new optimal embedding, U has a unique incident zero-edge $\{U, U'\}$ leading to that part of the graph whose embedding cannot yet be flattened out sufficiently with respect to U. We then go on transforming the new optimal embedding with respect to the separator $U \cap U'$ which may again lead to a sufficiently flat optimal embedding or, in failing to find one, lead on to U' via the part that is not flat enough. Now at some point this algorithm might turn back in U' and try to cross this last edge a second time.

Happily, this will immediately allow one to produce an optimal embedding that is sufficiently flat. As going on in one direction will be possible only for a finite number of times, this will complete the proof.

We start with the convenient case, where all parts can be flattened out sufficiently.

LEMMA 28. *Consider a tree decomposition $T = (\mathcal{N}, \mathcal{E})$ of a connected graph $G = (N, E)$, an optimal embedding $v_i \in \mathbb{R}^n$ $(i \in N)$ of (4), and a zero-node $S \in \mathcal{N}$ whose deletion splits T into m subtrees $(\mathcal{N}_j, \mathcal{E}_j) =: T_j$ $(j \in M := \{1, \dots, m\})$. Put*

$$\begin{aligned} \mathcal{L} &:= \operatorname{span}\{v_s : s \in S\}, \\ C_j &:= \bigcup_{U \in \mathcal{N}_j} U \setminus S, \\ \tilde{\delta}_j &:= \sum_{i \in C_j} \|p_{\mathcal{L}^\perp}(v_i)\| \quad (j \in M). \end{aligned}$$

If $\tilde{\delta}_{\hat{\jmath}} \le \sum_{j \in M \setminus \{\hat{\jmath}\}} \tilde{\delta}_j$ for all $\hat{\jmath} \in M$, then there is an optimal embedding v'_i $(i \in N)$ of dimension at most $|U'|$ for some $U' \in \{S, U : \{S, U\} \in \mathcal{E}\}$.

Proof. We distinguish two cases. In the first case assume that S has a neighbor U' in T, with $|U'| > |S|$, and apply Lemma 20 with respect to S and the C_j $(j \in M)$. The resulting optimal embedding v'_i has dimension at most $\dim \mathcal{L} + 2$, and since $\dim \mathcal{L} \le |S| - 1$ $(0 \in \operatorname{conv}\{v_s : s \in S\})$ the dimension is at most $|U'|$.

In the second case all neighbors U of S in T satisfy $|U| \le |S|$. By definition, no two nodes in $\mathcal{N}$ are identical, so each set C_j is separated from S by a subset $S_j := S \cap U_j$ induced by an edge $\{S, U_j\} \in \mathcal{E}$, with $|S_j| < |S|$. By applying Lemma 20 with respect to S and the C_j $(j \in M)$ we obtain the corresponding embedding v'_i $(i \in N)$ with a unique index $\bar{\jmath} \in M$ satisfying $b_{\bar{\jmath}} = d_1$. Because $S_{\bar{\jmath}} \subset S$ separates S and $C_{\bar{\jmath}}$, at most $|S| - 1$ nodes of S are incident to nodes in $C_{\bar{\jmath}}$. Therefore we may apply Lemma 25 with respect to S, the C_j $(j \in M)$, and the embedding v'_i $(i \in N)$ and obtain an optimal embedding v''_i $(i \in N)$ of dimension at most $|S|$. ∎

If, however, one of the sets is too big to be flattened out, we can find a unique edge that leads us towards a more balanced center in the big set.

LEMMA 29. *Given the setting of Lemma 28, assume that $\tilde{\delta}_{\hat{\jmath}} > \sum_{j \in M \setminus \{\hat{\jmath}\}} \tilde{\delta}_j$ for a $\hat{\jmath} \in M$. Let v'_i $(i \in N)$ be an optimal embedding arising from Lemma 21 for this S and the C_j $(j \in M)$. The (unique) edge $\{S, \widehat{U}\} \in \mathcal{E}$, with $\widehat{U} \in \mathcal{N}_{\hat{\jmath}}$, is a zero-edge with respect to this new optimal embedding.*

Proof. Since $\tilde{\delta}_{\hat{\jmath}} > 0$, neither the subtree $T_{\hat{\jmath}}$ nor $C_{\hat{\jmath}}$ are empty, so there is an edge $\{S, \widehat{U}\} \in \mathcal{E}$, with $\widehat{U} \in \mathcal{N}_{\hat{\jmath}}$. Suppose, for contradiction, that it is not a zero-edge with respect to the embedding v'_i $(i \in N)$. Then $S' := S \cap \widehat{U}$ separates G into $C_{\hat{\jmath}}$ and $N \setminus (S' \cup C_{\hat{\jmath}})$. By assumption, $0 \notin \operatorname{conv}\{v_s : s \in S'\}$ and $0 \in \operatorname{conv}\{v_s : s \in S\}$, so the separator-shadow Theorem 3 applied with respect to the separator S' implies that $v_i \in \operatorname{cone}\{v_s : s \in S'\} \subset \mathcal{L}$ for $i \in C_{\hat{\jmath}}$. But then $\tilde{\delta}_{\hat{\jmath}} = 0$. ∎

Note that $\widehat{U}$ is a zero-node of the embedding v'_i, and we could continue with transforming v'_i with respect to $\widehat{U}$ ending up in Lemma 28 or Lemma 29 again. However, in order to ensure that no edge is crossed twice, we need to look at the zero-edge itself first.

LEMMA 30. *Consider the setting of Lemma 29 with $\{S, \widehat{U}\} \in \mathcal{E}$ being the zero-edge with respect to embedding v'_i $(i \in N)$ satisfying $\widehat{U} \in \mathcal{N}_{\hat{\jmath}}$. Deleting this edge in T splits T into two subtrees $(\mathcal{N}'_j, \mathcal{E}'_j) =: T'_j$, with $j \in \{S, \widehat{U}\}$, so that $S \in \mathcal{N}'_S$ and*

$\widehat{U} \in \mathcal{N}'_{\widehat{U}}$. *Put*

$$\begin{aligned} S' &:= S \cap \widehat{U}, \\ \mathcal{L}' &:= \operatorname{span}\{v_s : s \in S'\}, \\ C'_j &:= \bigcup_{U \in \mathcal{N}'_j} U \setminus S', \\ \tilde{\delta}'_j &:= \sum_{i \in C_j} \|p_{\mathcal{L}'^{\perp}}(v'_i)\| \quad (j \in \{S, \widehat{U}\}). \end{aligned}$$

If $\tilde{\delta}'_S \geq \tilde{\delta}'_{\widehat{U}}$, *then there is an optimal embedding* v''_i $(i \in N)$ *of dimension at most* $|S|$.

Proof. If $\tilde{\delta}'_S = \tilde{\delta}'_{\widehat{U}}$, then Lemma 20 applied to embedding v'_i with respect to S' and C'_j for $j \in \{S, \widehat{U}\}$ yields an optimal embedding

$$v''_i := \begin{cases} v'_i & \text{for } \ i \in S', \\ p_{\mathcal{L}'}(v_i) + \|p_{\mathcal{L}'^{\perp}}(v_i)\| b & \text{for } \ i \in C'_S, \\ p_{\mathcal{L}'}(v_i) - \|p_{\mathcal{L}'^{\perp}}(v_i)\| b & \text{for } \ i \in C'_{\widehat{U}} \end{cases}$$

for some normalized $b \in \mathcal{L}'^{\perp}$, and the dimension is bounded by $\dim \mathcal{L}' + 1 \leq |S'| \leq |S|$.

For $\tilde{\delta}_S > \tilde{\delta}_{\widehat{U}}$ remember that the v'_i were constructed via Lemma 21. So with the definitions of $\bar{b}$ and the v'_i given there, we have

$$v'_i = p_{\mathcal{L}}(v_i) + \|p_{\mathcal{L}^{\perp}}(v_i)\| \bar{b} \quad \text{for } i \in C'_S = \bigcup_{j \in M \setminus \{\hat{\jmath}\}} C_j \cup S \setminus S'.$$

If $\bar{b} = 0$, then all of these v'_i lie in $\mathcal{L}$, and, by applying Lemma 21 to v'_i with respect to S' and C'_j for $j \in \{S, \widehat{U}\}$, the space of the new optimal embedding v''_i, $i \in N$, will be $\mathcal{L}$ enlarged by some direction h at most, so its dimension is bounded by $\dim \mathcal{L} + 1 \leq |S|$.

If $\bar{b} \neq 0$, then $p_{\mathcal{L}^{\perp}}(v_i) \neq 0$ for some $i \in C'_S$, and, by using $\bar{b} \in \mathcal{L}^{\perp}$, $\|\bar{b}\| = 1$, we get

$$\bar{b}^T v'_i = \bar{b}^T p_{\mathcal{L}}(v_i) + \|p_{\mathcal{L}^{\perp}}(v_i)\| \bar{b}^T \bar{b} = \|p_{\mathcal{L}^{\perp}}(v_i)\| \geq 0 \quad \text{for } i \in C'_S.$$

Since $\mathcal{L}' \subseteq \mathcal{L}$ we obtain $\bar{b} \in \operatorname{span}\{v'_i : i \in C'_S\} \cap \mathcal{L}'^{\perp} \setminus \{0\}$ and $\bar{b}^T v'_i \geq 0$ for $i \in C'_S$. So we are in the special case of Lemma 21. Thus, applying Lemma 21 to v'_i with respect to S' and C'_j for $j \in \{S, \widehat{U}\}$ yields a new optimal embedding v''_i, $i \in N$, with $v''_i \in \operatorname{span}\{v'_i : i \in C'_S\} \subseteq \mathcal{L} + \operatorname{span}\{\bar{b}\}$ for $i \in C'_S$, and therefore $v''_i \in \mathcal{L} + \operatorname{span}\{\bar{b}\}$ for all $i \in N$. The dimension of this new embedding is again bounded by $|S|$. □

In proving the finiteness of the algorithm below we will see that the values $\tilde{\delta}'_j$ of Lemma 30 do not change if the algorithm turns back in $\hat{U}$ to cross the same edge again, so that the condition $\tilde{\delta}'_S \geq \tilde{\delta}'_{\hat{U}}$ will be met the second time at the latest.

ALGORITHM 31.

Input*: A connected graph* $G = (N, E)$, *a tree decomposition* $T = (\mathcal{N}, \mathcal{E})$ *of* G, *an optimal embedding* v_i, $i \in N$, *of* (4).

Step* 0: Set* S *to a zero-vertex of* T *with respect to the embedding.*

Step* 1: By using the notation of Lemma* 28 *with respect to* S, *determine* $\tilde{\delta}_j$ *for* $j \in M$.

Step* 2: If* $\tilde{\delta}_{\hat{\jmath}} \leq \sum_{j \in M \setminus \{\hat{\jmath}\}} \tilde{\delta}_j$ *for all* $\hat{\jmath} \in M$, *apply the proof of Lemma* 28 *to find an optimal embedding of dimension at most the width of* T *plus one, and stop.*

Step 3*: Transform, as described in Lemma* 29*, the optimal embedding to* v'_i *(*$i \in N$*), and compute the corresponding zero-edge* $\{S, \widehat{U}\}$*. Determine* $\tilde{\delta}'_S$ *and* $\tilde{\delta}'_{\widehat{U}}$ *in the notation of Lemma* 30*.*

Step 4*: If* $\tilde{\delta}_S \geq \tilde{\delta}_{\widehat{U}}$*, apply the proof of Lemma* 30 *to find an optimal embedding of dimension at most the width of* T *plus one, and stop.*

Step 5*: Set* $S \leftarrow \widehat{U}$*,* $v_i \leftarrow v'_i$ *for* $i \in N$*, and goto Step* 1*.*

THEOREM 32. *Let* $G = (N, E)$ *be a connected graph and* $T = (\mathcal{N}, \mathcal{E})$ *a tree decomposition of* G*. Algorithm* 31 *is correct and stops with an optimal embedding for* (4) *of dimension at most width of* T *plus one in at most* $|\mathcal{N}|$ *iterations.*

Proof. Step 0 can be carried through by Lemma 26.

If in Step 1 the set M is empty ($\mathcal{N} = \{S\}$), then the condition in Step 2 is vacuously satisfied and the transformation of Lemma 28 is the identity. But in this case $S = |N|$ and any optimal embedding has dimension at most $|N| - 1$ by Observation 9, so the algorithm stops correctly.

We will prove that the algorithm steps over any edge at most once without stopping, and this will yield the iteration bound. Suppose that $\{S, \widehat{U}\}$ is the first edge of T to be considered a second time and that the algorithm just stepped from S to $\widehat{U}$ and now considers stepping back to S. Then $\widehat{U}$ transforms, by means of Lemma 29, the embedding v'_i that was generated by S via Lemma 29. By construction (see Lemma 21), both transformations have $\mathcal{L}' := \operatorname{span}\{v'_i : i \in S \cap \widehat{U}\}$ as an invariant subspace. Therefore the numbers $\tilde{\delta}'_S$ and $\tilde{\delta}'_{\widehat{U}}$ of Lemma 30 computed in Step 3 have identical values in both cases (but with names interchanged), so the condition of Step 4 is certainly satisfied the second time, and the algorithm stops.

The correctness of the statement regarding the dimension of the optimal embedding at termination is a consequence of the respective Lemmas 28 and 30. ∎

Acknowledgments. We thank two anonymous referees for their careful and considerate reports that helped to restructure and improve the presentation of the paper significantly.

REFERENCES

[1] K. L. BOYER AND K. SENGUPTA, *Modelbase partitioning using property matrix spectra*, Computer Vision and Image Understanding, 70 (1998), pp. 177–196.

[2] T. F. CHAN, P. CIARLET, JR., AND W. K. SZETO, *On the optimality of the median cut spectral bisection graph partitioning method*, SIAM J. Sci. Comput., 18 (1997), pp. 943–948.

[3] D. CVETKOVIĆ, M. DOOB, AND H. SACHS, *Spectra of Graphs. Theory and Application*, 3rd ed., J. A. Barth Verlag, Leipzig, 1995.

[4] E. DE KLERK, C. ROOS, AND T. TERLAKY, *Initialization in semidefinite programming via a self-dual skew-symmetric embedding*, Oper. Res. Lett., 20 (1997), pp. 213–221.

[5] R. DIESTEL, *Graph Theory*, 2nd ed., Springer, Berlin, 2000.

[6] M. FIEDLER, *Algebraic connectivity of graphs*, Czechoslovak Math. J., 23 (1973), pp. 298–305.

[7] M. FIEDLER, *Laplacian of graphs and algebraic connectivity*, Combin. Graph Theory, 25 (1989), pp. 57–70.

[8] M. FIEDLER, *Absolute algebraic connectivity of trees*, Linear Multilinear Algebra, 26 (1990), pp. 85–106.

[9] M. FIEDLER, *A Geometric Approach to the Laplacian Matrix of a Graph*, IMA Vol. Math. Appl. 50, Springer, New York, 1993, pp. 73–98.

[10] M. FIEDLER, *Some minimax problems for graphs*, Discrete Math., 121 (1993), pp. 65–74.

[11] C. HELMBERG, B. MOHAR, S. POLJAK, AND F. RENDL, *A spectral approach to bandwidth and separator problems in graphs*, Linear Multilinear Algebra, 39 (1995), pp. 73–90.

[12] X. JI AND H. ZHA, *Extracting Shared Topics of Multiple Documents*, Lect. Notes Comput. Sci. 2637, Springer, Berlin, 2003, pp. 100–110.

[13] M. JUVAN AND B. MOHAR, *Laplace eigenvalues and bandwidth-type invariants of graphs*, J. Graph Theory, 17 (1993), pp. 393–407.

[14] A. Kaveh and H. A. Rahimi Bondarabady, *Bisection for parallel computing using Ritz and Fiedler vectors*, Acta Mech., 167 (2004), pp. 131–144.
[15] J. Keuchel, C. Schnörr, C. Schellewald, and D. Cremers, *Binary partitioning, perceptual grouping, and restoration with semidefinite programming*, IEEE Trans. Pattern Analysis and Machine Intelligence, 25 (2003), pp. 1364–1379.
[16] L. Lovász, *On the Shannon capacity of a graph*, IEEE Trans. Inform. Theory, IT-25 (1979), pp. 1–7.
[17] B. Mohar, *The Laplacian spectrum of graphs*, in Graph Theory, Combinatorics, and Applications, John Wiley and Sons, New York, 1991, pp. 871–898.
[18] B. Mohar, *Graph Laplacians*, in Topics in Algebraic Graph Theory, Encyclopedia Math. Appl. 102, Beineke et al., eds., Cambridge University Press, London, 2004, pp. 113–136.
[19] B. Mohar and S. Poljak, *Eigenvalues in combinatorial optimization*, in Combinatorial and Graph Theoretic Problems in Linear Algebra, IMA Vol. Math. Appl. Vol. 50, R. Brualdi, S. Friedland, and V. Klee, eds., Springer, New York, 1993.
[20] A. Pothen, H. D. Simon, and K. Liou, *Partitioning sparse matrices with eigenvectors of graphs*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 430–452.
[21] H. D. Simon, A. Sohn, and R. Biswas, *Harp: A dynamic spectral partitioner*, J. Parallel and Distributed Computing, 50 (1998), pp. 83–103.
[22] A. Man-Cho So, Y. Ye, and J. Zhang, *A unified theorem on SDP rank reduction*, Math. Oper. Res., to appear.
[23] J. F. Sturm, *Using SeDuMi* 1.02*, A MATLAB Toolbox for Optimization over Symmetric Cones (updated for version* 1.05*)*, manual, Department of Econometrics, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands, 2001.
[24] J. Sun, S. Boyd, L. Xiao, and P. Diaconis, *The fastest mixing Markov process on a graph and a connection to a maximum variance unfolding problem*, SIAM Rev., 48 (2006), pp. 681–699.
[25] H. van der Holst, L. Lovász, and A. Schrijver, *The Colin de Verdière graph parameter*, in Graph Theory and Combinatorial Biology, Bolyai Soc. Math. Stud. 7, Lovász et al., ed., eds., Bolyai Mathematical Society, Budapest, 1999, pp. 29–85.
[26] K. Q. Weinberger and L. K. Saul, *Unsupervised learning of image manifolds by semidefinite programming*, Internat. J. Comput. Vision, 70 (2006), pp. 77–90.
[27] K. Q. Weinberger, F. Sha, and L. K. Saul, *Learning a kernel matrix for nonlinear dimensionality reduction*, in Proceedings of the Twenty-First International Conference on Machine Learning (ICML-04), Banff, Canada, ACM, New York, 2004, pp. 839–846.
[28] H. Wolkowicz, R. Saigal, and L. Vandenberghe, eds., *Handbook of Semidefinite Programming*, Internat. Ser. Oper. Res. Management Sci. 27, Kluwer Academic Publishers, Boston/Dordrecht/London, 2000.
[29] C. W. Wu, *On Rayleigh-Ritz ratios of a generalized Laplacian matrix of directed graphs*, Linear Algebra Appl., 402 (2005), pp. 207–227.

SIAM J. OPTIM.
Vol. 19, No. 2, pp. 503–523

SEMIDEFINITE RELAXATION BOUNDS FOR INDEFINITE HOMOGENEOUS QUADRATIC OPTIMIZATION*

SIMAI HE†, ZHI-QUAN LUO‡, JIAWANG NIE§, AND SHUZHONG ZHANG†

Abstract. This paper studies the relationship between the optimal value of a homogeneous quadratic optimization problem and its semidefinite programming (SDP) relaxation. We consider two quadratic optimization models: (1) $\min\{x^*Cx \mid x^*A_kx \geq 1,\ k = 0, 1, \ldots, m,\ x \in \mathbb{F}^n\}$ and (2) $\max\{x^*Cx \mid x^*A_kx \leq 1,\ k = 0, 1, \ldots, m,\ x \in \mathbb{F}^n\}$, where $\mathbb{F}$ is either the real field $\mathbb{R}$ or the complex field $\mathbb{C}$, and A_k, C are symmetric matrices. For the minimization model (1), we prove that if the matrix C and all but *one* of the A_k's are positive semidefinite, then the ratio between the optimal value of (1) and its SDP relaxation is upper bounded by $O(m^2)$ when $\mathbb{F} = \mathbb{R}$, and by $O(m)$ when $\mathbb{F} = \mathbb{C}$. Moreover, when two or more of the A_k's are indefinite, this ratio can be arbitrarily large. For the maximization model (2), we show that if C and at most one of the A_k's are indefinite while other A_k's are positive semidefinite, then the ratio between the optimal value of (2) and its SDP relaxation is bounded from below by $O(1/\log m)$ for both the real and the complex case. This result improves the bound based on the so-called approximate S-Lemma of Ben-Tal, Nemirovski, and Roos [*SIAM J. Optim.*, 13 (2002), pp. 535–560]. When two or more of the A_k's in (2) are indefinite, we derive a general bound in terms of the problem data and the SDP solution. For both optimization models, we present examples to show that the derived approximation bounds are essentially tight.

Key words. quadratic optimization, SDP relaxation, approximation ratio, probabilistic solution

AMS subject classifications. 90C20, 90C22, 68W20

DOI. 10.1137/070679041

1. Introduction.

In this paper we study the relationship between the optimal values of a homogeneous quadratic optimization problem and its *semidefinite programming* (SDP) relaxation. Two specific optimization models are considered.

The minimization model. Let A_k $(k = 0, 1, \ldots, m)$ and C be $n \times n$ real symmetric or complex Hermitian matrices. Consider

$$\begin{aligned} \min \quad & x^*Cx \\ \text{s.t.} \quad & x^*A_kx \geq 1,\ k = 0, 1, \ldots, m \\ & x \in \mathbb{F}^n, \end{aligned} \tag{1.1}$$

where $\mathbb{F}$ can be either the field of real numbers $\mathbb{R}$ or the field of complex numbers $\mathbb{C}$, and the superscript * represents Hermitian transpose (or regular transpose in case of real matrices). The quadratic optimization problems of form (1.1) are NP-hard [14], even when all the data matrices, C and A_k, $k = 1, \ldots, m$, are positive semidefinite. Homogeneous quadratic optimization problems (1.1) arise naturally in telecommunications and robust control applications; see [22, 17, 14] and the references therein.

*Received by the editors January 2, 2007; accepted for publication (in revised form) January 3, 2008; published electronically June 11, 2008.
http://www.siam.org/journals/siopt/19-2/67904.html

†Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, Hong Kong (smhe@se.cuhk.edu.hk, zhang@se.cuhk.edu.hk). The fourth author's research was supported by Hong Kong RGC earmarked grants CUHK418505 and CUHK418406.

‡Department of Electrical and Computer Engineering, University of Minnesota, 200 Union Street SE, Minneapolis, MN 55455 (luozq@ece.umn.edu). This author's research was supported in part by U.S. NSF grants DMS-0312416 and DMS-0610037.

§Department of Mathematics, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093 (njw@math.ucsd.edu).

In these applications, the optimization variables are naturally complex since they represent the in-phase and quadrature components of a complex signal. A popular approach to approximately solving the NP-hard quadratic problem (1.1) is to use the following SDP relaxation:

$$\begin{aligned} \min \quad & \mathrm{Tr}\,(CX) \\ \text{s.t.} \quad & \mathrm{Tr}\,(A_k X) \geq 1,\ k = 0, 1, \ldots, m \\ & X \in \mathbb{SF}^n_+, \end{aligned}$$

where $\mathrm{Tr}\,(\cdot)$ represents the trace of a matrix, and $\mathbb{SF}^n_+$ denotes the convex cone of positive semidefinite matrices in the space of all (Hermitian) symmetric matrices $\mathbb{SF}^n$. The above SDP is convex and can be solved efficiently using interior point methods. After the SDP relaxation problems are solved, we can apply a *probabilistic method* to the corresponding optimal SDP solution to extract rank-one feasible solutions for (1.1). Theoretically, even though the probabilistic solutions obtained in this manner are not globally optimal for (1.1), they can be shown to be high quality approximate solutions; see, e.g., [3, 14]. Recently, Luo et al. [14] considered problem (1.1) and gave bounds for the SDP approximation ratio for (1.1). When all of the matrices A_k and C are positive semidefinite, Luo et al. [14] showed that the ratio between the original optimal value and the SDP relaxation optimal value is bounded above by $O(m^2)$ when $\mathbb{F} = \mathbb{R}$ and by $O(m)$ when $\mathbb{F} = \mathbb{C}$, where the factors in the big O notations are absolute constants and independent of data matrices A_k and C. All of these bounds are shown to be tight in the worst case. However, the average performance can be much better than the stated worst-case bounds for randomly generated instances. The simulation studies in [14] showed that the average ratios are typically close to 1.

Our contributions. In section 3, we analyze the approximation ratio of the SDP relaxation for the homogeneous quadratic optimization problem (1.1) when some of the constraint matrices A_k are *indefinite*. We show that, for problem (1.1), the same upper bounds for the SDP approximation ratios as given in [14] ($O(m^2)$ when $\mathbb{F} = \mathbb{R}$ and $O(m)$ when $\mathbb{F} = \mathbb{C}$) still hold true even when *one* of the constraint matrices is indefinite. If there are more than one indefinite quadratic constraints, then we show by an example that the SDP approximation ratio can be infinity. Therefore, our bounds are essentially the best possible.

The maximization model. We also consider the quadratic optimization problem of the form

$$\begin{aligned} \max \quad & x^* C x \\ \text{s.t.} \quad & x^* A_k x \leq 1,\ k = 0, 1, \ldots, m \\ & x \in \mathbb{F}^n. \end{aligned} \tag{1.2}$$

This quadratic optimization problem is still NP-hard [3, 18], even when all of the matrices C and A_k are positive semidefinite. Problem (1.2) arises naturally in telecommunications and robust control applications; see [22, 17, 3] and the references therein. The SDP relaxation for (1.2) can be written as follows:

$$\begin{aligned} \max \quad & \mathrm{Tr}\,(CX) \\ \text{s.t.} \quad & \mathrm{Tr}\,(A_k X) \leq 1,\ k = 0, 1, \ldots, m \\ & X \in \mathbb{SF}^n_+. \end{aligned}$$

As in the minimization case, after the SDP relaxation problem is solved, some probabilistic method can be applied to extract a high quality rank-one feasible solution for (1.2). Various estimates exist for the quality of approximate solutions; see, e.g., [3, 18]. Specifically, Nemirovski, Roos, and Terlaky [18] proved that if all A_k's are positive semidefinite, then the ratio between the optimal value of the SDP relaxation problem and that of the original quadratic problem is bounded above by $O(\log m)$. More generally, Ben-Tal, Nemirovski, and Roos [3] established a so-called *approximate S-Lemma* which shows that the approximation ratio for the SDP relaxation is at most $O(\log(n^2 m))$ when all but one of the matrices A_k, $k = 0, 1, \ldots, m$ are positive semidefinite.

Our contributions. In section 4, we study the SDP approximation ratio for the homogeneous quadratic maximization problem (1.2) when some of the constraint matrices $\{A_k\}$ are *indefinite*. Our results are as follows. We strengthen the approximate S-Lemma of Ben-Tal, Nemirovski, and Roos [3] by improving their upper bound on the SDP approximation ratio from $O(\log(n^2 m))$ to $O(\log m)$ when *one* quadratic inequality is indefinite. In the process of establishing this new bound, we give a universal lower bound of 0.03 on the probability that a homogeneous quadratic form of n binary i.i.d. Bernoulli random variables lies below its mean. The previous best known lower bound for this probability was $1/(8n^2)$ due to Ben-Tal, Nemirovski, and Roos [3]. In this reference, the authors also conjectured that the actual lower bound should be 0.25. We also present a new and unifying upper bound on the ratio of the optimal value of SDP relaxation over that of the original quadratic maximization problem (1.2) *without* any positive definiteness assumptions. This new general bound involves the problem data and the SDP optimal solution, which is computable in polynomial time. We also present an example showing that this bound is essentially the best possible.

Related literature. In addition to the work of Ben-Tal, Nemirovski, and Roos [3], Luo et al. [14], and Nemirovski, Roos, and Terlaky [18], there is a sizeable literature on the quality bounds of SDP relaxation for solving nonconvex quadratic optimization problems. For instance, for problem (1.2), when $m = n$, $A_i = e_i e_i^T$ (there is no A_0), and C is positive semidefinite with nonpositive nondiagonal entries and row sums 0 (which corresponds to the maximum cut problem), Goemans and Williamson [8] showed that the ratio of the optimal value of SDP relaxation over that of the original quadratic maximization problem (1.2) is bounded below by $0.87856\ldots$. Furthermore, if C is only positive semidefinite, Nesterov [19] showed the same ratio is bounded below by $0.6366\ldots$. For closely related results, see Ye [24] and Bertsimas and Ye [4]. Recently, So, Ye, and Zhang [23] developed SDP relaxation methods for finding approximate low rank solutions for linear matrix inequalities. Their results unify and extend the approximation bounds of Nemirovski, Roos, and Terlaky [18] and Luo et al. [14] for the case when all of the data matrices are positive semidefinite. Beck and Teboulle [2] discussed the nonconvex problem of minimizing the ratio of two nonconvex quadratic functions over a possibly degenerate ellipsoid, and showed that the SDP relaxation can return exact solutions under a certain condition. There is also some work on solving quadratic optimization problems using other methods, e.g., Hiriart-Urruty and Jean-Baptiste [10], Jeyakumar, Rubinov, and Wu [12], and Madsen, Nielsen, and Pinar [15, 16].

Outline of this paper. Section 2 is devoted to analyzing the probability of a general random variable to be above (or below) its mean value. Section 3 concentrates on the SDP approximation bound for the quadratic minimization problem (1.1), while section 4 studies the SDP approximation bound for quadratic maximization

problem (1.2). Some concluding remarks are given in the last section (section 5).

2. Estimating the asymmetry of a random variable. To facilitate the technical analysis in subsequent sections, we establish in this section a bound on the probability for a general random variable to be above (or symmetrically, below) its mean value, using only the high order moment information of the random variable. This problem on its own is of importance in statistics and probability theory. The following lemma is a generalization of Theorem 2.1 in [13].

LEMMA 2.1. *Suppose that a random variable* Φ *satisfies* $\mathsf{E}\Phi = 0$, $\mathrm{Var}(\Phi) = 1$, *and* $\mathsf{E}|\Phi|^t \le \tau$ *for some* $t > 2$ *and* $\tau > 0$. *Then* $\mathrm{Prob}\{\Phi \ge 0\} > 0.25\tau^{-\frac{2}{t-2}}$ *and* $\mathrm{Prob}\{\Phi \le 0\} > 0.25\tau^{-\frac{2}{t-2}}$.

Proof. Let $p_1 = \mathrm{Prob}\{\Phi \ge 0\}$ and $p_2 = \mathrm{Prob}\{\Phi \le 0\}$. Also, let $Y_1 = \max(\Phi, 0)$ and $Y_2 = -\min(\Phi, 0)$. Since $\mathsf{E}\Phi = 0$, we know that $\mathsf{E}Y_1 - \mathsf{E}Y_2 = 0$. Let $s := \mathsf{E}Y_1 = \mathsf{E}Y_2$. By Hölder's inequality it follows that $(\mathsf{E}Y_1^t)^{1/(t-1)}(\mathsf{E}Y_1)^{(t-2)/(t-1)} \ge \mathsf{E}Y_1^2$ and $(\mathsf{E}Y_2^t)^{1/(t-1)}(\mathsf{E}Y_2)^{(t-2)/(t-1)} \ge \mathsf{E}Y_2^2$. Since $\mathsf{E}Y_1^t + \mathsf{E}Y_2^t = \mathsf{E}|\Phi|^t$, we have

$$\tau \ge \mathsf{E}|\Phi|^t = \mathsf{E}Y_1^t + \mathsf{E}Y_2^t \ge \frac{(\mathsf{E}Y_1^2)^{t-1} + (\mathsf{E}Y_2^2)^{t-1}}{s^{t-2}}.$$

Let $u = \mathsf{E}Y_1^2 \in [0,1]$. Since $\mathsf{E}Y_1^2 + \mathsf{E}Y_2^2 = \mathsf{E}\Phi^2 = \mathrm{Var}(\Phi) = 1$, it follows that $s^{t-2} \ge \frac{u^{t-1}+(1-u)^{t-1}}{\tau}$. On the other hand, by the Cauchy–Schwartz inequality, we have

$$s^2 = (\mathsf{E}Y_1)^2 = (\mathsf{E}(\mathbf{1}_{\{Y_1 \ge 0\}} Y_1))^2 \le \mathsf{E}(\mathbf{1}^2_{\{Y_1 \ge 0\}})\mathsf{E}Y_1^2 \le p_1 u,$$

which implies that

$$\begin{aligned} p_1 &\ge u^{-1}\left[\frac{u^{t-1} + (1-u)^{t-1}}{\tau}\right]^{\frac{2}{t-2}} \\ &= \frac{\left(u^{t-1} + (1-u)^{t-1}\right)^{\frac{2}{t-2}}}{u}\,\tau^{-\frac{2}{t-2}} \\ &\ge \left(u^{t-1} + (1-u)^{t-1}\right)^{\frac{2}{t-2}}\tau^{-\frac{2}{t-2}} \\ &\ge \left(2\left(\frac{1}{2}\right)^{t-1}\right)^{\frac{2}{t-2}}\tau^{-\frac{2}{t-2}} \\ &= 0.25\tau^{-\frac{2}{t-2}}, \end{aligned}$$

where the third inequality follows from the convexity of the function u^{t-1} when $t > 2$. Obviously, the equality cannot hold throughout. Therefore, $p_1 > 0.25\tau^{-\frac{2}{t-2}}$. By symmetry, we also have $p_2 > 0.25\tau^{-\frac{2}{t-2}}$. □

In case $t = 4$, Lemma 2.1 asserts that $\mathrm{Prob}\{\Phi \ge 0\} \ge \frac{1}{4\tau}$ and $\mathrm{Prob}\{\Phi \le 0\} \ge \frac{1}{4\tau}$. However, in this particular case, this specific bound can in fact be further sharpened.

LEMMA 2.2. *Suppose that a random variable* Φ *satisfies* $\mathsf{E}\Phi = 0$, $\mathrm{Var}(\Phi) = 1$, *and* $\mathsf{E}\Phi^4 \le \tau$. *Then* $\mathrm{Prob}\{\Phi \ge 0\} \ge \frac{2\sqrt{3}-3}{\tau} > \frac{9}{20\tau}$ *and* $\mathrm{Prob}\{\Phi \le 0\} \ge \frac{2\sqrt{3}-3}{\tau} > \frac{9}{20\tau}$.

Proof. It follows from the proof of Lemma 2.1 that

$$p_1 \ge \frac{u^3 + (1-u)^3}{\tau u} = \left(\frac{1}{u} + 3u - 3\right)\frac{1}{\tau} \ge \frac{2\sqrt{3}-3}{\tau} > \frac{9}{20\tau}.$$

By symmetry, $p_2 > \frac{9}{20\tau}$ holds as well. □

3. SDP relaxation bounds for the quadratic minimization model. Consider the homogeneous quadratic optimization

$$\begin{aligned} v_{qp}^{\min} := \min \quad & x^* C x \\ \text{s.t.} \quad & x^* A_k x \geq 1,\ k = 0, 1, \ldots, m \\ & x \in \mathbb{F}^n, \end{aligned} \tag{3.1}$$

where $C, A_0, A_1, \ldots, A_m \in \mathbb{SF}^n$ are symmetric matrices. This problem is NP-hard [14]. A natural SDP relaxation to the above quadratic optimization problem is

$$\begin{aligned} v_{sdp}^{\min} := \min \quad & \mathrm{Tr}\,(CZ) \\ \text{s.t.} \quad & \mathrm{Tr}\,(A_k Z) \geq 1,\ k = 0, 1, \ldots, m \\ & Z \in \mathbb{SF}^n_+. \end{aligned} \tag{3.2}$$

Obviously, the SDP relaxation provides a lower bound, i.e., $v_{sdp}^{\min} \leq v_{qp}^{\min}$. In the case $C = I_n$ and $A_0, A_1, \ldots, A_m$ are all positive semidefinite, Luo et al. [14] proved that $v_{qp}^{\min}/v_{sdp}^{\min} \leq \frac{27(m+1)^2}{\pi}$ for $\mathbb{F} = \mathbb{R}$, and $v_{qp}^{\min}/v_{sdp}^{\min} \leq 8(m+1)$ for $\mathbb{F} = \mathbb{C}$. Moreover, when two or more of $A_0, A_1, \ldots, A_m$ are indefinite, there is in general no data-independent upper bound on $v_{qp}^{\min}/v_{sdp}^{\min}$, as shown by the following example [14]:

$$\begin{aligned} \min \quad & x_1^2 + x_2^2 \\ \text{s.t.} \quad & x_2^2 \geq 1 \\ & x_1^2 + M x_1 x_2 \geq 1 \\ & x_1^2 - M x_1 x_2 \geq 1, \end{aligned}$$

where $M > 0$ is a constant. In the above example, $v_{sdp}^{\min} = 1$ and the last two constraints imply $x_1^2 \geq M|x_1||x_2| + 1$ which, together with the first constraint $x_2^2 \geq 1$, yield $x_1^2 \geq M|x_1| + 1$ or, equivalently, $|x_1| \geq (M + \sqrt{M^2+4})/2$. Therefore, $v_{qp}^{\min} \geq 1 + \frac{1}{4}(M + \sqrt{M^2+4})^2$. That is, $v_{qp}^{\min}/v_{sdp}^{\min} \geq 1 + \frac{1}{4}(M + \sqrt{M^2+4})^2$, which can be arbitrarily large, depending on the problem data $M > 0$.

In this section, we consider the homogeneous quadratic optimization (3.1) under the assumption that $C, A_1, A_2, \ldots, A_m \in \mathbb{SF}^n_+$ are positive semidefinite while $A_0 \in \mathbb{SF}^n$ can be indefinite. Throughout this section, we assume that (3.1) is feasible, and that there is $\mu_k \geq 0$, $k = 0, 1, \ldots, m$, such that $\sum_{k=0}^m \mu_k A_k \prec C$. This assumption guarantees that the SDP relaxation is primal feasible while its dual problem satisfies the Slater condition. Hence the strong duality holds and the primal problem (3.2) has an optimal solution that attains its infimum.

Our analysis shall treat the cases $\mathbb{F} = \mathbb{R}$ and $\mathbb{F} = \mathbb{C}$ separately, leading to strikingly different bounds.

3.1. The real case. Let us start with a useful lemma regarding a lower bound on worst asymmetric mass distributions for a χ^2-distribution around its mean vector. In fact this result is interesting on its own right.

LEMMA 3.1. *Let τ_i be any real numbers, $i = 1, \ldots, n$, and let $\eta \sim N(0, I_n)$ be an n-dimensional normal distribution with zero mean and covariance matrix I_n. Then*

we have

$$\text{Prob}\left\{\sum_{i=1}^{n}\tau_i(\eta_i^2-1)\geq 0\right\}>\frac{3}{100},\quad \text{Prob}\left\{\sum_{i=1}^{n}\tau_i(\eta_i^2-1)\leq 0\right\}>\frac{3}{100}.$$

Proof. Note that $\mathsf{E}(\eta_i^2-1)^2=\mathsf{E}(\eta_i^4-2\eta_i^2+1)=3-2+1=2$. Let $\Psi=\sum_{i=1}^n\tau_i(\eta_i^2-1)$, and $\Phi=\frac{\Psi}{\sqrt{2\sum_{i=1}^n\tau_i^2}}$. Then $\mathsf{E}\Phi=0$ and $\text{Var}(\Phi)=1$. Since $\mathsf{E}(\eta_i^2-1)^2=2$, and $\mathsf{E}(\eta_i^2-1)^4=60$, direct calculation shows that

$$\mathsf{E}\Psi^4=48\sum_{i=1}^{n}\tau_i^4+12\left(\sum_{i=1}^{n}\tau_i^2\right)^2\leq 60\left(\sum_{i=1}^{n}\tau_i^2\right)^2.$$

Therefore, we have

$$\mathsf{E}\Phi^4=\frac{\mathsf{E}\Psi^4}{4(\sum_{i=1}^n\tau_i^2)^2}\leq 15.$$

It follows from Lemma 2.2 that $\text{Prob}\{\Phi\geq 0\}>\frac{3}{100}$. Similarly, we have $\text{Prob}\{\Phi\leq 0\}>\frac{3}{100}$ by symmetry. ∎

Using Hölder's inequality, we also have $\mathsf{E}|\Psi|^3\leq 60^{\frac{3}{4}}(\sum_{i=1}^n\tau_i^2)^{\frac{3}{2}}$ and $\mathsf{E}|\Phi|^3\leq 15^{\frac{3}{4}}$, which can be used to lower $\text{Prob}\{\Phi\geq 0\}$ (cf. Theorem 2.1 in [13]). However, in this particular case, the bound so obtained is slightly worse than the one that we derived in Lemma 3.1.

LEMMA 3.2. *Let A,Z be two real symmetric matrices with $Z\succeq 0$ and $\text{Tr}\,(AZ)\geq 0$. Let $\xi\in N(0,Z)$ be a normal random vector with zero mean and covariance matrix Z. Then for any $0\leq\gamma\leq 1$ we have*

$$\text{Prob}\{\xi^TA\xi<\gamma\mathsf{E}(\xi^TA\xi)\}<1-\frac{3}{100}.$$

Proof. Let $r=\text{rank}(AZ)$, and let $Q\in\mathbb{R}^{n\times n}$ be an orthogonal matrix such that

$$Q^T(Z^{\frac{1}{2}}AZ^{\frac{1}{2}})Q=\text{diag}(\lambda_1,\ldots,\lambda_r,0,\ldots,0).$$

Since $\text{Tr}\,(AZ)\geq 0$ we have $\sum_{i=1}^r\lambda_i\geq 0$. Let $\bar\xi\in N(0,I_n)$ and $\xi:=Z^{\frac{1}{2}}Q\bar\xi$. Then ξ follows a Gaussian distribution $N(0,Z)$. Moreover, we have $\xi^TA\xi=\sum_{i=1}^r\lambda_i\bar\xi_i^2$, where $\bar\xi_i$, $i=1,\ldots,r$, are independent and follow the normal distribution $N(0,1)$. Therefore, we have $\mathsf{E}(\xi^TA\xi)=\sum_{i=1}^r\lambda_i$ and

$$\begin{aligned}\text{Prob}\{\xi^TA\xi<\gamma\mathsf{E}(\xi^TA\xi)\}&=\text{Prob}\left\{\sum_{i=1}^{r}\lambda_i\bar\xi_i^2<\gamma\sum_{i=1}^{r}\lambda_i\right\}\\&=\text{Prob}\left\{\sum_{i=1}^{r}\lambda_i(\bar\xi_i^2-1)<(\gamma-1)\sum_{i=1}^{r}\lambda_i\right\}\\&\leq\text{Prob}\left\{\sum_{i=1}^{r}\lambda_i(\bar\xi_i^2-1)<0\right\}<1-\frac{3}{100},\end{aligned}$$

where the first inequality follows from $\gamma\in[0,1]$ and $\sum_{i=1}^r\lambda_i\geq 0$, and the last step is due to Lemma 3.1. ∎

Now we are ready to establish the following quality bound for the SDP relaxation. The argument follows closely to those of [14].

THEOREM 3.3. *Consider the real quadratic program* (3.1) *and its SDP relaxation* (3.2)*, where* $\mathbb{F} = \mathbb{R}$*. Then, there holds*

$$\frac{v_{qp}^{\min}}{v_{sdp}^{\min}} \leq \frac{10^6 m^2}{\pi}.$$

Proof. Let $\hat{Z}$ be an optimal solution of the SDP relaxation (3.2) with rank r satisfying $\frac{(r+1)r}{2} \leq m$. The existence of such a matrix solution is well known; cf. Pataki [20]. Moreover, this low rank matrix can be constructed in polynomial time; cf. [11]. Clearly, $r < \sqrt{2m}$. Since $\hat{Z}$ is feasible, $\mathrm{Tr}\,(A_0\hat{Z}) \geq 1$. Define random vector $\xi = \mathcal{N}(0, \hat{Z})$. For any $0 < \gamma \leq 1$ and $\mu > 0$ we have

$$\begin{aligned}
&\mathrm{Prob}\left\{\min_{0\leq k\leq m} \xi^T A_k \xi \geq \gamma,\, \xi^T C\xi \leq \mu\,\mathrm{Tr}\,(C\hat{Z})\right\}\\
&= \mathrm{Prob}\left\{\xi^T A_k\xi \geq \gamma\,\forall k = 0,1,\ldots,m,\ \text{and}\ \xi^T C\xi \leq \mu\,\mathrm{Tr}\,(C\hat{Z})\right\}\\
&\geq \mathrm{Prob}\left\{\xi^T A_k\xi \geq \gamma\,\mathrm{Tr}\,(A_k\hat{Z})\,\forall k = 0,1,\ldots,m,\ \text{and}\ \xi^T C\xi \leq \mu\,\mathrm{Tr}\,(C\hat{Z})\right\}\\
&= \mathrm{Prob}\left\{\xi^T A_k\xi \geq \gamma\,\mathsf{E}(\xi^T A_k\xi)\,\forall k = 0,1,\ldots,m,\ \text{and}\ \xi^T C\xi \leq \mu\,\mathsf{E}(\xi^T C\xi)\right\}\\
&\geq 1 - \sum_{k=0}^{m} \mathrm{Prob}\left\{\xi^T A_k\xi < \gamma\,\mathsf{E}(\xi^T A_k\xi)\right\} - \mathrm{Prob}\left\{\xi^T C\xi > \mu\,\mathsf{E}(\xi^T C\xi)\right\}.
\end{aligned}$$

Since $A_k \succeq 0$ for $k = 1,\ldots,m$, it follows from Lemma 3.1 of [14] that

$$\mathrm{Prob}\left\{\xi^T A_k\xi < \gamma\mathsf{E}(\xi^T A_k\xi)\right\} \leq \max\left\{\sqrt{\gamma}, \frac{2(r-1)\gamma}{\pi-2}\right\}.$$

Although A_0 is indefinite, we can use Lemma 3.2 to obtain

$$\mathrm{Prob}\left\{\xi^T A_0\xi < \gamma\mathsf{E}(\xi^T A_0\xi)\right\} < 1 - \frac{3}{100}.$$

Also, since $C \succeq 0$, we can apply the Markov inequality to obtain

$$\mathrm{Prob}\left\{\xi^T C\xi > \mu\,\mathsf{E}(\xi^T C\xi)\right\} \leq \frac{1}{\mu}.$$

Combining the above estimates yields

$$\mathrm{Prob}\left\{\min_{0\leq k\leq m} \xi^T A_k\xi \geq \gamma,\, \xi^T C\xi \leq \mu\,\mathrm{Tr}\,(C\hat{Z})\right\} > \frac{3}{100} - m\max\left\{\sqrt{\gamma}, \frac{2(r-1)\gamma}{\pi-2}\right\} - \frac{1}{\mu}.$$

Let $\hat{\mu} = 100$ and $\hat{\gamma} = \frac{\pi}{10^4 m^2}$. Since $r < \sqrt{2m}$, we have $\sqrt{\hat{\gamma}} \geq \frac{2(r-1)\hat{\gamma}}{\pi-2}$. Then we have

$$\frac{3}{100} - m\max\left\{\sqrt{\hat{\gamma}}, \frac{2(r-1)\hat{\gamma}}{\pi-2}\right\} - \frac{1}{\hat{\mu}} = \frac{3}{100} - m\frac{\sqrt{\pi}}{100m} - \frac{1}{100} > \frac{1}{500}.$$

Therefore, there exists a vector $\xi \in \mathbb{R}^n$ such that

$$\xi^T A_k \xi \geq \hat{\gamma}, \quad k = 0, 1, \ldots, m, \quad \text{and} \quad \xi^T C \xi \leq \hat{\mu} \mathrm{Tr}\,(C\hat{Z}).$$

Now let $x = \frac{1}{\sqrt{\hat{\gamma}}}\xi$. Then, $x^T A_k x \geq 1$, $k = 0, 1, \ldots, m$, and

$$v_{qp}^{\min} \leq x^T C x = \frac{1}{\hat{\gamma}} \xi^T C \xi \leq \frac{\hat{\mu}}{\hat{\gamma}} \mathrm{Tr}\,(C\hat{Z}) = \frac{10^6 m^2}{\pi}\, v_{sdp}^{\min},$$

which establishes the desired bound. □

3.2. The complex case. Recall that the density function of a complex-valued normal distribution[1] $\eta \sim N_c(0, 1)$ is

$$\frac{1}{\pi} e^{-|u|^2} \, \forall u \in \mathbb{C}.$$

In polar coordinates, the density function becomes

$$\frac{\rho}{\pi} e^{-\rho^2} \, \forall \, \rho \in [0, +\infty) \quad \theta \in [0, 2\pi).$$

The argument θ is uniformly distributed in $[0, 2\pi)$, and the modulus ρ has the distribution

$$f(\rho) = \begin{cases} 2\rho e^{-\rho^2} & \text{if } \rho \geq 0, \\ 0 & \text{if } \rho < 0. \end{cases}$$

Thus squared modulus $|\eta|^2$ has the exponential distribution

$$\mathrm{Prob}\,\{|\eta|^2 \leq \alpha\} \leq 1 - e^{-\alpha}.$$

LEMMA 3.4. *For any real numbers τ_i and i.i.d. exponential random variables η_i with unit variance, $i = 1, \ldots, n$, there holds*

$$\mathrm{Prob}\left\{\sum_{i=1}^n \tau_i(\eta_i - 1) \geq 0\right\} > \frac{1}{20}, \qquad \mathrm{Prob}\left\{\sum_{i=1}^n \tau_i(\eta_i - 1) \leq 0\right\} > \frac{1}{20}.$$

Proof. Note that $\mathsf{E}(\eta_i - 1)^2 = 1$. Let $\Psi = \sum_{i=1}^n \tau_i(\eta_i - 1)$ and $\Phi = \frac{\Psi}{\sqrt{\sum_{i=1}^n \tau_i^2}}$. Clearly, $\mathsf{E}\Phi = 0$ and $\mathrm{Var}(\Phi) = 1$. Since $\mathsf{E}(\eta_i - 1)^4 = 9$, direct calculation shows that

$$\mathsf{E}\Psi^4 = 6\sum_{i=1}^n \tau_i^4 + 3\left(\sum_{i=1}^n \tau_i^2\right)^2 \leq 9\left(\sum_{i=1}^n \tau_i^2\right)^2.$$

This further implies

$$\mathsf{E}\Phi^4 = \frac{\mathsf{E}\Psi^4}{(\sum_{i=1}^n \tau_i^2)^2} \leq 9.$$

Using Lemma 2.2 we have $\mathrm{Prob}\,\{\Phi \geq 0\} > \frac{1}{20}$. Similarly, $\mathrm{Prob}\,\{\Phi \leq 0\} > \frac{1}{20}$. □

[1]For a discussion on the complex normal distribution and the related references, see Zhang and Huang [26].

Interestingly, it is possible to find a closed formula (see, e.g., [7] and [1]) for the above probability. In particular, if all the τ_i's are distinctive, then

$$\text{Prob}\left\{\sum_{i=1}^{n}\tau_i(\eta_i-1)\geq 0\right\}=\sum_{i=1}^{n}\frac{e^{-\frac{1}{\tau_i}}}{\prod_{j\neq i}\left(1-\frac{\tau_j}{\tau_i}\right)}.$$

Therefore, we have

$$\frac{1}{20}<\sum_{i=1}^{n}\frac{e^{-\frac{1}{\tau_i}}}{\prod_{j\neq i}\left(1-\frac{\tau_j}{\tau_i}\right)}<\frac{19}{20}$$

for any distinctive real values τ_i, $i=1,\ldots,n$.

LEMMA 3.5. *Let A, Z be two Hermitian matrices satisfying $Z\succeq 0$ and $\text{Tr}\,(AZ)\geq 0$. Let $\xi\sim N_c(0,Z)$ be a complex normal random vector. Then, for any $0\leq\gamma\leq 1$, we have*

$$\text{Prob}\left\{\xi^*A\xi<\gamma\mathsf{E}(\xi^*A\xi)\right\}<1-\frac{1}{20}.$$

Proof. Let $Q\in\mathbb{C}^{n\times n}$ be a unitary matrix such that

$$Q^*(Z^{\frac{1}{2}}AZ^{\frac{1}{2}})Q=\text{diag}(\lambda_1,\ldots,\lambda_r,0,\ldots,0),$$

where $r=\text{rank}(AZ)$. Since $\text{Tr}\,(AZ)\geq 0$, it follows that $\sum_{i=1}^{r}\lambda_i\geq 0$. Let $\hat{\xi}\in\mathbb{C}^n$ be a random Gaussian vector drawn from the complex normal distribution $N_c(0,I_n)$. Then the random vector $\xi=Z^{\frac{1}{2}}Q\hat{\xi}$ follows the Gaussian distribution $N_c(0,Z)$. As a result, there holds

$$\begin{aligned}\text{Prob}\left\{\xi^*A\xi<\gamma\mathsf{E}(\xi^*A\xi)\right)&=\text{Prob}\left\{\sum_{i=1}^{r}\lambda_i|\hat{\xi}_i|^2<\gamma\sum_{i=1}^{n}\lambda_i\right\}\\&=\text{Prob}\left\{\sum_{i=1}^{n}\lambda_i(|\hat{\xi}_i|^2-1)<(\gamma-1)\sum_{i=1}^{n}\lambda_i\right\}\\&\leq\text{Prob}\left\{\sum_{i=1}^{n}\lambda_i(|\hat{\xi}_i|^2-1)<0\right\},\end{aligned}$$

where the last step follows from $\gamma\in[0,1]$ and $\sum_{i=1}^{r}\lambda_i\geq 0$. Since $|\xi_i|^2$ is exponentially distributed, by Lemma 3.4 we have

$$\text{Prob}\left\{\sum_{i=1}^{n}\lambda_i(|\hat{\xi}_i|^2-1)\geq 0\right\}>\frac{1}{20},$$

which proves the lemma. ∎

THEOREM 3.6. *Consider* (3.1) *and* (3.2)*, where $\mathbb{F}=\mathbb{C}$. Then*

$$\frac{v_{qp}^{\min}}{v_{sdp}^{\min}}\leq 2400m.$$

Proof. It is known that in this case, if $v_{sdp}^{\min}$ is finite and $m \le 3$, then $v_{qp}^{\min}/v_{sdp}^{\min} = 1$ (cf., e.g., [11] and [25]). Below we shall consider only the case where $m \ge 4$. Let $\hat{Z}$ be a low rank optimal solution of the SDP relaxation (3.2) such that $r = \mathrm{rank}(\hat{Z}) \le \sqrt{m}$ (see [11, section 5]). The feasibility of $\hat{Z}$ implies that $\mathrm{Tr}\,(A_0\hat{Z}) \ge 1$. Similar to Theorem 3.3, we can use the union bound to obtain the following inequality:

$$\begin{aligned}&\mathrm{Prob}\left\{\min_{0\le k\le m} \xi^* A_k\xi \ge \gamma,\, \xi^* C\xi \le \mu\,\mathrm{Tr}\,(C\hat{Z})\right\}\\ &\ge 1 - \sum_{k=0}^{m} \mathrm{Prob}\,\{\xi^* A_k\xi < \gamma\,\mathsf{E}(\xi^* A_k\xi)\} - \mathrm{Prob}\,\{\xi^* C\xi > \mu\,\mathsf{E}(\xi^* C\xi)\}.\end{aligned}$$

Since $A_k \succeq 0$, $k = 1, \ldots, m$, it follows from Lemma 3.4 in [14] that

$$\mathrm{Prob}\,\{\xi^* A_k\xi < \gamma\mathsf{E}(\xi^* A_k\xi)\} \le \max\left\{\frac{4}{3}\gamma, 16(r-1)^2\gamma^2\right\}.$$

Although A_0 is indefinite, Lemma 3.5 asserts that

$$\mathrm{Prob}\,\{\xi^* A_0\xi < \gamma\mathsf{E}(\xi^* A_0\xi)\} < 1 - \frac{1}{20}.$$

Therefore, combining these estimates and using the Markov inequality, we have

$$\begin{aligned}&\mathrm{Prob}\left\{\min_{0\le k\le m} \xi^* A_k\xi \ge \gamma, \xi^* C\xi \le \mu,\, \mathrm{Tr}\,(C\hat{Z})\right\}\\ &> \frac{1}{20} - m\max\left\{\frac{4}{3}\gamma, 16(r-1)^2\gamma^2\right\} - \frac{1}{\mu}.\end{aligned}$$

Now choose $\hat{\mu} = 60$ and $\hat{\gamma} = \frac{1}{40m}$. In this case, $\frac{4}{3}\hat{\gamma} \ge \hat{1}6(r-1)^2\hat{\gamma}^2$. We also have a strict lower bound of the above probability:

$$\mathrm{Prob}\left\{\min_{0\le k\le m} \xi^* A_k\xi \ge \hat{\gamma},\, \xi^* C\xi \le \hat{\mu}\,\mathrm{Tr}\,(C\hat{Z})\right\} > 0.$$

This implies that there exists $\xi \in \mathbb{C}^n$ such that

$$\xi^* A_k\xi \ge \hat{\gamma},\ k = 0, 1, \ldots, m; \quad \xi^* C\xi \le \hat{\mu}\mathrm{Tr}\,(C\hat{Z}).$$

Now let $x := \frac{1}{\sqrt{\hat{\gamma}}}\xi$. Then $x^* A_k x \ge 1$, $k = 0, 1, \ldots, m$, and so

$$v_{qp}^{\min} \le x^* C x \le \frac{\xi^* C\xi}{\hat{\gamma}} \le \frac{\hat{\mu}\mathrm{Tr}\,(C\hat{Z})}{\hat{\gamma}} = 2400m \cdot v_{sdp}^{\min}.$$

The theorem is proven. □

Notice that there are examples (see [14]) which show that the worst-case ratios of $v_{qp}^{\min}/v_{sdp}^{\min}$ are indeed $O(m^2)$ and $O(m)$ in the real and complex case, respectively, even in the absence of the indefinite constraint $x^* A_0 x \ge 1$. Thus, the bounds of Theorems 3.3 and 3.6 are essentially tight.

What happens if there is more than one indefinite quadratic constraint? The following example shows that in this case the SDP relaxation does not admit any *finite* quality bound.

Example 3.7.

$$\begin{aligned} \min \quad & x_4^2 \\ \text{s.t.} \quad & x_1x_2 + x_3^2 + x_4^2 \geq 1 \\ & -x_1x_2 + x_3^2 + x_4^2 \geq 1 \\ & \tfrac{1}{2}x_1^2 - x_3^2 \geq 1 \\ & \tfrac{1}{2}x_2^2 - x_3^2 \geq 1 \\ & x_1, x_2, x_3, x_4 \in \mathbb{R}. \end{aligned}$$

The first two constraints are equivalent to $|x_1x_2| \leq x_3^2 + x_4^2 - 1$. At the same time, the last two constraints imply $|x_1x_2| \geq 2(x_3^2 + 1)$. Combining these two inequalities yields

$$x_3^2 + x_4^2 - 1 \geq 2(x_3^2 + 1),$$

which further implies $x_4^2 \geq 3$. Therefore, we must have $v_{qp}^{\min} \geq 3$ in this case. However,

$$\begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

is feasible for the corresponding SDP relaxation problem and attains an objective value of 0. Thus, it must be optimal, and thus $v_{sdp}^{\min} = 0$. Hence, $v_{qp}^{\min}/v_{sdp}^{\min} = \infty$ in this case.

4. SDP relaxation bounds for the quadratic maximization model. In this section, we discuss the approximation bound of SDP relaxation for the nonconvex homogeneous quadratic maximization problem (1.2). Subsection 4.1 considers the case that one constraint is indefinite, and subsection 4.2 considers the case that two or more constraints are indefinite.

4.1. One indefinite constraint. In this subsection, consider the quadratic maximization problem

$$\begin{aligned} v_{qp}^{\max} := \max \quad & x^*Cx \\ \text{s.t.} \quad & x^*A_kx \leq 1,\ k = 0, 1, \ldots, m \\ & x \in \mathbb{F}^n, \end{aligned} \tag{4.1}$$

where $A_k \in \mathbb{SF}_+^n$, $k = 1, \ldots, m$, are positive semidefinite, while $C, A_0 \in \mathbb{SF}^n$ may be indefinite. For convenience, from now on we shall focus on the case $\mathbb{F} = \mathbb{R}^n$. Unlike the case of minimization form, this choice does not significantly affect the quality of SDP approximation ratios, since in the complex case the bounds are of the same order of magnitude. We assume that there is $\mu_k \geq 0$, $k = 0, 1, \ldots, m$, such that

$$\sum_{k=0}^{m} \mu_k A_k \succ 0.$$

Under this condition, the SDP relaxation satisfies the dual Slater condition. Thus the primal-dual optimal solutions exist and the primal-dual optimal objective values are attainable. Let the SDP relaxation optimal value be

$$\begin{aligned} v_{sdp}^{\max} := \max \quad & \mathrm{Tr}\,(CX) \\ \text{s.t.} \quad & \mathrm{Tr}\,(A_k X) \le 1,\ k = 0, 1, \ldots, m \\ & X \succeq 0. \end{aligned} \tag{4.2}$$

Obviously $v_{qp}^{\max} \le v_{sdp}^{\max}$.

LEMMA 4.1. *Let w_{ij} $(1 \le i < j \le n)$ be any real numbers, and let ξ_i $(1 \le i \le n)$ be random variables such that* Prob $\{\xi_i = -1\}$ = Prob $\{\xi_i = 1\} = 0.5$. *Then there holds*

$$\text{Prob}\left\{\sum_{1\le i<j\le n} w_{ij}\xi_i\xi_j \le 0\right\} > \frac{3}{100}.$$

Proof. Let $\Psi = \sum_{1\le i<j\le n} w_{ij}\xi_i\xi_j$. Then $\mathsf{E}\Psi = 0$, $\mathsf{E}(\Psi^2) = \sum_{1\le i<j\le n} w_{ij}^2$ and

$$\mathsf{E}(\Psi^4) = \sum_{1\le i<j\le n} w_{ij}^4 + 6 \sum_{1\le i<j<k\le n} (w_{ij}^2 w_{ik}^2 + w_{ij}^2 w_{jk}^2 + w_{ik}^2 w_{jk}^2) + W,$$

where

$$\begin{aligned} W = {} & 24 \sum_{1\le i<j<k<\ell\le n} (w_{ij}w_{ik}w_{j\ell}w_{k\ell} + w_{ij}w_{i\ell}w_{jk}w_{k\ell} + w_{ik}w_{i\ell}w_{jk}w_{j\ell}) \\ & + 6 \sum_{1\le i<j<k<\ell\le n} \left(w_{ij}^2 w_{k\ell}^2 + w_{ik}^2 w_{j\ell}^2 + w_{i\ell}^2 w_{jk}^2\right) \\ & \le 30 \sum_{1\le i<j<k<\ell\le n} \left(w_{ij}^2 w_{k\ell}^2 + w_{ik}^2 w_{j\ell}^2 + w_{i\ell}^2 w_{jk}^2\right). \end{aligned}$$

Therefore, we have $\mathsf{E}(\Psi^4) \le 15(\sum_{1\le i<j\le n} w_{ij}^2)^2$. Now let $\Phi = \frac{\Psi}{\sqrt{\sum_{1\le i<j\le n} w_{ij}^2}}$. Then $\mathsf{E}(\Phi) = 0$, $\mathrm{Var}(\Phi) = 1$, and $\mathsf{E}(\Phi^4) \le 15$. By Lemma 2.2, we have

$$\text{Prob}\,\{\Phi \le 0\} > \frac{3}{100}.$$

The desired result follows. □

Lemma 4.1 represents a significant advancement in settling an open question of Ben-Tal, Nemirovski, and Roos [3, Conjecture A.5] who conjectured that

$$\text{Prob}\left\{\sum_{1\le i<j\le n} w_{ij}\xi_i\xi_j \le 0\right\} \ge \frac{1}{4} \qquad \forall\ w_{ij}.$$

We managed to show a smaller constant bound of 3/100, instead of 1/4. The above inequality was needed to establish the so called *approximate S-lemma*—an extension of the well-known S-lemma, which is important in the context of robust optimization and is closely related to our analysis in this section. In their work [18], Nemirovski,

Roos, and Terlaky derived a weaker lower bound of $1/8n^2$, which goes to zero as $n \to \infty$.

We can now use Lemma 4.1 to analyze the performance of SDP relaxation for (4.2). Let $\hat{X} = UU^T$ be one optimal solution of (4.2), where $U \in \mathbb{R}^{n\times r}$ and $r = \text{rank}(\hat{X})$. Suppose $Q \in \mathbb{R}^{n\times r}$ is the orthogonal matrix such that $\hat{C} := Q^T U^T C U Q$ is diagonal. Let ξ_k, $k = 1, \ldots, r$, be i.i.d. random variables taking values -1 or 1 with equal probabilities, and let

$$x(\xi) := \frac{1}{\sqrt{\max\limits_{0\le k\le m} \xi^T \hat{A}_k \xi}}\, UQ\xi,$$

where $\hat{A}_k = Q^T U^T A_k U Q$. Note that the above random vector $x(\xi)$ is always well-defined, since the assumption $\sum_{k=0}^m \mu_k A_k \succ 0$ implies

$$\max_{0\le k\le m} \xi^T \hat{A}_k \xi > 0 \text{ for any } \xi \neq 0.$$

Let $\mu = \min\{m, \max_i \text{rank}(A_i \hat{X})\}$. We have the following estimate of the SDP approximation ratio.

THEOREM 4.2. *It holds that*

$$v_{qp}^{\max} \le v_{sdp}^{\max} \le 2\log(67\, m\mu)\, v_{qp}^{\max}.$$

Proof. Notice that $\hat{C} = Q^T U^T C U Q$ is diagonal, and hence

$$x(\xi)^T C x(\xi) = \frac{1}{\max\limits_{0\le k\le m} \xi^T \hat{A}_k \xi}\, \xi^T Q^T U^T C U Q \xi = \frac{1}{\max\limits_{0\le k\le m} \xi^T \hat{A}_k \xi} \text{Tr}\,(CX).$$

Therefore, for any $\alpha > 1$ we have

$$\begin{aligned}
&\text{Prob}\left\{x(\xi)^T C x(\xi) \ge \frac{1}{\alpha}\,\text{Tr}\,(CX)\right\} \\
&= \text{Prob}\left\{\max_{0\le k\le m} \xi^T \hat{A}_k \xi \le \alpha\right\} \\
&= 1 - \text{Prob}\left\{\max_{0\le k\le m} \xi^T \hat{A}_k \xi > \alpha\right\} \\
&\ge 1 - \text{Prob}\left\{\max_{1\le k\le m} \xi^T \hat{A}_k \xi > \alpha\right\} - \text{Prob}\left\{\xi^T \hat{A}_0 \xi > \alpha\right\}.
\end{aligned}$$

Since $\text{Tr}\,(A_0) \le 1$ and so $\alpha - \text{Tr}\,(A_0) \ge 0$, it follows from Lemma 4.1 that

$$\text{Prob}\left\{\xi^T \hat{A}_0 \xi > \alpha\right\} \le \text{Prob}\left\{\sum_{1\le i<j\le m} (\hat{A}_0)_{ij}\xi_i\xi_j > 0\right\} < 1 - \frac{3}{100}.$$

Since $\hat{A}_k \succeq 0$ for $k = 1, \ldots, m$, and $\text{Tr}\,(\hat{A}_k) \le 1$, it follows from (12) in [18] that

$$\text{Prob}\left\{\max_{1\le k\le m} \xi^T \hat{A}_k \xi > \alpha\right\} < 2m\mu e^{-\frac{1}{2}\alpha}.$$

Hence we have

$$\text{Prob}\left\{x(\xi)^T C x(\xi) \geq \frac{1}{\alpha}\,\text{Tr}\,(CX)\right\} > \frac{3}{100} - 2m\mu e^{-\frac{1}{2}\alpha}.$$

Letting $\hat{\alpha} = 2\log(67\,m\mu)$ ensures the above probability will be positive. Therefore, there exists a random vector ξ such that $\text{Tr}\,(CX) \leq \hat{\alpha}\, x(\xi)^T C x(\xi)$, and the theorem is proven. □

We point out that Theorem 4.2 is an improvement of the so-called approximate S-lemma of Ben-Tal, Nemirovski, and Roos [3, Lemma A.6]. In particular, they showed that $\alpha \leq 2\log(16n^2\,m\mu)$, in contrast to our bound $\alpha \leq 2\log(67\,m\mu)$.

Notice that in (4.1) there is only one indefinite inequality. Can we allow more than one indefinite constraint? The following example shows that the answer is "no" if we wish to have a data-independent worst-case approximation ratio. (Data-dependent approximation ratio bounds will be discussed in section 4.2, where we do allow multiple indefinite constraints.)

Example 4.3. Consider

$$\begin{aligned} \max \quad & x_1^2 + \tfrac{1}{M}x_2^2 \\ \text{s.t.} \quad & Mx_1x_2 + x_2^2 \leq 1 \\ & -Mx_1x_2 + x_2^2 \leq 1 \\ & M(x_1^2 - x_2^2) \leq 1, \end{aligned}$$

where $M > 0$ is an arbitrarily large positive constant. Its SDP relaxation is

$$\begin{aligned} \max \quad & X_{11} + \tfrac{1}{M}X_{22} \\ \text{s.t.} \quad & MX_{12} + X_{22} \leq 1,\ -MX_{12} + X_{22} \leq 1,\ M(X_{11} - X_{22}) \leq 1 \\ & \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} \succeq 0. \end{aligned}$$

For this quadratic program, the first two constraints imply that $|x_1x_2| \leq \frac{1-x_2^2}{M} \leq \frac{1}{M}$, and so $x_1^2 \leq \frac{1}{M^2x_2^2}$. The third inequality assures that $x_1^2 \leq \frac{1}{M} + x_2^2$. Therefore, $x_1^2 \leq \min\{\frac{1}{M^2x_2^2}, \frac{1}{M} + x_2^2\} \leq \frac{\sqrt{5}+1}{2M} \approx \frac{1.618}{M}$. Moreover, $x_2^2 \leq 1$, and so $v_{qp}^{\max} \leq \frac{2.618}{M}$.

The SDP relaxation satisfies both primal and dual Slater conditions, so the primal-dual optimal solutions exist. A feasible solution for the SDP relaxation (primal problem) is the 2 x 2 identity matrix, with the objective value being $1 + \frac{1}{M} > 1$. On the other hand, since $X_{22} \leq M|X_{12}| + X_{22} \leq 1$, and $X_{11} \leq X_{22} + \frac{1}{M}$, an upper bound for the SDP optimal value is $1 + \frac{2}{M}$. Therefore, for this example the ratio $\frac{v_{sdp}^{\max}}{v_{qp}^{\max}} \geq \frac{M}{2.618} \approx 0.382M$ can be arbitrarily large, depending on the size of M.

If there are at most two homogeneous quadratic constraints, and, moreover, if the SDP relaxation has a primal-dual complementary optimal solution, then the SDP optimal value will be equal to the optimal value of the quadratic model; see, e.g., Ye and Zhang [25, Corollary 2.6]. In other words, if there are no more than two inequality constraints, then under the primal-dual Slater condition, we will have $v_{sdp}^{\max}/v_{qp}^{\max} = 1$. In this sense, Example 4.3 is the smallest possible in size. By removing the requirement that the SDP relaxation has a finite optimal value, then it is possible to construct an example which involves only two inequality constraints.

Example 4.4. Consider

$$\begin{aligned} \max \quad & x_1x_2 + x_1^2 \\ \text{s.t.} \quad & x_1x_2 \le 1 \\ & x_1^2 - x_2^2 \le 1, \end{aligned}$$

with the SDP relaxation

$$\begin{aligned} \max \quad & X_{12} + X_{11} \\ \text{s.t.} \quad & X_{12} \le 1,\ X_{11} - X_{22} \le 1, \\ & \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} \succeq 0. \end{aligned}$$

In terms of polar coordinates, $(x_1, x_2) \longrightarrow (r\cos\theta, r\sin\theta)$, the original quadratic problem can be turned into

$$\begin{aligned} \max \quad & r^2(\sin 2\theta + \cos 2\theta + 1)/2 \\ \text{s.t.} \quad & r^2 \sin 2\theta \le 2 \\ & r^2 \cos 2\theta \le 1. \end{aligned}$$

By a further change of variables $(r^2\cos 2\theta, r^2 \sin 2\theta) \longrightarrow (y_1, y_2)$, we can reformulate the original quadratic problem as

$$\begin{aligned} \max \quad & \tfrac{1}{2}\left(y_1 + y_2 + \sqrt{y_1^2 + y_2^2}\right) \\ \text{s.t.} \quad & y_1 \le 2 \\ & y_2 \le 1. \end{aligned}$$

This optimization problem has a unique optimal solution at $(y_1^*, y_2^*) = (2, 1)$ with the optimal value being $\frac{3+\sqrt{5}}{2} \approx 2.618$. The SDP relaxation problem is clearly unbounded, as any positive multiple of the identity matrix is feasible. Therefore, $v_{sdp}^{\max}/v_{qp}^{\max} = +\infty$. This example is possible because the dual of the SDP relaxation problem is infeasible.

4.2. Multiple indefinite constraints. Unlike the minimization form (1.1) for which the SDP approximation ratio can be infinite when there are more than one indefinite constraints (see Example 3.7), the maximization form (1.2) can still admit a finite SDP approximation ratio in this case. In particular, consider a general homogeneous quadratic maximization problem

$$\begin{aligned} \max \quad & x^T C x \\ \text{s.t.} \quad & x^T A_k x \le 1,\ k = 0, 1, \ldots, m \\ & x \in \mathbb{F}^n. \end{aligned} \tag{4.3}$$

Suppose that $\mathcal{I}, \mathcal{D}$ are two index sets, $\mathcal{I} \cup \mathcal{D} = \{0, 1, \ldots, m\}$ and $\mathcal{I} \cap \mathcal{D} = \emptyset$, such that $A_k \succeq 0$ for $k \in \mathcal{D}$ and A_k indefinite for $k \in \mathcal{I}$. The SDP relaxation for (4.3) is

$$\begin{aligned} \max \quad & \mathrm{Tr}\,(CX) \\ \text{s.t.} \quad & \mathrm{Tr}\,(A_k X) \le 1,\ k = 0, 1, \ldots, m \\ & X \succeq 0. \end{aligned} \tag{4.4}$$

We begin our analysis with a technical lemma which bounds the probability of an exponential tail. Similar bounds exist in the literature, e.g., [6]. However, the lemma below serves our needs exactly; for completeness we include a proof here.

LEMMA 4.5. *Let $\{\lambda_i\}_{i=1}^n$ be any given real numbers, and $\{\eta_i\}_{i=1}^n$ be i.i.d. random variables drawn from either the real or complex valued zero mean Gaussian distribution with unit variance. Let $\sigma = \sqrt{\sum_{i=1}^n \lambda_i^2}$ and $\delta = \max\{\max\{\lambda_i \mid 1 \le i \le n\}, 0\}$. Then, for any $\alpha > 0$ there holds*

$$\begin{aligned}&\text{Prob}\left\{\sum_{i=1}^n \lambda_i\eta_i^2 - \sum_{i=1}^n \lambda_i \ge \alpha\sigma\right\}\\ &\quad\le \begin{cases} \exp\left(-\min\left\{\alpha, \frac{\sigma}{\delta}\right\}\frac{\alpha}{8}\right) & \text{if } \eta_i \sim N(0,1) \text{ is real Gaussian,}\\ \exp\left(-\min\left\{\alpha, \frac{\sigma}{\delta}\right\}\frac{\alpha}{4}\right) & \text{if } \eta_i \sim N_c(0,1) \text{ is complex Gaussian.}\end{cases}\end{aligned}$$

Proof. We will prove only the real Gaussian case; the complex case is similar and therefore omitted. Let $\beta := \frac{1}{4}\min\{\frac{1}{\delta}, \frac{\alpha}{\sigma}\}$. Then, $2\beta\lambda_i \le 1/2$ for all $i = 1, \ldots, n$, and $\beta\sigma = \frac{1}{4}\min\{\frac{\sigma}{\delta}, \alpha\}$. Note that for any $t \le 1/2$ the following inequality holds:

$$\frac{1}{1-t} \le e^{t+t^2}. \tag{4.5}$$

Let $\zeta := e^{\beta\sum_{i=1}^n \lambda_i\eta_i^2}$. Since $\{\eta_i^2\}_{i=1}^n$ are standard i.i.d. χ^2 random variables, it follows that

$$\begin{aligned}\mathsf{E}(\zeta) &= \prod_{i=1}^n \mathsf{E}\left(e^{\beta\lambda_i\eta_i^2}\right) = \prod_{i=1}^n \frac{1}{\sqrt{1-2\beta\lambda_i}} = \left(\prod_{i=1}^n \frac{1}{1-2\beta\lambda_i}\right)^{\frac{1}{2}}\\ &\le \left(\prod_{i=1}^n e^{2\beta\lambda_i + 4\beta^2\lambda_i^2}\right)^{\frac{1}{2}} = e^{2\beta^2\sigma^2 + \beta\sum_{i=1}^n \lambda_i},\end{aligned}$$

where the inequality is due to (4.5). This together with the Markov inequality implies

$$\begin{aligned}\text{Prob}\left\{\sum_{i=1}^n \lambda_i\eta_i^2 - \sum_{i=1}^n \lambda_i \ge \alpha\sigma\right\} &= \text{Prob}\left\{\zeta \ge e^{\beta(\alpha\sigma + \sum_{i=1}^n \lambda_i)}\right\}\\ &\le \frac{\mathsf{E}(\zeta)}{e^{\beta(\alpha\sigma + \sum_{i=1}^n \lambda_i)}}\\ &\le e^{2\beta^2\sigma^2 - \beta\sigma\alpha} = e^{\beta\sigma(2\beta\sigma - \alpha)} \le e^{\beta\sigma(\frac{\alpha}{2} - \alpha)}\\ &= e^{-\min\left\{\alpha, \frac{\sigma}{\delta}\right\}\frac{\alpha}{8}}.\end{aligned}$$

The lemma is proven. ∎

We are now ready to pursue the performance analysis for the real case $\mathbb{F} = \mathbb{R}$. Assume that (4.4) has an optimal solution $\hat{X}$. Denote the set of (real) eigenvalues of $A_k\hat{X}$ as $\lambda_1^k, \ldots, \lambda_n^k$, $k = 0, 1, \ldots, m$. Since $\text{Tr}\,(A_k\hat{X}) \le 1$, it follows that $\sum_{i=1}^n \lambda_i^k \le 1$. Moreover, $\|A_k\hat{X}\|_F^2 \ge \sum_{i=1}^n (\lambda_i^k)^2$, $k = 0, 1, \ldots, m$, where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix.

Let ξ be a random vector drawn from the Gaussian distribution $N(0, \hat{X})$. For any $\alpha > 1$ and $0 \le k \le m$, we consider the probability of the event $\text{Prob}\{\xi^T A_k\xi > \alpha\}$.

By diagonalization, we have $\text{Prob}\{\xi^T A_k \xi > \alpha\} = \text{Prob}\{\sum_{i=1}^n \lambda_i^k \eta_i^2 > \alpha\}$, where $\eta = (\eta_1, \ldots, \eta_n)^T$ is a random vector following the normal distribution $N(0, I_n)$.

If we let $\sigma^k := \sqrt{\sum_{i=1}^n (\lambda_i^k)^2} \le \|A_k \hat{X}\|_F$ and $\delta^k := \max\{0, \max\{\lambda_i^k \mid 1 \le i \le n\}\}$, then Lemma 4.5 leads to

$$\text{Prob}\{\xi^T A_k \xi > \alpha\} \tag{4.6}$$
$$\le \exp\left(-\min\left\{\frac{\alpha - \sum_{i=1}^n \lambda_i^k}{\sigma^k}, \frac{\sigma^k}{\delta^k}\right\} \frac{\alpha - \sum_{i=1}^n \lambda_i^k}{8\sigma^k}\right) \quad \forall\, 0 \le k \le m.$$

Alternatively, we can bound the tail probability using Chebyshev's inequality. In particular, since $\text{Var}(\sum_{i=1}^n \lambda_i^k \eta_i^2) = 2\sum_{i=1}^n (\lambda_i^k)^2 \le 2\|A_k \hat{X}\|_F^2$, it follows from Chebyshev's inequality that

$$\begin{aligned}
\text{Prob}\left\{\sum_{i=1}^n \lambda_i^k \eta_i^2 > \alpha\right\} &= \text{Prob}\left\{\sum_{i=1}^n \lambda_i^k \eta_i^2 - \sum_{i=1}^n \lambda_i^k > \alpha - \sum_{i=1}^n \lambda_i^k\right\} \\
&\le \text{Prob}\left\{\left|\sum_{i=1}^n \lambda_i^k \eta_i^2 - \sum_{i=1}^n \lambda_i^k\right| > \alpha - \sum_{i=1}^n \lambda_i^k\right\} \\
&\le \frac{\text{Var}(\sum_{i=1}^n \lambda_i^k \eta_i^2)}{\left(\alpha - \sum_{i=1}^n \lambda_i^k\right)^2} \le \frac{2\|A_k \hat{X}\|_F^2}{(\alpha - 1)^2} \quad \forall\, 0 \le k \le m, \qquad (4.7)
\end{aligned}$$

where we have used the fact $\alpha > 1 \ge \sum_{i=1}^n \lambda_i^k$. Applying Lemma 3.1 and using (4.6)–(4.7) gives

$$\begin{aligned}
&\text{Prob}\left\{\xi^T A_k \xi \le \alpha,\, k = 0, 1, \ldots, m;\, \xi^T C \xi \ge \text{Tr}\,(C\hat{X})\right\} \\
&\ge 1 - \text{Prob}\left\{\xi^T C \xi < \text{Tr}\,(C\hat{X})\right\} - \sum_{k=0}^m \text{Prob}\left\{\xi^T A_k \xi > \alpha\right\} \\
&\ge \frac{3}{100} - \sum_{k=0}^m \min\left\{\exp\left(-\min\left\{\frac{\alpha - \sum_{i=1}^n \lambda_i^k}{\sigma^k}, \frac{\sigma^k}{\delta^k}\right\} \frac{\alpha - \sum_{i=1}^n \lambda_i^k}{8\sigma^k}\right), \frac{2\|A_k \hat{X}\|_F^2}{(\alpha - 1)^2}\right\}.
\end{aligned}$$

Notice that $\delta^k \le \sigma^k$ and $\sum_{i=1}^n \lambda_i^k \le 1$ for any k. Therefore, we have, for any $\alpha > 1$,

$$\begin{aligned}
&\text{Prob}\left\{\xi^T A_k \xi \le \alpha,\, k = 0, 1, \ldots, m;\, \xi^T C \xi \ge \text{Tr}\,(C\hat{X})\right\} \\
&\ge \frac{3}{100} - \sum_{i \in \mathcal{D}} \exp\left(-\min\left\{\frac{\alpha - 1}{\sigma^k}, 1\right\} \frac{\alpha - 1}{8\sigma^k}\right) \\
&\quad - \sum_{i \in \mathcal{I}} \min\left\{\exp\left(-\min\left\{\frac{\alpha - 1}{\sigma^k}, 1\right\} \frac{\alpha - 1}{8\sigma^k}\right), \frac{2\|A_k \hat{X}\|_F^2}{(\alpha - 1)^2}\right\}.
\end{aligned}$$

Let us choose

$$\hat{\alpha} = 1 + \max\left\{20 + 8\log|\mathcal{D}|, \min\left\{(20 + 8\log|\mathcal{I}|) \max_{k \in \mathcal{I}} \|A_k \hat{X}\|_F, \sqrt{200 \sum_{k \in \mathcal{I}} \|A_k \hat{X}\|_F^2}\right\}\right\}.$$

Since $\sigma^k \le \sum_{i=1}^n \lambda_i^k \le 1$ for $k \in \mathcal{D}$, it follows from the choice of $\hat{\alpha}$ that

$$\exp\left(-\min\left\{\frac{\hat{\alpha}-1}{\sigma^k}, 1\right\}\frac{\hat{\alpha}-1}{8\sigma^k}\right)$$
$$= \exp\left(-\frac{\hat{\alpha}-1}{8\sigma^k}\right) \le \exp\left(-\frac{\hat{\alpha}-1}{8}\right) \le \frac{1}{100|\mathcal{D}|} \quad \forall\, k \in \mathcal{D}$$

and

$$\sum_{i\in\mathcal{I}} \min\left\{\exp\left(-\min\left\{\frac{\hat{\alpha}-1}{\sigma^k}, 1\right\}\frac{\hat{\alpha}-1}{8\sigma^k}\right), \frac{2\|A_k\hat{X}\|_F^2}{(\hat{\alpha}-1)^2}\right\} \le \frac{1}{100}.$$

This further implies that

$$\text{Prob}\left\{\xi^T A_k \xi \le \hat{\alpha},\, k = 0, 1, \ldots, m;\, \xi^T C \xi \ge \text{Tr}\,(C\hat{X})\right\} \ge \frac{1}{100}.$$

Summarizing, we obtain the following worst-case performance ratio bounds on the SDP relaxation for a real-valued homogeneous (indefinite) quadratic maximization problem. (We also state the complex case without proof.)

THEOREM 4.6. *For the quadratic optimization problem* (4.3) *with* $\mathbb{F} = \mathbb{R}$ *and its SDP relaxation* (4.4)*, suppose that an optimal solution, say* $\hat{X}$*, for* (4.4) *exists. Then,*

$$\frac{v_{sdp}^{\max}}{v_{qp}^{\max}} \le 1 + \max$$
$$\left\{20 + 8\log|\mathcal{D}|, \min\left\{(20 + 8\log|\mathcal{I}|)\max_{k\in\mathcal{I}}\|A_k\hat{X}\|_F, \sqrt{200\sum_{k\in\mathcal{I}}\|A_k\hat{X}\|_F^2}\right\}\right\}.$$

Similarly, for the complex case $\mathbb{F} = \mathbb{C}$*, we have*

$$\frac{v_{sdp}^{\max}}{v_{qp}^{\max}} \le 1 + \max$$
$$\left\{15 + 4\log|\mathcal{D}|, \min\left\{(15 + 4\log|\mathcal{I}|)\max_{k\in\mathcal{I}}\|A_k\hat{X}\|_F, \sqrt{40\sum_{k\in\mathcal{I}}\|A_k\hat{X}\|_F^2}\right\}\right\}.$$

Let us consider two special cases of Theorem 4.6. First, if $\mathcal{I} = \emptyset$, then Theorem 4.6 becomes $\frac{v_{sdp}^{\max}}{v_{qp}^{\max}} \le 20 + 8\log m$ (in the real case), which recovers the approximation result of Nemirovski, Roos, and Terlaky [18]. The second case is $\mathcal{D} = \emptyset$, where Theorem 4.6 becomes

$$\frac{v_{sdp}^{\max}}{v_{qp}^{\max}} \le 1 + \min\left\{(20 + 8\log(m+1))\max_{0\le k\le m}\|A_k\hat{X}\|_F, \sqrt{200\sum_{k=0}^{m}\|A_k\hat{X}\|_F^2}\right\}.$$

Below is an example showing that this bound is also tight (in the order of magnitude). Specifically, consider Example 4.3 again:

$$\begin{aligned} \max \quad & x_1^2 + \tfrac{1}{M}x_2^2 \\ \text{s.t.} \quad & Mx_1x_2 + x_2^2 \le 1 \\ & -Mx_1x_2 + x_2^2 \le 1 \\ & M(x_1^2 - x_2^2) \le 1. \end{aligned}$$

In this case we know that the SDP relaxation has an optimal solution $\hat{X} = \left[\begin{smallmatrix} 1+\frac{1}{M} & 0 \\ 0 & 1 \end{smallmatrix}\right]$, while the approximation ratio is $v_{sdp}^{\max}/v_{qp}^{\max} = O(M)$. There are three constraints, all indefinite, $\mathcal{I} = \{1, 2, 3\}$, with

$$A_1 = \begin{bmatrix} 0 & \frac{M}{2} \\ \frac{M}{2} & 1 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & -\frac{M}{2} \\ -\frac{M}{2} & 1 \end{bmatrix}, \quad A_3 = \begin{bmatrix} M & 0 \\ 0 & -M \end{bmatrix},$$

and so one may compute that

$$A_1\hat{X} = \begin{bmatrix} 0 & \frac{M}{2} \\ \frac{M}{2}+\frac{1}{2} & 1 \end{bmatrix}, \; A_2\hat{X} = \begin{bmatrix} 0 & -\frac{M}{2} \\ -\frac{M}{2}-\frac{1}{2} & 1 \end{bmatrix}, \; A_3\hat{X} = \begin{bmatrix} M+1 & 0 \\ 0 & -M \end{bmatrix}.$$

Thus, $\|A_k\hat{X}\|_F^2 = O(M^2)$, for $k = 1, 2, 3$. Theorem 4.6 predicts that $v_{sdp}^{\max}/v_{qp}^{\max} \leq O(M)$, and this upper bound is exactly attained in this example.

5. Discussions. This paper studies the quality bounds of SDP relaxations for some classes of nonconvex quadratic optimization problems. Our analysis reveals interesting differences in the quality bounds for the optimization models expressed in either maximization or minimization form, and for optimization variables defined over either the real or complex field. It provides a complete picture on the performance of the SDP relaxation techniques for quadratic optimization problems involving indefinite constraints.

Theoretically, the minimization model (1.1) and maximization model (1.2) are intrinsically different; they cannot be directly transformed from one to the other. In general, the feasible region of problem (1.1) can be nonconvex, unbounded, or even disconnected, while its objective function is usually assumed to be convex. In contrast, the maximization model (1.2) typically has a convex and bounded feasible region, but the nonconvexity of the objective function makes it difficult. These essential differences have led to the qualitatively different behaviors in the respective SDP approximation ratios.

It is equally interesting to note that the choice of field in which the optimization variables reside can also impact the quality of SDP relaxation. In a natural way, every complex quadratic program can be turned into an equivalent real quadratic program by doubling the dimension. Such a transformation will not affect the resulting approximation ratio. Since the SDP approximation ratio is weaker in the real case, we cannot derive the desired approximation ratio for the complex case by this simple reduction. It is worth noting that the tighter SDP approximation ratio for the complex case has been observed in digital communication applications [22, 17, 14], where the signals are naturally complex due to their in-phase and quadrature components.

An interesting byproduct of our work is a universal lower bound of Prob($\sum_{i=1}^n \tau_i (\eta_i - 1) \geq 0$) for the i.i.d. exponential random variables η_i (Lemma 4.1). The lower bounds of this type are interesting on their own and are related to the well-known inequality of Grünbaum [9] in convex analysis. In particular, by a different analytic argument, it is possible to further improve the universal lower bound obtained in this paper as follows:

$$\text{Prob}\left(\sum_{i=1}^n \tau_i(\eta_i - 1) \geq 0\right) = \sum_{i=1}^n \frac{e^{-\frac{1}{\tau_i}}}{\prod_{j\neq i}\left(1 - \frac{\tau_j}{\tau_i}\right)} > \frac{1}{e}, \tag{5.1}$$

where τ_i, $i = 1, \ldots, n$, are any real numbers. (The above equality can be derived by evaluating a multidimensional integral.) The inequality (5.1) admits a simple

geometric interpretation. For the joint standard exponential distribution on $\mathbb{R}^n_+$, the center of gravity of $\mathbb{R}^n_+$ is $x^c := \mathsf{E}(\eta) = (1, 1, \ldots, 1)^T$, and the inequality (5.1) can be interpreted as follows:

$$\text{Prob}(\mathbb{R}^n_+ \cap \mathcal{H}_+) \geq e^{-1} \quad \text{for any hyperplane } \mathcal{H} \text{ passing through } x^c. \tag{5.2}$$

Here $\mathcal{H}_+$ denotes the positive side of the hyperplane $\mathcal{H}$. The inequality (5.2) is an extension of the Grünbaum inequality [9]:

$$\text{Volume}\,(\mathcal{C} \cap \mathcal{H}_+) \geq e^{-1}\ \text{Volume}\,(\mathcal{C})$$

for any bounded convex set $\mathcal{C}$ in $\mathbb{R}^n$, and for any hyperplane $\mathcal{H}$ passing through the center of gravity of $\mathcal{C}$

$$x^c = \frac{1}{\text{Volume}(\mathcal{C})} \int_{\mathcal{C}} dx.$$

In particular, if we assign the uniform distribution to $\mathcal{C}$, then the mean vector of this distribution is given by the center of gravity x^c, and the probability in (5.2) can be expressed in terms of volume. In this way, Grünbaum's inequality can be equivalently written as (5.2). This shows that the inequality (5.2) generalizes Grünbaum's theorem [9] from the uniform distribution over a compact, convex set to the exponential distribution over $\mathbb{R}^n_+$. Interestingly, it is possible to establish the inequality (5.2) for any log-concave distributions defined over any (possibly unbounded) convex set in $\mathbb{R}^n$. The proof of this inequality relies on a result of Bobkov [5, Lemma 3.3] and a result of Prekopa [21] on the projection of any log-concave distribution. We plan to report the details of this proof in the future.

Acknowledgment. The authors wish to thank Yuval Peres for suggesting reference [13] to us.

REFERENCES

[1] S. V. Amari and R. B. Misra, *Closed-form expressions for distribution of sum of exponential random variables*, IEEE Trans. Reliability, 46 (1997), pp. 519–522.

[2] A. Beck and M. Teboulle, *A convex optimization approach for minimizing the ratio of indefinite quadratic functions over an ellipsoid*, Math. Programming, to appear.

[3] A. Ben-Tal, A. Nemirovski, and C. Roos, *Robust solutions of uncertain quadratic and conic-quadratic problems*, SIAM J. Optim., 13 (2002), pp. 535–560.

[4] D. Bertsimas and Y. Ye, *Semidefinite relaxations, multivariate normal distributions, and order statistics*, in Handbook of Combinatorial Optimization, Vol. 3, D.-Z. Du and P. M. Pardalos, eds., Kluwer Academic Publishers, Boston, 1998, pp. 1–19.

[5] S. G. Bobkov, *On concentration of distributions of random weighted sums*, Ann. Probab., 31 (2003), pp. 195–215.

[6] S. G. Bobkov and M. Ledoux, *Poincaré's inequalities and Talagrand's concentration phenomenon for the exponential distribution*, Probab. Theory Related Fields, 107 (1997), pp. 383–400.

[7] D. R. Cox, *Renewal Theory*, John Wiley & Sons, New York, 1962.

[8] M. Goemans and D. Williamson, *Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming*, J. Assoc. Comput. Math., 42 (1995), pp. 1115–1145.

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

[10] J.-B. Hiriart-Urruty, *Global optimality conditions in maximizing a convex quadratic function under convex quadratic constraints*, J. Global Optim., 21 (2001), pp. 445–455.

[11] Y. Huang and S. Zhang, *Complex matrix decomposition and quadratic programming*, Math. Oper. Res., 32 (2007), pp. 758–768.

[12] V. Jeyakumar, A. M. Rubinov, and Z. Y. Wu, *Sufficient global optimality conditions for non-convex quadratic minimization problems with box constraints*, J. Global Optim., 36 (2006), pp. 471–481.

[13] H. Kläver and N. Schmitz, *An inequality for the asymmetry of distributions and a Berry-Esseen theorem for random summation*, J. Ineq. Pure Appl. Math., Vol. 7, No. 1, Article 2, 2006.

[14] Z.-Q. Luo, N. D. Sidiropoulos, P. Tseng, and S. Zhang, *Approximation bounds for quadratic optimization with homogeneous quadratic constraints*, SIAM J. Optim., 18 (2007), pp. 1–28.

[15] K. Madsen, H. B. Nielsen, and M. C. Pinar, *A finite continuation algorithm for bound constrained quadratic programming*, SIAM J. Optim., 9 (1999), pp. 62–83.

[16] K. Madsen, H. B. Nielsen, and M. C. Pinar, *Bound constrained quadratic programming via piecewise quadratic functions*, Math. Program., 85 (1999), pp. 135–156.

[17] E. Matskani, N. D. Sidiropoulos, Z.-Q. Luo, and L. Tassiulas, *Convex approximation techniques for joint multiuser downlink beamforming and admission control*, IEEE Trans. Wireless Communication, to appear.

[18] A. Nemirovski, C. Roos, and T. Terlaky, *On maximization of quadratic form over intersection of ellipsoids with common center*, Math. Program., 86 (1999), pp. 463–473.

[19] Yu. Nesterov, *Semidefinite relaxation and nonconvex quadratic optimization*, Optim. Methods Softw., 9 (1998), pp. 141–160.

[20] G. Pataki, *On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues*, Math. Oper. Res., 23 (1998), pp. 339–358.

[21] A. Prekopa, *On logarithmic concave measures and functions*, Acta Sci. Math. (Szeged), 34 (1973), pp. 335–343.

[22] N. D. Sidiropoulos, T. N. Davidson, and Z.-Q. Luo, *Transmit beamforming for physical layer multicasting*, IEEE Trans. Signal Process., 54 (2006), pp. 2239–2251.

[23] A. So, Y. Ye, and J. Zhang, *A Unified Theorem on SDP Rank Reduction*, working paper, 2006; available online from http://www.stanford.edu/∼yyye/.

[24] Y. Ye, *Approximating quadratic programming with bound and quadratic constraints*, Math. Program., 84 (1999), pp. 219–226.

[25] Y. Ye and S. Zhang, *New results on quadratic minimization*, SIAM J. Optim., 14 (2003), pp. 245–267.

[26] S. Zhang and Y. Huang, *Complex quadratic optimization and semidefinite programming*, SIAM J. Optim., 16 (2006), pp. 871–890.

SIAM J. OPTIM.
Vol. 19, No. 2, pp. 524–551

ON RATES OF CONVERGENCE FOR STOCHASTIC OPTIMIZATION PROBLEMS UNDER NON–INDEPENDENT AND IDENTICALLY DISTRIBUTED SAMPLING*

TITO HOMEM-DE-MELLO†

Abstract. In this paper we discuss the issue of solving stochastic optimization problems by means of sample average approximations. Our focus is on rates of convergence of estimators of optimal solutions and optimal values with respect to the sample size. This is a well-studied problem in case the samples are independent and identically distributed (i.e., when standard Monte Carlo simulation is used); here we study the case where that assumption is dropped. Broadly speaking, our results show that, under appropriate assumptions, the rates of convergence for *pointwise estimators* under a sampling scheme carry over to the optimization case, in the sense that convergence of approximating optimal solutions and optimal values to their true counterparts has the same rates as in pointwise estimation. We apply our results to two well-established sampling schemes, namely, Latin hypercube sampling and randomized quasi-Monte Carlo (QMC). The novelty of our work arises from the fact that, while there has been some work on the use of variance reduction techniques and QMC methods in stochastic optimization, none of the existing work—to the best of our knowledge—has provided a theoretical study on the effect of these techniques on rates of convergence for the optimization problem. We present numerical results for some two-stage stochastic programs from the literature to illustrate the discussed ideas.

Key words. stochastic optimization, two-stage stochastic programming with recourse, sample average approximation, Monte Carlo simulation, quasi-Monte Carlo methods, Latin hypercube sampling, variance reduction techniques

AMS subject classifications. 90C15, 65C05

DOI. 10.1137/060657418

1. Introduction.

In this paper we consider stochastic optimization problems of the form

$$\min_{x\in X} \{g(x) := \mathbb{E}[G(x,\boldsymbol{\xi})]\}, \tag{1.1}$$

where X is a subset of $\mathbb{R}^n$, $\boldsymbol{\xi}$ is a random vector in $\mathbb{R}^s$, and $G : \mathbb{R}^n \times \mathbb{R}^s \rightarrowtail \mathbb{R}$ is a real-valued measurable function. We refer to (1.1) as the "true" optimization problem. The class of problems falling into the framework of (1.1) is quite large and includes two-stage stochastic programs as a particular subclass.

Oftentimes the expectation in (1.1) cannot be calculated exactly, particularly when G does not have a closed form. In those cases, approximations based on sampling are usually the alternative. One such approximation can be constructed as follows. Consider a family $\{\hat{g}_N(\cdot)\}$ of random approximations of the function $g(\cdot)$, each $\hat{g}_N(\cdot)$ being defined as

$$\hat{g}_N(x) := \frac{1}{N}\sum_{j=1}^{N} G(x,\boldsymbol{\xi}^j), \tag{1.2}$$

where $\{\boldsymbol{\xi}^1,\ldots,\boldsymbol{\xi}^N\}$ is a sample from the distribution of $\boldsymbol{\xi}$. When $\boldsymbol{\xi}^1,\ldots,\boldsymbol{\xi}^N$—viewed as random variables—are independent and identically distributed (i.i.d.) the quantity $\hat{g}_N(x)$ is called a *(standard) Monte Carlo* estimator of $g(x)$.

*Received by the editors April 13, 2006; accepted for publication (in revised form) January 31, 2008; published electronically June 11, 2008.
http://www.siam.org/journals/siopt/19-2/65741.html
†Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208-3119 (tito@northwestern.edu).

Given the family of estimators $\{\hat{g}_N(\cdot)\}$ defined in (1.2), one can construct the corresponding approximating program

$$\min_{x \in X} \hat{g}_N(x). \tag{1.3}$$

Let $\hat{x}_N$ and $\hat{\nu}_N$ denote, respectively, an optimal solution and the optimal value of (1.3). Then $\hat{x}_N$ and $\hat{\nu}_N$ provide approximations, respectively, to an optimal solution x^* and the optimal value ν^* of the true problem (1.1). Note that the optimization in (1.3) is performed for a fixed sample; for that reason, this is called an *external sampling* approach. When $\hat{g}_N(\cdot)$ is a standard Monte Carlo estimator of $g(\cdot)$, such an approach is found in the literature under the names of sample average approximation method, stochastic counterpart, and sample-path optimization, among others.

The external sampling approach with standard Monte Carlo simulation has been implemented in various settings; see, for instance, [14, 23, 43]. One advantage of that approach lies in its nice convergence properties; for example, it is possible to show that, when x^* is the unique optimal solution, $\hat{x}_N \to x^*$ and $\hat{\nu}_N \to \nu^*$ under fairly general assumptions (see, e.g., [10, 22, 44, 48, 49]). Two properties have proven particulary useful in terms of establishing *rates of convergence*: The first establishes that, under proper conditions, $P(|g(\hat{x}_N) - g(x^*)| \leq \varepsilon)$ and $P(\|\hat{x}_N - x^*\| \leq \varepsilon)$ converge to one *exponentially fast* in the sample size N for any fixed $\varepsilon > 0$ (see [6, 21]). Under some further conditions one can say more, namely, that $P(\hat{x}_N = x^*)$ converges to one exponentially fast in the sample size N [53]. Exponential rates of convergence have interesting consequences in terms of complexity of the underlying problems; see [51] for a discussion.

Another useful property establishes that the sequence of optimal values $\{\hat{\nu}_N\}$ satisfies a certain kind of *central limit theorem* (CLT). More specifically, one has

$$N^{1/2}(\hat{\nu}_N - \nu^*) \xrightarrow{d} \text{Normal}(0, \sigma^*),$$

where "$\xrightarrow{d}$" denotes convergence in distribution and $\sigma^* := \text{Var}[G(x^*)]$ [48]. An immediate conclusion from the above result is that the rate of convergence of optimal values of (1.3) is of order $N^{-1/2}$. A compilation of these and other related results can be found in [50].

It is no surprise that the sequence of approximating optimal values converges at rate $N^{-1/2}$. Indeed, consider the estimator $\hat{g}_N$ defined in (1.2), and fix $x \in X$. Under mild conditions, it follows from the CLT that $\sqrt{N}[\hat{g}_N(x) - g(x)]/\sigma(x)$ converges in distribution to the standard Normal, where $\sigma^2(x)$ is the variance of $G(x)$. This implies that the error $\hat{g}_N(x) - g(x)$ converges to zero at the rate $N^{-1/2}$. That is, even the pointwise estimators converge at rate $N^{-1/2}$. In many practical cases, the value of N necessary to obtain a reasonably small error under this scheme becomes prohibitively large, especially if evaluation of $G(x, \boldsymbol{\xi})$ for a given $\boldsymbol{\xi}$ is computationally expensive. This motivates the use of *variance reduction techniques* that can yield estimators with smaller variance than the ones obtained with standard sampling. Consequently, the same error can be obtained with less computational effort, which is a crucial step for the use of sampling-based methods in large-scale problems.

Several variance reduction techniques have been developed in the simulation and statistics literature, notably importance sampling, control variates, stratified sampling, and others (see, e.g., [11, 25]). However, incorporation of these techniques into a stochastic optimization algorithm is still at an early stage. Existing work [1, 7,

16, 19, 26, 52] already shows that significant benefits can be gained by implementing some of these methods, but these papers provide only *empirical* evidence of the gain.

Another approach to obtain better pointwise estimators is to choose the sample points in an appropriate manner. Such is the case of *quasi-Monte Carlo methods* (QMC); see [32] for a comprehensive discussion. This class of methods has been gaining popularity in the past few years, as it has been observed that these techniques can provide rates of convergence for pointwise estimators superior to the $N^{-1/2}$ obtained with standard Monte Carlo methods. However, because estimating the actual error of a QMC estimator relative to the quantity being estimated can be difficult, some procedures to randomize a QMC sequence have been proposed in the literature. We provide a brief review of the basic ideas of QMC methods in section 3.2.

A few papers study the optimization problem $\min_{x\in X} \hat{g}_N(x)$ under QMC: In [20], empirical results are provided for the use of Hammersley sequences (one form of QMC) in stochastic optimization problems. In [39, 40], the authors use the fact that the empirical measure defined by a QMC sequence converges weakly to the uniform distribution to show that, under mild assumptions, the estimator function $\hat{g}_N$ constructed with QMC points *epiconverges* to the true function g, which guarantees convergence of optimal values and optimal solutions under appropriate further conditions; in [24] those results are applied to the case where the QMC sequence is randomized with the so-called Cranley–Patterson procedure. The numerical results in those papers also suggest considerable gains in terms of rates of convergence when using QMC methods. A different type of point sequences is studied in [41] whereby the sampling points are chosen in a way to minimize the Wasserstein distance between the original distribution and the empirical distribution generated by the points. A related approach is used in [15], which deals with the Fortet–Mourier metrics instead. Again, the numerical results presented in [42] suggest a considerable advantage of these techniques over standard Monte Carlo methods.

The above discussion shows that, while there has been some important work on the use of variance reduction techniques and QMC methods in stochastic optimization, none of these papers has provided a theoretical study on the effect of these techniques on *rates of convergence*. The reason is that, without the i.i.d. assumption, many of the classical results in probability theory cannot be applied. One exception is the work in [6], which provides results on the exponential rate of convergence of optimal solutions even without the i.i.d. assumption. However, that paper does not focus on any particular sampling technique; rather, they assume that certain conditions that allow for the application of the Gartner–Ellis theorem in large deviations theory (see, e.g., [8]) are satisfied.

Another potential way to study convergence rates in general settings (i.e., without the i.i.d. assumption) is by means of *stability* theory. Broadly speaking, stability theory in the context of stochastic optimization quantifies how much the optimal value and the optimal solutions of the problem change when the underlying probability measures are perturbed. For example, by writing the optimal values of (1.1) and (1.3), respectively, as $\nu(F)$ and $\nu(F_N)$—where F is the distribution of $\boldsymbol{\xi}$ and F_N is the empirical distribution defined by a sample—it is possible to show that, under certain assumptions, $|\nu(F) - \nu(F_N)|$ is bounded by an appropriately defined distance between F and F_N. In the particular case of i.i.d. sampling, one can write the latter distance in terms of the sample size N, which leads to a different way to view the $N^{-1/2}$ rate obtained via the CLT. We refer to [45] for a thorough exposition of stability results in stochastic programming.

In this paper we propose a study of rates of convergence for optimal solutions and optimal values of the approximating problem (1.3) *without imposing that the sample be independent or identically distributed*. More specifically, we show that, under certain conditions, if the proposed sampling scheme yields an exponential rate of convergence for pointwise estimators, then the convergence of optimal solutions will also have an exponential rate. Moreover, in case of discrete or piecewise linear problems, if the proposed sampling scheme yields a CLT for pointwise estimators, then the convergence of optimal values will obey the CLT as well. Unless stated otherwise the setting is fairly general—i.e., the decision space can be continuous or discrete, and the distributions of the underlying random variables can be continuous or discrete, although some of the results will not be valid in some of these cases.

We illustrate the ideas for the particular cases of Latin hypercube sampling (LHS) and a specific variation of randomized QMC called scrambled (t, m, s)-nets. We show that, for a particular class of functions, the exponential feature of the rate of convergence is preserved under LHS for pointwise estimators and therefore for estimators of optimal solutions. We also use CLT-type results available for LHS and randomized QMC to illustrate the convergence results for estimators of optimal values. In particular, we show that, under LHS, the estimators $\hat{\nu}_N$ of optimal values converge at a rate of order $N^{-1/2}$, the same as standard Monte Carlo methods; for QMC, under appropriate assumptions the sequence $\{\hat{\nu}_N\}$ converges at a rate of order $[(\log N)^{s-1}/N^3]^{1/2}$, which asymptotically is better than $N^{-1/2}$.

We then apply our results to two-stage stochastic linear programs and discuss the validity of our assumptions in that context. Numerical results are presented for two problems from the literature to illustrate the ideas.

The remainder of the paper is organized as follows: In section 2 we describe our main results for rates of convergence of estimators of optimal solutions and optimal values. In section 3 we apply these results to LHS and randomized QMC. We illustrate the ideas for two-stage stochastic programs in section 4 and present numerical results in section 5. Concluding remarks are presented in section 6.

2. Rates of convergence. We discuss separately the results on rates of convergence for optimal solutions and optimal values. Throughout this paper, S^* and S_N denote the set of optimal solutions of (1.1) and (1.3), respectively. Before we study the two cases, we shall make a general assumption.

Assumption A1. For each $x \in X$, $\hat{g}_N(x) \to g(x)$ with probability one (denoted w.p.1).

Assumption A1 is very natural, as it requires the estimators to be *consistent*. In the i.i.d. case, this is just the standard strong law of large numbers, which holds if $\mathbb{E}[|\hat{g}_N(x)|] < \infty$ for each $x \in X$.

2.1. Convergence of estimators of optimal solutions. We start by making the following probabilistic assumption on the estimators $\{\hat{g}_N(x)\}$.

Assumption B1. For each $x \in X$, there exist a number $C_x > 0$ and a function $\gamma_x(\cdot)$ such that $\gamma_x(0) = 0$, $\gamma_x(z) > 0$ if $z > 0$, and

$$P\left(|\hat{g}_N(x) - g(x)| \geq \delta\right) \leq C_x\, e^{-N\gamma_x(\delta)} \qquad \text{for all } N \geq 1 \text{ and all } \delta > 0. \tag{2.1}$$

That is, the probability that the deviation between $\hat{g}_N(x)$ and $g(x)$ is bigger than δ goes to zero exponentially fast with N. Notice that (2.1) implies that $\hat{g}_N(x)$ converges in probability to $g(x)$, which is also ensured by Assumption A1.

Instead of (2.1), we can impose the following weaker condition.

Assumption B1′. For each $x \in X$, there exists a function $\gamma_x(\cdot)$ such that $\gamma_x(0) = 0$, $\gamma_x(z) > 0$ if $z > 0$, and

$$\limsup_{N\to\infty} \frac{1}{N} \log P\left(|\hat{g}_N(x) - g(x)| \geq \delta\right) \leq -\gamma_x(\delta) \qquad \text{for all } \delta > 0. \tag{2.2}$$

Some of our results will be stated by assuming that B1 holds; alternatively, B1′ can be used, though in such cases the corresponding result will be stated in asymptotic form as well.

We study now a sufficient condition for Assumption B1 to hold. The main concept behind it arises from the theory of *large deviations*, a well-studied field. For a thorough exposition of the theory, we refer to any of the classical texts in the area, e.g., [8]. We present here a result from [9].

PROPOSITION 2.1. *Consider the sample $\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^N$ used in* (1.2)*, and define the extended real-valued function*

$$\phi_N(x,t) := \frac{1}{N} \log \mathbb{E}\left[e^{tN\hat{g}_N(x)}\right]. \tag{2.3}$$

Suppose that for each $x \in X$ there exists an extended real-valued function ϕ_x^* such that $\phi_N(x,\cdot) \leq \phi_x^*(\cdot)$ for all N, and assume that ϕ_x^* satisfies the following conditions: (i) $\phi_x^*(0) = 0$*;* (ii) $\phi_x^*(\cdot)$ *is continuously differentiable and strictly convex on a neighborhood of zero; and* (iii) $(\phi_x^*)'(0) = g(x)$*. Then, Assumption* B1 *holds, with the constants C_x all equal to 2 and the functions $\gamma_x(\cdot)$ given by $\gamma_x(\delta) := \min\{I_x(g(x)+\delta), I_x(g(x)-\delta)\}$, where $I_x(z) = \sup_{t\in\mathbb{R}}\{tz - \phi_x^*(t)\}$.*

A simple setting where the conditions of Proposition 2.1 are satisfied is when the functions $\phi_N(x,\cdot)$, $N = 1, 2, \ldots$, are bounded by the log-moment-generating function of some random variable W_x (i.e., $\phi_x^*(t) = \log \mathbb{E}[e^{tW_x}]$) such that $\mathbb{E}[W_x] = g(x)$. Clearly, condition (i) holds in that case. Moreover, if there exists an open neighborhood $\mathcal{N}$ of zero such that $\phi_x^*(\cdot)$ is finite on $\mathcal{N}$, then it is well known that ϕ_x^* is infinitely differentiable on $\mathcal{N}$ (see, e.g., p. 35 of [8]) and (iii) holds. In that case, Proposition 1 in [54] ensures that ϕ_x^* is strictly convex on $\mathcal{N}$.

Note that when the samples $\{\boldsymbol{\xi}^i\}$ are i.i.d. we have

$$\begin{aligned}\phi_N(x,t) &= \frac{1}{N}\log(\mathbb{E}[e^{tN\hat{g}_N(x)}]) = \frac{1}{N}\log(\{\mathbb{E}[e^{tG(x,\xi)}]\}^N) = \log(\mathbb{E}[e^{tG(x,\xi)}]) \\ &= \log M_x(t),\end{aligned}$$

where $M_x(t) := \mathbb{E}[e^{tG(x,\boldsymbol{\xi})}]$ is the moment-generating function of $G(x,\boldsymbol{\xi})$ evaluated at t. Hence, in that case we have $\phi_N(x,t) = \phi_x^*(t) := \log M_x(t)$ for all N, so the resulting function I_x in Proposition 2.1 is the rate function associated with $G(x,\boldsymbol{\xi})$. Inequality (2.1) then yields the well-known Chernoff upper bounds on the deviation probabilities. It is also well known (Cramér's theorem) that in that case $\gamma_x(\delta)$ in Proposition 2.1 is an *asymptotically exact* rate, in the sense that (2.2) holds with equality.

One important consequence of the above developments is the following: Suppose that the function ϕ_x^* in Proposition 2.1 is dominated by the log-moment-generating function of the random variable $G(x,\boldsymbol{\xi})$, i.e., $\phi_x^*(t) \leq \phi_x^{\mathrm{MC}}(t) := \log \mathbb{E}[e^{tG(x,\boldsymbol{\xi})}]$. This immediately implies that the rate function I_x dominates the rate function associated with the random variable $G(x,\boldsymbol{\xi})$, which as seen earlier is the asymptotically exact rate function obtained with i.i.d. (i.e., Monte Carlo) sampling. In other words, if one

uses a sampling technique that yields functions $\phi_N(x,\cdot)$ for which one can find ϕ_x^* in Proposition 2.1 such that $\phi_x^*(\cdot) \le \phi_x^{\mathrm{MC}}(\cdot)$, then *the pointwise convergence rate for this sampling technique—in the sense of* (2.1)*—is at least as good as the rate obtained with standard Monte Carlo methods.* We will use this basic argument repeatedly in the course of this paper.

In the subsections below we will study the convergence of optimal solutions in two different settings—one when the function $G(\cdot,\boldsymbol{\xi})$ is Lipschitz and the other when either $G(\cdot,\boldsymbol{\xi})$ is piecewise linear or the feasible set X is finite.

2.1.1. The Lipschitz case. We now make an assumption on the integrand G viewed as a function of its first argument.

Assumption A2. The feasibility set X is compact, and there exists an integrable function $L : \mathbb{R}^s \rightarrowtail \mathbb{R}$ such that, for almost every $\boldsymbol{\xi}$ and all $x, y \in X$,

$$|G(x,\boldsymbol{\xi}) - G(y,\boldsymbol{\xi})| \le L(\boldsymbol{\xi})\|x-y\|. \tag{2.4}$$

Clearly, Assumption A2 ensures that the function $G(\cdot,\boldsymbol{\xi})$ is continuous for almost every $\boldsymbol{\xi}$. Moreover, it implies that $\hat{g}_N(\cdot)$ and $g(\cdot)$ are also Lipschitz continuous with constants equal to $\hat{L}_N := N^{-1}\sum_{j=1}^N L(\boldsymbol{\xi}^j)$ and $\mathbb{E}[L(\boldsymbol{\xi})]$, respectively. From classical results in convex analysis (e.g., [18, Theorem IV.3.1.2]), we see that if (i) the feasibility set X is compact and contained in the relative interior of the domain of $G(\cdot,\boldsymbol{\xi})$ for almost every $\boldsymbol{\xi}$, and (ii) $G(\cdot,\boldsymbol{\xi})$ is *convex* for almost every $\boldsymbol{\xi}$, then the existence of $L(\boldsymbol{\xi})$ in (2.4) is assured, so in that case only integrability of $L(\boldsymbol{\xi})$ needs to be checked.

Recall that $\hat{x}_N$ is an optimal solution of (1.3) and S^* is the set of optimal solutions of (1.1). Below, $\mathrm{dist}(z,A)$ denotes the usual Euclidean distance function between a point z and a set A, i.e., $\mathrm{dist}(z,A) := \inf_{y\in A}\|z-y\|$. The following result is known (see, e.g., [46, pp. 67–70]), but we state it here for reference.

PROPOSITION 2.2. *Suppose that Assumptions* A1 *and* A2 *hold. Then*

(i) $\hat{g}_N(x) \to g(x)$ *uniformly on* X *w.p.*1*;*

(ii) $\hat{\nu}_N \to \nu^*$ *w.p.*1*;*

(iii) $\mathrm{dist}(\hat{x}_N, S^*) \to 0$ *w.p.*1.

Theorem 2.3 below shows a probabilistic rate of convergence of optimal solutions. In preparation for that result, we state the following assumption, which is similar to Assumption B1 but applied to the random variable $L(\boldsymbol{\xi})$ in Assumption A2. Conditions under which such an assumption holds are similar to those given in Proposition 2.1.

Assumption $\mathrm{B1}_L$. Let $\hat{L}_N$ be the estimator of $\mathbb{E}[L(\boldsymbol{\xi})]$ defined as $\hat{L}_N := N^{-1}\sum_{j=1}^N L(\boldsymbol{\xi}^j)$, where as before $\{\boldsymbol{\xi}^1,\dots,\boldsymbol{\xi}^N\}$ is a sample from the distribution of $\boldsymbol{\xi}$. There exist a number $C_L > 0$ and a function $\gamma_L(\cdot)$ such that $\gamma_L(0) = 0$, $\gamma_L(z) > 0$ if $z > 0$, and

$$P\left(|\hat{L}_N - \mathbb{E}[L(\boldsymbol{\xi})]| \ge \delta\right) \le C_L\, e^{-N\gamma_L(\delta)} \qquad \text{for all } N \ge 1 \text{ and all } \delta > 0. \tag{2.5}$$

THEOREM 2.3. *Consider problem* (1.3)*, and suppose that Assumptions* A2*,* B1*, and* $\mathrm{B1}_L$ *hold. Then, given* $\varepsilon > 0$*, there exist constants* $K > 0$ *and* $\alpha > 0$ *such that*

$$P\left(\mathrm{dist}(\hat{x}_N, S^*) \ge \varepsilon\right) \le Ke^{-\alpha N} \qquad \textit{for all } N \ge 1.$$

The constants K *and* α *depend on the families of estimators* $\{\hat{g}_N(\cdot)\}$ *and* $\{\hat{L}_N\}$ *only through, respectively, the constants* C_x *and* C_L *and the exponent functions* $\gamma_x(\cdot)$ *and*

$\gamma_L(\cdot)$ *in* (2.1) *and* (2.5). *More specifically,*

$$\alpha = \min\left(\min_{k=1,\ldots,r}\{\gamma_{x_k}(\delta/3)\},\ \gamma_L(\delta/3)\right),$$
$$K = (r+1)\max\left(\max_{k=1,\ldots,r}\{C_{x_k}\},\ C_L\right),$$

where $\delta > 0$, r *is a finite number, and* $x_1, \ldots, x_r$ *are points in* X.

The proof of Theorem 2.3 will be based on the following lemma.

LEMMA 2.4. *Suppose that Assumptions* A2, B1, *and* B1$_L$ *hold. Then, for any* $\delta > 0$, *there exist positive constants* $A = A(\delta)$ *and* $\alpha = \alpha(\delta)$ *such that*

$$P\left(|\hat{g}_N(x) - g(x)| \geq \delta\right) \leq Ae^{-\alpha N} \qquad \text{for all } x \in X \text{ and all } N \geq 1. \tag{2.6}$$

Moreover, there exists a positive constant K *(also dependent on* δ*) such that*

$$P\left(|\hat{g}_N(x) - g(x)| < \delta \text{ for all } x \in X\right) \geq 1 - Ke^{-\alpha N} \qquad \text{for all } N \geq 1. \tag{2.7}$$

Proof. Let $\eta := \delta/(3\mathbb{E}[L(\boldsymbol{\xi})] + \delta)$, and denote by $B(x, \eta)$ the open ball with center x and radius η. Let $\mathcal{X} = \{x_1, \ldots, x_r\}$ be a collection of points in X such that $X \subset \cup_{k=1}^r B(x_k, \eta)$. Notice that the existence of $\mathcal{X}$ is ensured by the compactness of X.

Consider now an arbitrary point $x \in X$. By construction, there exists some $x_k \in \mathcal{X}$ such that $\|x - x_k\| < \eta$. Thus, from (2.4) we have

$$|\hat{g}_N(x) - \hat{g}_N(x_k)| \leq \frac{1}{N}\sum_{j=1}^{N}\left|G(x, \boldsymbol{\xi}^j) - G(x_k, \boldsymbol{\xi}^j)\right| < \qquad \hat{L}_N\eta = \frac{\delta}{3}\,\frac{\hat{L}_N}{\mathbb{E}[L(\boldsymbol{\xi})] + \delta/3}, \tag{2.8}$$

$$|g(x) - g(x_k)| \leq \quad \mathbb{E}\left[|G(x, \boldsymbol{\xi}) - G(x_k, \boldsymbol{\xi})|\right] \quad < \mathbb{E}[L(\boldsymbol{\xi})]\eta < \ \delta/3. \tag{2.9}$$

Moreover, by Assumptions B1 and B1$_L$ we have

$$P\left(|\hat{g}_N(x_k) - g(x_k)| \geq \delta/3\right) \leq C_{x_k}\, e^{-N\gamma_{x_k}(\delta/3)}, \tag{2.10}$$
$$P(|\hat{L}_N - \mathbb{E}[L(\boldsymbol{\xi})]| \geq \delta/3) \leq C_L\, e^{-N\gamma_L(\delta/3)}. \tag{2.11}$$

Finally, since

$$|\hat{g}_N(x) - g(x)| \leq |\hat{g}_N(x) - \hat{g}_N(x_k)| + |\hat{g}_N(x_k) - g(x_k)| + |g(x) - g(x_k)|,$$

it follows that

$$\begin{aligned}\{|\hat{g}_N(x) - g(x)| < \delta\} \ &\supseteq\ \{|\hat{g}_N(x) - \hat{g}_N(x_k)| < \delta/3\} \cap \{|\hat{g}_N(x_k) - g(x_k)| < \delta/3\} \\ &\quad \cap \{|g(x_k) - g(x)| < \delta/3\} \\ &\supseteq \{|\hat{L}_N - \mathbb{E}[L(\boldsymbol{\xi})]| < \delta/3\} \cap \{|\hat{g}_N(x_k) - g(x_k)| < \delta/3\},\end{aligned} \tag{2.12}$$

and then from (2.10)–(2.11) we have

$$\begin{aligned}P\left(|\hat{g}_N(x) - g(x)| \geq \delta\right) &\leq P\left(|\hat{g}_N(x_k) - g(x_k)| \geq \delta/3\right) + P(|\hat{L}_N - \mathbb{E}[L(\boldsymbol{\xi})]| \geq \delta/3) \\ &\leq C_{x_k}\, e^{-N\gamma_{x_k}(\delta/3)} + C_L\, e^{-N\gamma_L(\delta/3)}.\end{aligned}$$

By taking

$$\alpha := \min\left(\min_{k=1,\ldots,r}\{\gamma_{x_k}(\delta/3)\},\ \gamma_L(\delta/3)\right), \tag{2.13}$$

$$A := 2\max\left(\max_{k=1,\ldots,r}\{C_{x_k}\},\ C_L\right), \tag{2.14}$$

inequality (2.6) follows.

To show (2.7), notice that from (2.12) we have

$$\begin{aligned} &P(|\hat{g}_N(x) - g(x)| < \delta \text{ for all } x \in X) \\ &\quad \geq P\left(\{|\hat{g}_N(x_k) - g(x_k)| < \delta/3, \quad k = 1,\ldots,r\} \cap \{|\hat{L}_N - \mathbb{E}[L(\boldsymbol{\xi})]| < \delta/3\}\right) \\ &\quad \geq 1 - \sum_{k=1}^{r} P\left(|\hat{g}_N(x_k) - g(x_k)| \geq \delta/3\right) - P(|\hat{L}_N - \mathbb{E}[L(\boldsymbol{\xi})]| \geq \delta/3), \end{aligned} \tag{2.15}$$

where the last inequality stems from a direct application of Bonferroni's inequality. It follows from (2.10), (2.11), and (2.15) that

$$P\left(|\hat{g}_N(x) - g(x)| < \delta \text{ for all } x \in X\right) \geq 1 - \frac{r+1}{2} A e^{-\alpha N},$$

so by taking

$$K := \frac{r+1}{2} A \tag{2.16}$$

we obtain (2.7). □

We return now to the proof of Theorem 2.3.

Proof. Let $\varepsilon > 0$ be given. Assumption A2 implies the existence of some $\delta > 0$ such that $\text{dist}(\hat{x}_N, S^*) < \varepsilon$ whenever $|\hat{g}_N(x) - g(x)| < \delta$ for all $x \in X$; see, e.g., [46, p. 69] for a proof. By Lemma 2.4, the event $\{|\hat{g}_N(x) - g(x)| < \delta \text{ for all } x \in X\}$ occurs with probability at least $1 - Ke^{-\alpha N}$ (where both K and α depend on δ). It follows that

$$P\left(\text{dist}(\hat{x}_N, S^*) \geq \varepsilon\right) \leq Ke^{-\alpha N}$$

as asserted. Notice that δ does not depend on the particular approximation $\hat{g}_N(\cdot)$; therefore, from (2.13), (2.14), and (2.16) we see that the constants K and α depend on $\{\hat{g}_N(\cdot)\}$ and $\{\hat{L}_N\}$ only through, respectively, the constants C_x and C_L and the exponent functions $\gamma_x(\cdot)$ and $\gamma_L(\cdot)$ in Assumptions B1 and B1$_L$. □

In essence, Theorem 2.3 says that the existence of an exponential rate of convergence for *pointwise estimators* is enough to ensure an exponential rate of convergence for *optimal solutions* of the corresponding approximating problems, regardless of the sampling scheme adopted. Although reasonably intuitive, such a result has not—to the best of our knowledge—been stated or proved anywhere in the literature.

It is important to remark that the second part of Theorem 2.3 suggests that a better pointwise convergence rate leads to a better rate of convergence of optimal solutions. Indeed, suppose that one has at hand two families of approximations, say, $\{\bar{g}_N(x)\}$ and $\{\tilde{g}_N(x)\}$, whose respective exponent functions $\bar{\gamma}_x(\cdot)$ and $\tilde{\gamma}_x(\cdot)$ in (2.1) are such that $\bar{\gamma}_x(\cdot) \geq \tilde{\gamma}_x(\cdot)$ for all $x \in X$. Then the corresponding constants $\bar{\alpha}$ and $\tilde{\alpha}$ will be such that $\bar{\alpha} \geq \tilde{\alpha}$, which suggests that the family $\{\bar{g}_N(\cdot)\}$ yields a better rate

of convergence of $\hat{x}_N$ to S^*. Of course, Theorem 2.3 gives only an *upper bound* on the deviation probability $P\left(\text{dist}(\hat{x}_N, S^*) \geq \varepsilon\right)$, so no definitive statements can be made.

Nevertheless, we shall see later specific situations where the pointwise rate of convergence yields an asymptotically exact rate of convergence for the optimization problem; in those cases, the superiority of one sampling scheme over another can be established.

2.1.2. The finite/piecewise linear case. We derive now results that parallel the ones in section 2.1.1 but with the following assumption in place of Assumption A2.

Assumption A3. Either (i) the feasibility set X is finite or (ii) X is compact, convex, and polyhedral, the function $G(\cdot, \boldsymbol{\xi})$ is convex piecewise linear for every value of $\boldsymbol{\xi}$, and the distribution of $\boldsymbol{\xi}$ has finite support.

Assumption A3 is useful in the context of discrete stochastic optimization (case (i)) or stochastic linear programs (case (ii)). The proposition below shows consistency of the estimators. In the proposition (and elsewhere in this paper), the statement "w.p.1 for N large enough" means that, with probability one, there exists an N_0 such that, on each sample path of the underlying process, the condition holds for all $N > N_0$. The value of such N_0 depends on the particular sample path. The proof of the proposition follows a similar argument to that of Theorem 2.6 below and is therefore omitted.

PROPOSITION 2.5. *Suppose that Assumptions* A1 *and* A3 *hold. Then*

(i) $\hat{g}_N(x) \to g(x)$ *uniformly on* X *w.p.*1*;*

(ii) $\hat{\nu}_N \to \nu^*$ *w.p.*1*;*

(iii) $\hat{x}_N \in S^*$ *w.p.*1 *for* N *large enough.*

THEOREM 2.6. *Consider problem* (1.3)*, and suppose that Assumptions* A3 *and* B1 *hold. Then there exist constants* $K > 0$ *and* $\alpha > 0$ *such that*

$$P\left(\hat{x}_N \notin S^*\right) \leq Ke^{-\alpha N} \qquad \text{for all } N \geq 1.$$

Moreover, the constants K *and* α *depend on the family of estimators* $\{\hat{g}_N(\cdot)\}$ *only through the constants* C_x *and the exponent functions* $\gamma_x(\cdot)$ *in* (2.1)*.*

The proof of Theorem 2.6 will be based on the following lemma.

LEMMA 2.7. *Suppose that Assumption* B1 *holds and that the set* X *is finite. Then, for any* $\delta > 0$*, there exist positive constants* $A = A(\delta)$ *and* $\alpha = \alpha(\delta)$ *such that*

$$P\left(|\hat{g}_N(x) - g(x)| \geq \delta\right) \leq Ae^{-\alpha N} \qquad \text{for all } x \in X \text{ and all } N \geq 1. \tag{2.17}$$

Moreover, there exists a positive constant K *(also dependent on* δ*) such that*

$$P\left(|\hat{g}_N(x) - g(x)| < \delta \text{ for all } x \in X\right) \geq 1 - Ke^{-\alpha N} \qquad \text{for all } N \geq 1. \tag{2.18}$$

Proof. By setting $\alpha := \min_{x \in X} \gamma_x(\delta)$ and $A := \max_{x \in X} C_x$ in (2.1), we immediately show (2.17). The proof of (2.18) follows a very similar argument to that in the proof of Lemma 2.4 and is therefore omitted. □

We return now to the proof of Theorem 2.6.

Proof. Suppose initially that X is finite. Let δ be defined as $(1/2)\min_{x \in X \setminus S^*} g(x) - \nu^*$. It is clear that, if $|\hat{g}_N(x) - g(x)| < \delta$ for all $x \in X$, we have $\hat{g}_N(x) < \hat{g}_N(y)$ for all $x \in S^*$ and all $y \in X \setminus S^*$, i.e., $\hat{x}_N \in S^*$. Now suppose that the conditions in part (ii) of Assumption A3 hold. Then, from Lemma 2.4 in [53], we know that there exists a finite set of points $\{x_1, \ldots, x_\ell\} \cup \{y_1, \ldots, y_q\}$ such that $x_i \in S^*$, $y_j \in X \setminus S^*$ and, if $\hat{g}_N(x_i) < \hat{g}_N(y_j)$ for all $i \in \{1, \ldots, \ell\}$ and all $j \in \{1, \ldots, q\}$, then $\hat{x}_N \in S^*$ (in fact, the set S_N forms a face of S^*). Therefore, we can use the same argument as in the

case where X is finite. We remark that similar results were derived in [23, 53] in the i.i.d. context.

Next, by Lemma 2.7, the event $\{|\hat{g}_N(x) - g(x)| < \delta \text{ for all } x \in X\}$ occurs with probability at least $1 - Ke^{-\alpha N}$ (where both K and α depend on δ). It follows that

$$P(\hat{x}_N \notin S^*) \leq Ke^{-\alpha N}$$

as asserted. As argued in the proof of Theorem 2.3, δ does not depend on the particular approximation $\hat{g}_N(\cdot)$, so the constants K and α depend on $\{\hat{g}_N(\cdot)\}$ only through the constants C_x and the exponent functions $\gamma_x(\cdot)$ in Assumption B1. □

We conclude this section by mentioning that an analogous form of Theorems 2.3 and 2.6 can be derived in case Assumption B1′ holds instead of B1. We state the result below for completeness; the proof follows very similar steps to the proofs of those theorems and is therefore omitted.

THEOREM 2.8. *Consider problem* (1.3)*, and suppose that Assumption* B1′ *holds.*

1. *Suppose that Assumptions* A2 *and* B1_L *hold. Then, given* $\varepsilon > 0$*, there exists a constant* $\alpha > 0$ *such that*

$$\limsup_{N\to\infty} \frac{1}{N} \log P(\text{dist}(\hat{x}_N, S^*) \geq \varepsilon) \leq -\alpha. \tag{2.19}$$

2. *Suppose that Assumption* A3 *holds. Then there exists a constant* $\alpha > 0$ *such that*

$$\limsup_{N\to\infty} \frac{1}{N} \log P(\hat{x}_N \notin S^*) \leq -\alpha. \tag{2.20}$$

2.2. Convergence of estimators of optimal values. We consider now the convergence of the optimal value of (1.3). In the previous section we showed that an exponential rate of convergence for pointwise estimators leads to an exponential rate of convergence for solutions of (1.3); here we will show that, in the context of Assumption A3, a CLT-type result for pointwise estimators leads to a CLT-type result for the optimal value of (1.3). Outside the context of A3, however, one needs more than CLT for pointwise estimators.

We start by making the following probabilistic assumptions on the estimators $\{\hat{g}_N(x)\}$.

Assumption B2. For each $x \in S^*$, the random variable $W_N(x)$ defined as

$$W_N(x) := \frac{\hat{g}_N(x) - g(x)}{\sigma_N(x)}, \tag{2.21}$$

where $\sigma_N^2(x) := \text{Var}[\hat{g}_N(x)]$, is such that $W_N(x)$ converges in distribution to a standard Normal (denoted $W_N(x) \xrightarrow{d} \text{Normal}(0,1)$).

Of course, Assumption B2 holds in case of i.i.d. sampling under very mild assumptions—in that case it corresponds to the classical CLT (with $\sigma_N(x) = \sqrt{\text{Var}[G(x,\boldsymbol{\xi})]/N}$). However, as we shall see later, B2 holds in other contexts as well. Note that we impose Assumption B2 only on the set S^* of optimal solutions to (1.1).

The lemma below states a property that will be used in what follows.

LEMMA 2.9. *Suppose that Assumptions* A1 *and* A3 *hold. Then*

$$\hat{g}_N(\hat{x}_N) - \min_{x^* \in S^*} \hat{g}_N(x^*) = 0 \quad \textit{w.p.1 for } N \textit{ large enough.}$$

Proof. We have already seen in Proposition 2.5 that, under Assumptions A1 and A3, we have $\hat{x}_N \in S^*$ w.p.1 for N large enough. Consider now an arbitrary sample

path where such a condition holds. Then there exists N_0 such that $\hat{x}_N \in S^*$ for all $N > N_0$. That is, for each $N > N_0$ there exists some point $x^*(N) \in S^*$ such that $\hat{x}_N = x^*(N)$. It follows that

$$\hat{g}_N(\hat{x}_N) - \hat{g}_N(x^*(N)) = 0 \quad \text{for all } N > N_0.$$

By definition, $\hat{x}_N$ minimizes $\hat{g}_N(\cdot)$ over X. Together with the above equality, this implies that

$$\hat{g}_N(\hat{x}_N) \leq \min_{x^* \in S^*} \hat{g}_N(x^*) \leq \hat{g}_N(x^*(N)) = \hat{g}_N(\hat{x}_N) \quad \text{for all } N > N_0$$

and hence

$$\hat{g}_N(\hat{x}_N) - \min_{x^* \in S^*} \hat{g}_N(x^*) = 0 \quad \text{for all } N > N_0. \qquad \square$$

We then have the following result for rates of convergence.

THEOREM 2.10. *Consider problem* (1.3)*, and suppose that Assumptions* A1 *and* A3 *hold. Suppose also that the estimators* $\hat{g}_N(x)$ *have the* same variance *on the set* S^* *of optimal solutions to* (1.1)*, i.e., the function* $\sigma_N^2(\cdot)$ *is constant on* S^**, and let* $(\sigma_N^*)^2$ *denote that common value. Then*

$$\frac{\hat{\nu}_N - \nu^*}{\sigma_N^*} - \min_{x^* \in S^*} W_N(x^*) \xrightarrow{d} 0. \tag{2.22}$$

If, in addition, Assumption B2 *holds and problem* (1.1) *has a* unique *optimal solution (call it* x^**), then*

$$\frac{\hat{\nu}_N - \nu^*}{\sigma_N(x^*)} \xrightarrow{d} \text{Normal}(0,1). \tag{2.23}$$

Proof. By Lemma 2.9 we have

$$\frac{\hat{g}_N(\hat{x}_N) - \nu^*}{\sigma_N^*} - \frac{\min_{x^* \in S^*} \hat{g}_N(x^*) - \nu^*}{\sigma_N^*} = 0 \quad \text{w.p.1 for } N \text{ large enough.}$$

Since convergence w.p.1 implies convergence in distribution, it follows that

$$\frac{\hat{g}_N(\hat{x}_N) - \nu^*}{\sigma_N^*} - \frac{\min_{x^* \in S^*} \hat{g}_N(x^*) - \nu^*}{\sigma_N^*} \xrightarrow{d} 0$$

and hence

$$\frac{\hat{g}_N(\hat{x}_N) - \nu^*}{\sigma_N^*} - \min_{x^* \in S^*} \frac{\hat{g}_N(x^*) - \nu^*}{\sigma_N^*} \xrightarrow{d} 0.$$

Note that the term inside the min operation is actually $W_N(x^*)$. Moreover, by definition $\hat{g}_N(\hat{x}_N) = \hat{\nu}_N$, which then shows (2.22).

Suppose now that B2 holds and that $S^* = \{x^*\}$. Then, since $W_N(x^*) \xrightarrow{d}$ Normal$(0,1)$, by using a classical result in convergence of distributions (see, e.g., [3, Theorem 3.1]), we conclude that

$$\frac{\hat{\nu}_N - \nu^*}{\sigma_N(x^*)} \xrightarrow{d} \text{Normal}(0,1). \qquad \square$$

The above result can be slightly strengthened in case the set S^* is finite (say, $S^* = \{x^1, \ldots, x^\ell\}$) and a multivariate version of Assumption B2 holds—namely, that for some deterministic sequence $\{\tau_N\}$ such that $\tau_N \to \infty$ the multivariate process $\tau_N(\hat{g}_N(x^1) - g(x^1), \ldots, \hat{g}_N(x^\ell) - g(x^\ell))$ converges in distribution to a random vector Y with Normal distribution with mean vector zero and covariance matrix Σ. In that case, by using a very similar argument to that used in [23], one can show directly that $\tau_N(\hat{\nu}_N - \nu^*)$ converges in distribution to $\min_{x^* \in S^*} Y(x^*)$. We chose to present our result in the above form because it requires only a univariate CLT.

As mentioned earlier, outside the context of Assumption A3 stronger conditions are required. One possibility is to assume that Assumption A2 holds and that a version of Assumption B2 for functional spaces holds for the space $C(X)$ of continuous functions defined on X. As discussed in [48], Assumption A2 suffices to ensure that each $G(\cdot, \boldsymbol{\xi})$ is a random element of the space $C(X)$, and hence $\hat{g}_N(\cdot) := N^{-1} \sum_{j=1}^N G(\cdot, \boldsymbol{\xi})$ is also a random element of $C(X)$. The validity of a CLT in that functional space, in turn, implies that a convergence result such as (2.23) holds. This approach works well in the i.i.d. context; see [48] for a discussion. However, we are not aware of other contexts where a CLT in a functional space exists, so we do not elaborate further on this topic.

3. Applications. We discuss now the application of the results developed in section 2 to two classes of non-i.i.d. sampling techniques—namely, LHS and randomized QMC methods. Note that these techniques are devised to sample s-dimensional random vectors U that are uniformly distributed on $[0,1]^s$ and have independent components. Given an s-dimensional random vector $\boldsymbol{\xi}$ with arbitrary distribution and not necessarily independent components, it is always possible to write $\boldsymbol{\xi} = \Psi(U)$ for some mapping $\Psi : [0,1]^s \mapsto \mathbb{R}^s$, which is constructed by inverting the conditional distribution of ξ_j, given $\xi_1, \ldots, \xi_{j-1}$, $j = 1, \ldots, s$; for details see, for example, [47]. In practice, it is difficult to generate Ψ for a general multivariate distribution, so such a method is typically used when either the distribution has a special form or the components of $\boldsymbol{\xi}$ are independent. In the latter case, $\Psi(u_1, \ldots, u_s) = (F_1^{-1}(u_1), \ldots, F_s^{-1}(u_s))$, where F_j^{-1} is the inverse of the cumulative distribution function F_j of ξ_j, defined as $F_j^{-1}(u) := \inf\{y \in \Xi_j : F_j(y) \geq u\}$, and Ξ_j denotes the support of the distribution F_j. For the remainder of this paper *we assume that the components of* $\boldsymbol{\xi}$ *are independent.* Moreover, since $G(x, \boldsymbol{\xi}) = G(x, \Psi(U))$, we will restrict the domain of $G(x, \cdot)$ to $\Xi_1 \times \ldots \Xi_s$. Sometimes we will refer to the function $G(x, \Psi(\cdot))$, which is defined on $[0,1]^s$.

3.1. Latin hypercube sampling. Stratified sampling techniques have been used in statistics and simulation for years (see, for instance, [11] and references therein). Generally speaking, the idea is to partition the sample space and fix the number of samples on each component of the partition, which should be proportional to the probability of that component. This way we ensure that the number of sampled points on each region will be approximately equal to the *expected* number of points to fall in that region. It is intuitive that such a procedure yields a smaller variance than crude Monte Carlo methods; for proofs see [11]. Notice, however, that, though theoretically appealing, implementing such a procedure is far from trivial, since the difficulty is to determine the partition as well as to compute the corresponding probabilities.

There are many variants of this basic method, one of the most well known being the so-called LHS, introduced in [31]. The LHS method operates as follows: Suppose that we want to draw N samples from a random vector $\boldsymbol{\xi}$ with s independent

components $\xi_1, \ldots, \xi_s$, each of which has a Uniform(0,1) distribution. The algorithm consists of repeating the two steps below for each dimension $j = 1, \ldots, s$:

1. Generate

$$Y^1 \sim U\left(0, \frac{1}{N}\right), \qquad Y^2 \sim U\left(\frac{1}{N}, \frac{2}{N}\right), \ldots, Y^N \sim U\left(\frac{N-1}{N}, 1\right);$$

2. let $\xi_j^i := Y^{\pi(i)}$, where π is a random permutation of $1, \ldots, N$.

In [31], it is shown that each sample ξ_j^i (viewed as a random variable) has *the same distribution* as ξ_j, which in turn implies that the estimators generated by the LHS method are unbiased. In case of arbitrary distributions, the above procedure is easily modified by drawing the sample as before and applying the inversion method discussed at the beginning of section 3 to generate the desired random variates.

It is also shown in [31] that, under some conditions, the LHS method does indeed reduce the variance compared to crude Monte Carlo methods. The papers [37, 55] show that, asymptotically (i.e., as the sample size N goes to infinity), LHS is never worse than crude Monte Carlo methods, even without the assumptions of [31]. More specifically, $V_{\text{LHS}} \leq N/(N-1)V_{\text{MC}}$, where V_{LHS} and V_{MC} are the variances under LHS and crude Monte Carlo methods, respectively.

3.1.1. Exponential rate of convergence. Suppose that the objective function $g(\cdot)$ in (1.1) is approximated by a sample average calculated by using the LHS method; i.e., for each $i = 1, \ldots, s$, $\xi_i^1, \ldots, \xi_i^N$ are samples of ξ_i (the ith component of $\boldsymbol{\xi}$) constructed by using the LHS method. Call the resulting estimator in (1.2) $\hat{g}_N^{\text{LHS}}(x)$. To study convergence properties of the approximating problem in (1.3), we shall use the tools of section 2. Our first goal is to show that the family $\{\hat{g}_N^{\text{LHS}}(\cdot)\}$ satisfies Assumption B1, so that we can apply Theorems 2.3 and 2.6 to ensure an exponential rate of convergence.

We shall restrict our attention to functions satisfying the following assumption.

Assumption C1. For each $x \in X$, the function $G(x, \cdot)$ is monotone in each component. That is, for each $i = 1, \ldots, s$ and each $\delta > 0$ we have

$$\text{either} \qquad G(x, z + \delta e_i) \geq G(x, z) \qquad \text{for all } z \in \mathbb{R}^s \tag{3.1}$$

$$\text{or} \qquad G(x, z + \delta e_i) \leq G(x, z) \qquad \text{for all } z \in \mathbb{R}^s, \tag{3.2}$$

where as customary e_i denotes the vector with 1 in the ith component and zeros otherwise.

An important case where such an assumption is satisfied is that of two-stage stochastic linear programs with fixed recourse. In section 4 we discuss that case in detail.

An alternative assumption is the following.

Assumption C1′. For each $x \in X$, the function $G(x, \cdot)$ is *additive*; i.e., there exist functions $G_1, \ldots, G_s$ (all of them mapping $\mathbb{R}^n \times \mathbb{R}$ to $\mathbb{R}$) such that $G(x, \boldsymbol{\xi}) = G_1(x, \xi_1) + \cdots + G_s(x, \xi_s)$. Moreover, $|\mathbb{E}[G_j(x, \xi_j)]| < \infty$, the functions $G_j(x, F_j^{-1}(\cdot))$ have at most a finite number of singularities (i.e., points where the function approaches $\pm\infty$), and the set of points at which $G_j(x, F_j^{-1}(\cdot))$ is discontinuous has Lebesgue measure zero.

The importance of Assumptions C1 and C1′ in the present context is given by the results below.

THEOREM 3.1. *Suppose that* (i) *Assumption* C1 *holds and* (ii) *for each $x \in X$, the moment-generating function of $G(x, \boldsymbol{\xi})$ (denoted $\phi_x^{\text{MC}}(t) := \mathbb{E}[e^{tG(x,\xi)}]$) is finite*

everywhere. Consider the LHS estimators $\hat{g}_N^{\mathrm{LHS}}(\cdot)$ *above defined and the corresponding problem* $\min_{x\in X} \hat{g}_N^{\mathrm{LHS}}(x)$. *Let* $\hat{x}_N^{\mathrm{LHS}}$ *denote an optimal solution of that problem. Then*

1. *if Assumption* A2 *holds with a uniform Lipschitz constant* L *(i.e.,* $L(\cdot) \equiv L$*), then given* $\varepsilon > 0$ *there exist constants* $\tilde{K} > 0$ *and* $\tilde{\alpha} > 0$ *such that*

$$P\left(\operatorname{dist}(\hat{x}_N^{\mathrm{LHS}}, S^*) \geq \varepsilon\right) \leq \tilde{K} e^{-\tilde{\alpha} N} \qquad \text{for all } N \geq 1; \tag{3.3}$$

2. *if Assumption* A3 *holds, then there exists a constant* $\tilde{\alpha} > 0$ *such that*

$$P\left(\hat{x}_N^{\mathrm{LHS}} \notin S^*\right) \leq \tilde{K} e^{-\tilde{\alpha} N} \qquad \text{for all } N \geq 1. \tag{3.4}$$

Moreover, in either case the exponent $\tilde{\alpha}$ *is at least as large as the corresponding exponent obtained for standard Monte Carlo methods.*

Proof. Let $\phi_N(x,t) := \frac{1}{N} \log \mathbb{E}[e^{tN\hat{g}_N^{\mathrm{LHS}}(x)}]$. If conditions (i) and (ii) above hold, then by Proposition 6 in [9] we have $\phi_N(x,t) \leq \phi_x^{\mathrm{MC}}(t)$ for all x and all t, and hence it follows from Proposition 2.1 that Assumption B1 holds for $\{\hat{g}_N^{\mathrm{LHS}}(\cdot)\}$. Moreover, in case 1 of the theorem (i.e., when Assumption A2 holds with $L(\cdot) \equiv L$) Assumption $\mathrm{B1}_L$ is trivially satisfied. The two cases of the theorem then parallel Theorems 2.3 and 2.6, which shows (3.3) and (3.4).

The last assertion of the theorem is a consequence of the remark following the proof of Theorem 2.3. Indeed, the arguments in the previous paragraph show that the constants C_x and the exponent functions $\gamma_x(\cdot)$ used to show (2.1) are the same for both LHS and standard Monte Carlo methods. □

Although Theorem 3.1 guarantees only the same bounds for both LHS and standard Monte Carlo methods, a closer look at the proof of the inequality $\phi_N(x,\cdot) \leq \phi_x^{\mathrm{MC}}(\cdot)$ in [9] shows that such an inequality is essentially a consequence of Jensen's inequality, which often holds strictly; hence, generally speaking, LHS tends to behave better than Monte Carlo methods.

In case Assumption C1′ holds instead of C1, we have the following stronger result.

THEOREM 3.2. *Suppose that the assumptions of Theorem* 3.1 *are satisfied, but Assumption* C1′ *holds instead of* C1. *Then the conclusions of Theorem* 3.1 *hold. In addition, we have the following:*

1. *If Assumption* A2 *holds with a uniform Lipschitz constant* L *(i.e.,* $L(\cdot) \equiv L$*), then*

$$\lim_{N\to\infty} \frac{1}{N} \log P\left(\operatorname{dist}(\hat{x}_N^{\mathrm{LHS}}, S^*) \geq \varepsilon\right) = -\infty. \tag{3.5}$$

2. *If Assumption* A3 *holds, then*

$$\lim_{N\to\infty} \frac{1}{N} \log P\left(\hat{x}_N^{\mathrm{LHS}} \notin S^*\right) = -\infty. \tag{3.6}$$

Proof. The proof of the first part of the theorem follows the same steps as the proof of Theorem 3.1 (except that Proposition 4 in [9] is invoked instead of Proposition 6).

To show the second part, by writing each random variable ξ_j as $F_j^{-1}(U_j)$ (where $U_j \sim U(0,1)$), we have that conditions (i) and (ii) of Theorem 3.1 ensure that the assumptions of Theorem 3 in [9] are satisfied. The latter result, in turn, ensures that Assumption B1′ holds with the function $\gamma_x = \infty$ everywhere except at zero, where it is equal to zero. Then (3.5) and (3.6) follow from (2.19) and (2.20) in Theorem 2.8. □

The strength of Theorem 3.2, of course, lies in the asymptotic results (3.5)–(3.6), which show that in the additive case the rate of convergence under LHS is *superexponential*.

3.1.2. Central limit theorem. We study now the convergence of optimal values of the approximating problem (1.3) under LHS. To do so we shall apply the results of section 2.2. Before that, however, we need to review some results related to the ANOVA decomposition of a function.

Let $U = (U_1, \ldots, U_s)$ be an s-dimensional random vector of independent components with uniform distribution on $[0,1]^s$ and $f : [0,1]^s \longrightarrow \mathbb{R}$ an arbitrary measurable function, and consider the problem of estimating $I := \mathbb{E}[f(U)]$. It is shown in [55] that, when $\mathbb{E}[f(U)^2] < \infty$, f can be decomposed as

$$f(u_1, \ldots, u_s) = \mathbb{E}[f(U)] + \sum_{k=1}^{s} f_k(u_k) + r(u_1, \ldots, u_s), \tag{3.7}$$

where $f_k(u_k) = \mathbb{E}[f(U) \,|\, U_k = u_k] - \mathbb{E}[f(U)]$ and $r(u)$ is the *residual* term, which satisfies $\mathbb{E}[r(U) \,|\, U_j = u_j] = 0$ for all j and all u_j. [55] also shows that the residual term can be viewed as a "residual from additivity" in the following sense. We say that a function $g : [0,1]^s \longrightarrow \mathbb{R}$ is *additive* if there exist unidimensional functions $g_1, \ldots, g_s$ and a constant C such that g can be written as $g(u_1, \ldots, u_s) = C + \sum_k g_k(u_k)$ for almost all $u \in [0,1]^s$ (where "almost all" refers to the Lebesgue measure). Then the additive function $f_a : [0,1]^s \longrightarrow \mathbb{R}$, defined as $f_a(u_1, \ldots, u_s) = \mathbb{E}[f(U)] + \sum_{k=1}^{s} f_k(u_k)$, is the best additive fit to f in the L^2-norm; i.e., it minimizes $\mathbb{E}[(f(U) - g(U))^2]$ over all additive functions g. Note that if f is itself additive, then the residual $r(u) = f(u) - f_a(u)$ will be equal to zero almost everywhere (a.e.); conversely, if $r(u) = 0$ a.e., then $f(u) = f_a(u)$ a.e., so in that case f is additive.

The variance of the estimator I_{LHS} (defined as $I_{\mathrm{LHS}} := N^{-1} \sum_{i=1}^{N} f(U^i)$, where $U^1, \ldots, U^N$ are samples drawn with LHS) satisfies

$$\sigma_N^2 := \mathrm{Var}[I_{\mathrm{LHS}}] = N^{-1} \mathbb{E}\left[(r(U))^2\right] + o(N^{-1}); \tag{3.8}$$

see [55]. By using (3.7) and (3.8), it is shown in [33] that, when f is bounded, a CLT holds for the estimator I_{LHS} under LHS. More specifically, it is shown that

$$N^{1/2}(I_{\mathrm{LHS}} - I) \;\xrightarrow{d}\; \mathrm{Normal}(0, \sigma^2), \quad \text{where } \sigma^2 := \mathbb{E}[(r(U))^2]. \tag{3.9}$$

Next, notice that from (3.8) we can write

$$\frac{I_{\mathrm{LHS}} - I}{\sigma_N} = \frac{N^{1/2}(I_{\mathrm{LHS}} - I)}{\left[\sigma^2 + \frac{o(N^{-1})}{N^{-1}}\right]^{1/2}}.$$

Since $N^{1/2}(I_{\mathrm{LHS}} - I) \xrightarrow{d} \mathrm{Normal}(0, \sigma^2)$ and the deterministic sequence $\{[\sigma^2 + \frac{o(N^{-1})}{N^{-1}}]^{1/2}\}$ converges to σ, it follows from a classical result in probability theory (see, e.g., [3, p. 29]) that, when $\sigma > 0$,

$$\frac{I_{\mathrm{LHS}} - I}{\sigma_N} \;\xrightarrow{d}\; \frac{1}{\sigma}\mathrm{Normal}(0, \sigma^2) = \mathrm{Normal}(0, 1). \tag{3.10}$$

Notice that the condition $\mathbb{E}[f(U)^2] < \infty$ also implies that a strong law of large numbers holds for LHS, i.e.,

$$|I_{\mathrm{LHS}} - I| \;\to\; 0 \quad \text{w.p.1}; \tag{3.11}$$

for a proof, see [27].

By applying (3.10) to our setting we see that Assumption B2 holds for LHS when, for every $x \in S^*$, the random variable $G(x, \boldsymbol{\xi})$ is bounded and the function $G(x, \Psi(\cdot))$ has a nonzero ANOVA residual. As seen above, the latter condition means that $G(x, \Psi(\cdot))$ is not additive; i.e., it cannot be written in the form $C + \sum_k g_k(u_k)$. It is easy to see that this is equivalent to saying that the function $G(x, \cdot)$ is not additive; note that here we extend the definition of additivity to the domain of $G(x, \cdot)$, which as discussed before is restricted to the support of $\boldsymbol{\xi}$.

On the other hand, if $\mathbb{E}[G(x, \boldsymbol{\xi})^2] < \infty$ for all $x \in X$, then (3.11) implies that Assumption A1 holds. Thus, under additional assumptions we can apply Theorem 2.10 and Propositions 2.2 and 2.5. We summarize the result in the theorem below.

THEOREM 3.3. *Consider the LHS estimators $\hat{g}_N^{\mathrm{LHS}}(\cdot)$ defined above and the corresponding problem $\min_{x \in X} \hat{g}_N^{\mathrm{LHS}}(x)$. Let $\hat{x}_N^{\mathrm{LHS}}$ and $\hat{\nu}_N^{\mathrm{LHS}}$ denote, respectively, an optimal solution and the optimal value of that problem. Suppose that $\mathbb{E}[G(x, \boldsymbol{\xi})^2] < \infty$ for all $x \in X$.*

1. *If Assumption* A2 *holds, then* $\operatorname{dist}(\hat{x}_N^{\mathrm{LHS}}, S^*) \to 0$ *w.p.*1 *and* $\hat{\nu}_N^{\mathrm{LHS}} \to \nu^*$ *w.p.*1.
2. *If Assumption* A3 *holds, then $\hat{x}_N^{\mathrm{LHS}} \in S^*$ w.p.*1 *for N large enough and $\hat{\nu}_N^{\mathrm{LHS}} \to \nu^*$ w.p.*1. *In addition, if problem* (1.1) *has a* unique *optimal solution (call it x^*), the random variable $G(x^*, \boldsymbol{\xi})$ is bounded, and the function $G(x^*, \cdot)$ is not additive, then*

$$\frac{\hat{\nu}_N^{\mathrm{LHS}} - \nu^*}{\sigma_N(x^*)} \xrightarrow{d} \mathrm{Normal}(0, 1),$$

where $\sigma_N^2(x^*)$ is the variance of $\hat{g}_N^{\mathrm{LHS}}(x^*)$. Moreover, there exists a positive constant C such that

$$\sigma_N^2(x^*) = N^{-1}C + o(N^{-1}). \tag{3.12}$$

Theorem 3.3 shows that the rate of convergence of optimal values under LHS (under the conditions of case 2 of the theorem) is $N^{-1/2}$. Thus, compared to standard Monte Carlo methods we can see that, although LHS will likely reduce the variance of pointwise estimators, it cannot improve the *rate* of convergence unless $G(x^*, \cdot)$ is additive; in that case, we expect the convergence rate to be much faster. Indeed, recall from Theorem 3.2 that, under the assumptions of that theorem (which include additivity), the convergence of optimal solutions is superexponential. Note also that, when S^* is finite (but not necessarily a singleton), $G(x^*, \boldsymbol{\xi})$ is bounded, and $G(x^*, \cdot)$ is not additive for all $x^* \in S^*$, the stronger result discussed in the paragraph following the proof of Theorem 2.10 applies with $\tau_N = N^{1/2}$, since the aforementioned CLT result proved in [33] is also valid in a multivariate context.

3.2. Randomized QMC. For completeness, we provide in this section a brief review of QMC techniques. We follow mostly [32], which we refer to for a comprehensive treatment of QMC concepts. Let U be an s-dimensional random vector with uniform distribution on $[0, 1]^s$ and $f : [0, 1]^s \longrightarrow \mathbb{R}$ an arbitrary function, and consider the problem of estimating $I := \mathbb{E}[f(U)]$.

The basic idea of QMC is to calculate a sample average estimate as in the standard Monte Carlo method, but, instead of drawing a random sample from the uniform distribution on $[0, 1]^s$, a certain set of points $u^1, \ldots, u^N$ on space $[0, 1]^s$ is carefully chosen. The deterministic estimate

$$I_{\mathrm{QMC}} := \frac{1}{N} \sum_{i=1}^{N} f(u^i) \tag{3.13}$$

is constructed. A key result is the so-called Koksma–Hlawka inequality, which, roughly speaking, states that the quality of the approximation given by I_{QMC} depends on the quality of the chosen points (measured by the difference between the corresponding empirical measure and the uniform distribution, which is quantified by the so-called *star discrepancy*) as well as on the nature of the function f (measured by its total variation). A great deal of the research on QMC methods aims at determining ways to construct *low-discrepancy sequences*, i.e., sequences of points $u^1, u^2, \ldots$ for which the star discrepancy is small for all N. A particular type of sequence that has proven valuable is defined in terms of (t, m, s)-nets. We need some definitions before delving into more details, which we do next.

Let $b \geq 2$ be an arbitrary integer called the base. An *elementary interval in base* b (in dimension s) is a subinterval E of $[0,1]^s$ of the form

$$E = \prod_{j=1}^{s} \left[\frac{a_j}{b^{d_j}}, \ \frac{a_j + 1}{b^{d_j}} \right]$$

for nonnegative integers $\{a_j\}$ and $\{d_j\}$ such that $a_j < b^{d_j}$ for all j. The volume of E is $b^{-\sum_j d_j}$. Next, let t and m be nonnegative integers such that $t \leq m$. A finite sequence of b^m points is a (t, m, s)-*net* in base b if every elementary interval in base b of volume b^{t-m} contains exactly b^t points of the sequence. A sequence of points $u^1, u^2, \ldots$ is a (t, s)-*sequence* in base b if, for all integers $k \geq 0$ and $m > t$, the set of points consisting of the u^n such that $kb^m \leq n < (k+1)b^m$ is a (t, m, s)-net in base b.

The advantage of (t, m, s)-nets becomes clear from a result due to Niederreiter [32, Theorems 4.10 and 4.17], who shows that the error $|I_{\mathrm{QMC}} - I|$ is (i) of order $(\log N)^{s-1}/N$ when I_{QMC} is computed from a (t, m, s)-net in base b with $m > 0$ and (ii) of order $(\log N)^s/N$ when I_{QMC} is computed from the first $N \geq 2$ terms of a (t, s)-sequence in base b. Note that in case (i) N must be equal to b^m, whereas in case (ii) N is arbitrary, which explains the weaker bound. In either case, it is clear that, asymptotically, the error is smaller than $N^{-1/2}$ given by standard Monte Carlo methods.

Despite the advantage of QMC with respect to error rates, the method has two major drawbacks:

(a) The bounds provided by the Koksma–Hlawka inequality involve difficult-to-compute quantities such as the total variation of f; i.e., they yield qualitative (rather than quantitative) results. Hence, obtaining an exact estimate of the error may be difficult.

(b) A comparison of the functions $(\log N)^s/N$ and $N^{-1/2}$ shows that, even though asymptotically the error from QMC is smaller than the error from standard Monte Carlo methods, such an advantage does not appear until N is very large, unless s is small.

These difficulties have long been realized by the QMC community, and various remedies have been proposed. A common way to overcome difficulty (a) above is to incorporate some randomness into the choice of QMC points. By doing so, errors can be estimated by using standard methods, e.g., via multiple independent replications. This is the main idea of *randomized* QMC methods (RQMC); see [12, 38] for detailed discussions.

One particular technique we are interested in using relies on "scrambling" the decimal digits of each point of a (t, s)-sequence in a proper way. This idea was proposed in [34] and has gained popularity due to the nice properties of the randomized sequence. We shall use these properties below.

3.2.1. Using QMC in optimization. Consider again the family of estimators defined in (1.2). Suppose that $\{\boldsymbol{\xi}^i\}$ is generated by a (t,s)-sequence, and call the resulting family $\{\hat{g}_N^{\mathrm{QMC}}(x)\}$.

Let us fix $x \in X$ for a moment. As argued by the authors of [40]—who in turn cite a result in [30]—the empirical measure defined by a QMC sequence converges weakly to the uniform distribution, provided that the star discrepancy of that sequence goes to zero, which is the case of (t,s)-sequences. It follows from [3, Theorem 2.7] that, if $G(x,\boldsymbol{\xi})$ is bounded, then $\hat{g}_N^{\mathrm{QMC}}(x) \to g(x)$ as $N \to \infty$. Now suppose that $\{\boldsymbol{\xi}^i\}$ is generated by a scrambled (t,s)-sequence, and call the corresponding estimator $\hat{g}_N^{\mathrm{RQMC}}$. In [35] it is shown that scrambled (t,s)-sequences are (t,s)-sequences with probability one, which then implies that

$$\hat{g}_N^{\mathrm{RQMC}}(x) \to g(x) \quad \text{w.p.1.} \tag{3.14}$$

Moreover, $\hat{g}_N^{\mathrm{RQMC}}(x)$ is an unbiased estimator of $g(x)$, i.e., $\mathbb{E}[\hat{g}_N^{\mathrm{RQMC}}(x)] = g(x)$. Notice that the term "with probability one" above refers to the probability space where the random variables defining the permutations that are part of the scrambling algorithm lie. We assume that this probability space is the same as the one where the random vectors $\boldsymbol{\xi}$ are defined.

For some of the results that follow we will need the following assumption.

Assumption D1. The following conditions hold for each $x \in S^*$:

$$\left| \frac{\partial^s}{\partial u_1 \dots \partial u_s} G(x, \Psi(u_1, \dots u_s)) - \frac{\partial^s}{\partial u_1 \dots \partial u_s} G(x, \Psi(v_1, \dots v_s)) \right| \le B\|u - v\|^\beta \tag{3.15}$$

(for some $B > 0$ and some $\beta \in (0,1]$), and

$$\int_{[0,1]^s} \left[\frac{\partial^s}{\partial u_1 \dots \partial u_s} G(x, \Psi(u_1, \dots u_s)) \right]^2 du \; > \; 0, \tag{3.16}$$

where $\Psi(u_1, \dots u_s) = (F_1^{-1}(u_1), \dots, F_s^{-1}(u_s))$.

A few remarks about cases where Assumption D1 is satisfied are now in order. Suppose momentarily that G is infinitely differentiable in the second argument and that each F_j^{-1} is differentiable as well. Then we have

$$\begin{aligned} &\frac{\partial}{\partial u_1} G(x, F_1^{-1}(u_1), \dots, F_s^{-1}(u_s)) \\ &\qquad = \frac{\partial}{\partial \xi_1} G(x, \xi_1, F_2^{-1}(u_2) \dots, F_s^{-1}(u_s)) \Big|_{\xi_1 = F_1^{-1}(u_1)} \frac{\partial}{\partial u_1} F_1^{-1}(u_1), \end{aligned}$$

so, by repeating the calculation for the higher-order mixed derivatives, we obtain

$$H(u_1, \dots, u_s) := \frac{\partial^s}{\partial u_1 \dots \partial u_s} G(x, F_1^{-1}(u_1), \dots, F_s^{-1}(u_s)) \tag{3.17}$$

$$= \frac{\partial^s}{\partial \xi_1 \dots \partial \xi_s} G(x, \xi_1, \dots, \xi_s) \Big|_{\substack{\xi_j = F_j^{-1}(u_j) \\ j=1,\dots,s}} \frac{\partial}{\partial u_1} F_1^{-1}(u_1) \dots \frac{\partial}{\partial u_s} F_s^{-1}(u_s). \tag{3.18}$$

It follows that, if the gradient of the function H defined in (3.17) is uniformly bounded for all $u \in [0,1]^s$, then H is Lipschitz (see, e.g, [2, Corollary 40.6]); i.e., (3.15) holds.

A sufficient condition for uniform boundedness of $\nabla H(u)$ on $[0,1]^s$ is its continuity on $[0,1]^s$. Equation (3.18) shows that continuous differentiability of G (up to order $s+1$) and F_j^{-1}, $j=1,\ldots,s$ (up to second order), on the closed set $[0,1]^s$ suffices for that. Of course, imposing a continuous differentiability assumption on F_j^{-1} restricts the type of distributions that can be used; we shall return to that issue shortly.

Condition (3.16) essentially says that interactions of order up to s are significant, at least on a set of positive probability. For example, (3.16) does not hold if $G(x,\cdot)$ is linear for $x \in S^*$, since the mixed derivatives of any order bigger than 1 are equal to zero. Situations like that suggest that the *effective dimension* (see [35]) of the problem is less than s—indeed, in the linear case the effective dimension is 1. In that case, one should apply QMC only to the most significant variables, for which mutual interaction is significant.

By applying the above results on RQMC to the general context of section 2.2, we obtain the following.

THEOREM 3.4. *Consider the RQMC estimators $\hat{g}_N^{\mathrm{RQMC}}(\cdot)$ above defined and the corresponding problem $\min_{x\in X} \hat{g}_N^{\mathrm{RQMC}}(x)$. Let $\hat{x}_N^{\mathrm{RQMC}}$ and $\hat{\nu}_N^{\mathrm{RQMC}}$ denote an optimal solution and the optimal value of that problem, respectively. Suppose that $G(x,\boldsymbol{\xi})$ is bounded for all $x \in X$.*

1. *If Assumption* A2 *holds, then* $\mathrm{dist}(\hat{x}_N^{\mathrm{RQMC}}, S^*) \to 0$ *w.p.*1 *and* $\hat{\nu}_N^{\mathrm{RQMC}} \to \nu^*$ *w.p.*1.
2. *If Assumption* A3 *holds, then $\hat{x}_N^{\mathrm{RQMC}} \in S^*$ w.p.*1 *for N large enough and $\hat{\nu}_N^{\mathrm{RQMC}} \to \nu^*$ w.p.*1. *If, in addition, Assumption* D1 *holds, problem* (1.1) *has a* unique *optimal solution (call it x^*), and the samples $\{\boldsymbol{\xi}^i\}$ are generated by a scrambled $(0,m,s)$-net (i.e., $t=0$), then*

$$\frac{\hat{\nu}_N^{\mathrm{RQMC}} - \nu^*}{\sigma_N(x^*)} \xrightarrow{d} \mathrm{Normal}(0,1), \tag{3.19}$$

where $\sigma_N^2(x^*)$ is the variance of $\hat{g}_N^{\mathrm{RQMC}}(x^*)$. Moreover, in the latter case there exist positive constants c and C such that

$$c\,\frac{(\log_b N)^{s-1}}{N^3} \le \sigma_N^2(x^*) \le C\,\frac{(\log_b N)^{s-1}}{N^3} \tag{3.20}$$

as $m \to \infty$.

Proof. Let us fix $x \in X$. The assertion in case 1 and the first assertion in case 2 follow directly from (3.14) (which implies that Assumption A1 holds) and Propositions 2.2 and 2.5.

Consider now the random variable $W(x)$ defined as

$$W(x) := \frac{\hat{g}_N^{\mathrm{RQMC}}(x) - g(x)}{\sigma_N(x)}, \tag{3.21}$$

where $\sigma_N^2(x) := \mathrm{Var}[\hat{g}_N^{\mathrm{RQMC}}(x)]$. Here we resort to a key result on scrambled (t,m,s)-nets proved in [28]—building upon previous results in [35, 36]—that says that a CLT holds for pointwise estimators constructed with a scrambled $(0,m,s)$-net. Assumption D1 translates the conditions in [28] into our notation. It follows that, under D1, $W(x)$ converges in distribution to the standard normal for each $x \in S^*$; i.e., Assumption B2 holds and hence the conclusion follows from Theorem 2.10. □

Theorem 3.4 shows the benefits of using RQMC methods in optimization. Essentially, it says that, in the setting of case 2 of the theorem, the convergence rate

of optimal values is of order $[(\log N)^{s-1}/N^3]^{1/2}$, which asymptotically is better than the $N^{-1/2}$ obtained with standard Monte Carlo methods. This suggests that RQMC methods can be very efficacious for stochastic optimization. Note, however, that, strictly speaking, (3.19) applies only to the case where X is finite, since the assumption of finite support of $\boldsymbol{\xi}$ in the second case of Assumption A3 conflicts with the smoothness condition in Assumption D1. We will discuss the smoothness issue in more detail later. Note also that (3.19) is valid only for scrambled $(0, m, s)$-nets, which restricts the choice of the base b—as shown in [32], when $m \geq 2$ a $(0, m, s)$-net in base b can exist only if $b \geq s - 1$.

4. Two-stage stochastic programs. In this section we discuss the application of the results outlined in the previous sections to two-stage stochastic linear programs (see, e.g., [4] for a comprehensive discussion of this class of problems). We consider problems of the form

$$\min_{x \in X} c^t x + \mathbb{E}[Q(x, \boldsymbol{\xi})], \tag{4.1}$$

where X is a convex polyhedral set,

$$Q(x, \boldsymbol{\xi}) = \inf \left\{ q^t y \,:\, Wy \leq h - Tx,\ y \geq 0 \right\}, \tag{4.2}$$

and $\boldsymbol{\xi} = (h, T)$. As before, $\boldsymbol{\xi}$ is an s-dimensional random vector with arbitrary distribution. Let $G(x, \boldsymbol{\xi})$ denote the function $c^t x + Q(x, \boldsymbol{\xi})$; then we see that the the above problem falls in the framework of (1.1).

The use of Monte Carlo sampling to solve two-stage problems has been extensively studied in the literature, from both algorithmic (e.g., [17, 19, 26, 52]) and theoretical perspectives (see, for instance, [50] for a compilation of results).

Note that the function $Q(x, \boldsymbol{\xi})$ can be written in the form $Q(x, \boldsymbol{\xi}) = \tilde{Q}(h - Tx)$, where

$$\tilde{Q}(z) = \inf \left\{ q^t y \,:\, Wy \leq z,\ y \geq 0 \right\}. \tag{4.3}$$

By duality, we see that the function $\tilde{Q}(\cdot)$ can be represented in the form

$$\tilde{Q}(z) = \sup\{u^t z : W^t u \leq q,\ u \leq 0\}. \tag{4.4}$$

For the sake of simplicity we assume that (i) for every vector z the system $Wy \leq z$, $y \geq 0$, has a solution (the recourse is complete) and (ii) the system $W^t u \leq q,\ u \leq 0$, has a solution (dual feasibility). Under these assumptions, $\tilde{Q}(\cdot)$ is a finite-valued, piecewise linear convex function. This in turn implies that the function $G(x, \boldsymbol{\xi})$ is also piecewise linear convex in each argument and can be written as

$$G(x, \boldsymbol{\xi}) = \max_{k=1,\ldots,r} c^t x + (v^k)^t (h - Tx), \tag{4.5}$$

where $v^1, \ldots, v^r$ are the vertices of the polyhedron $\{u \,:\, W^t u \leq q,\ u \leq 0\}$. Furthermore, by standard subdifferential calculus we have that the subdifferential set of $G(x, \boldsymbol{\xi})$ with respect to x is given by

$$\partial_x G(x, \boldsymbol{\xi}) = \operatorname{conv}\{c - T^t v^k \,:\, G(x, \boldsymbol{\xi}) = c^t x + (v^k)^t (h - Tx),\ k = 1, \ldots, r\}, \tag{4.6}$$

where "conv" denote the convex hull of the set.

In the discussion that follows we assume that the feasibility set X is compact.

4.1. LHS results. In order to apply the results for LHS discussed in section 3.1, we need to verify that the corresponding assumptions are satisfied.

Assume momentarily that the matrix T is deterministic, so that $\boldsymbol{\xi} = h$. Consider Assumption A2. It follows from (4.6) that $\partial_x G(x, \boldsymbol{\xi})$ is uniformly bounded for all x and all $\boldsymbol{\xi}$, and thus, by a version of the mean-value theorem for subdifferentiable functions (see, e.g., [18, Theorem VI.2.3.3]), we conclude that A2 holds with a constant L such that $L(\cdot) \equiv L$. Next, notice that from (4.3) we have $G(x, \boldsymbol{\xi}) = \min\{q^t y : Wy \le \boldsymbol{\xi} - Tx,\ y \ge 0\}$. Thus, for any $\delta > 0$ we have $G(x, \boldsymbol{\xi} + \delta e_i) \le G(x, \boldsymbol{\xi})$; i.e., Assumption C1 holds.

It follows from the above discussion and from Theorem 3.1 that, if the moment-generating function of $G(x, \boldsymbol{\xi})$ is finite everywhere for all x, then given $\varepsilon > 0$ there exist constants $\tilde{K} > 0$ and $\tilde{\alpha} > 0$ such that

$$P\left(\operatorname{dist}(\hat{x}_N^{\mathrm{LHS}}, S^*) \ge \varepsilon\right) \le \tilde{K} e^{-\tilde{\alpha} N} \qquad \text{for all } N \ge 1.$$

Moreover, the exponent $\tilde{\alpha}$ is at least as large as the corresponding exponent obtained for standard Monte Carlo methods. This suggests that convergence under LHS will indeed be faster than under standard Monte Carlo methods.

As mentioned earlier, $G(\cdot, \boldsymbol{\xi})$ is piecewise linear. Thus, if $\boldsymbol{\xi}$ has finite support, then Assumption A3 holds, so from Theorem 3.1 we have

$$P\left(\hat{x}_N^{\mathrm{LHS}} \notin S^*\right) \le \tilde{K} e^{-\tilde{\alpha} N} \qquad \text{for all } N \ge 1.$$

It is fruitful to compare the above result with the i.i.d. case derived in [53]. Indeed, when problem (4.1) has a unique solution x^*, a slightly modified proof of Theorem 3.2 in [53] shows that there exists $\beta > 0$ such that

$$\limsup_{N \to \infty} \frac{1}{N} \log P\left(\hat{x}_N \ne x^*\right) = -\beta, \tag{4.7}$$

where $\hat{x}_N$ is the solution obtained with standard Monte Carlo methods. Moreover, the constant β is given by the minimum of a number of pointwise rates $-\gamma_x(\delta_0)$ in (2.1) (for a fixed $\delta_0 > 0$) over a *finite* number of x's. Since finite support of $\boldsymbol{\xi}$ implies that the moment-generating function of $G(x, \boldsymbol{\xi})$ is finite everywhere for all x, it follows from Proposition 6 in [9] that the pointwise rates $-\gamma_x(\delta_0)$ under LHS are no worse than under Monte Carlo methods. It follows that, when LHS is applied, an equation similar to (4.7) holds and the resulting constant β is *no worse* than under Monte Carlo methods. Since the rate in (4.7) is exact, we conclude that LHS can only improve upon Monte Carlo methods in this setting.

Next, we apply Theorem 3.3 to the present context. As seen above, Assumption A2 holds if T is deterministic. Note that A2 holds even if we allow T to be random (i.e., $\boldsymbol{\xi} = (h, T)$), as long as the distribution of $\boldsymbol{\xi}$ has bounded support, since in that case $\partial_x G(x, \boldsymbol{\xi})$ in (4.6) is uniformly bounded for all x. Moreover, under such a condition (4.5) clearly implies that $G(x, \boldsymbol{\xi})$ is bounded for each x. Theorem 3.3 then ensures that $\operatorname{dist}(\hat{x}_N^{\mathrm{LHS}}, S^*) \to 0$ w.p.1 and $\hat{\nu}_N^{\mathrm{LHS}} \to \nu^*$ w.p.1. Now suppose again that the distribution of $\boldsymbol{\xi}$ has finite support, so Assumption A3 holds. Then $\hat{x}_N^{\mathrm{LHS}} \in S^*$ w.p.1 for N large enough and $\hat{\nu}_N^{\mathrm{LHS}} \to \nu^*$ w.p.1. It follows that, if problem (1.1) has a unique optimal solution x^* and $G(x^*, \cdot)$ is not additive, then

$$\frac{\hat{\nu}_N^{\mathrm{LHS}} - \nu^*}{\sigma_N(x^*)} \ \xrightarrow{d}\ \mathrm{Normal}(0, 1),$$

where $\sigma_N(x^*) := \mathrm{Var}[\hat{g}_N^{\mathrm{LHS}}(x^*)] = N^{-1}C + o(N^{-1})$ for some positive constant C. Note that the nonadditivity assumption is reasonable in this setting, since at the optimal solution x^* typically it happens that the maximum in (4.5) is achieved by more than one k, so $G(x^*, \cdot)$ is not linear.

4.2. QMC results. We now apply the results from section 3.2 to the two-stage stochastic programming model described above. As seen above, a sufficient condition for Assumption A2 to hold is that the distribution of $\boldsymbol{\xi}$ have bounded support. In that case, we have from Theorem 3.4 that $\mathrm{dist}(\hat{x}_N^{\mathrm{RQMC}}, S^*) \to 0$ w.p.1 and $\hat{\nu}_N^{\mathrm{RQMC}} \to \nu^*$ w.p.1. When $\boldsymbol{\xi}$ has finite support (i.e., Assumption A3 holds), we obtain a stronger result, namely, that $\hat{x}_N^{\mathrm{RQMC}} \in S^*$ w.p.1 for N large enough.

The second part of Theorem 3.4—which deals with convergence rates—unfortunately is not applicable in this context. The reason, as pointed out in the discussion following Theorem 3.4, is that the smoothness condition stated in Assumption D1 cannot hold in this case, since $G(x, \cdot)$ is nondifferentiable for each x. Moreover, the assumption that $\boldsymbol{\xi}$ has finite support causes the inverse cumulative distribution function (cdfs) F_j^{-1} to be discontinuous. Note, however, that, as recognized in [28], the smoothness condition in Assumption D1 is only *sufficient* for the proof of the CLT for scrambled $(0, m, s)$-nets. Indeed, as the numerical experiments in section 5 show, it appears that the rates obtained with Theorem 3.4 are sometimes valid in the stochastic programming context considered above even though the smoothness condition does not apply.

5. Numerical experiments. To illustrate the ideas set forth in the previous sections, we discuss now some numerical experiments conducted with two small problems available in the literature. The first problem is `APL1P`, a model for electric power capacity expansion on a transportation network that was first described in [19]. The second problem is `LandS`, a modification of a simple problem in electrical investment planning originally presented in [29]. The modified version we study is the one discussed in [26].

APL1P. `APL1P` has 2 decision variables with 2 constraints (plus lower bound constraints) on the first stage and 9 decision variables with 5 constraints (plus lower bound constraints) on the second stage. The random variables appear on both the right-hand side and the technology matrix of the second stage. There are $s = 5$ independent random variables. The number of realizations per random variable yields a total of $4{\times}5{\times}4{\times}4{\times}4 = 1280$ scenarios. With current computing power, this problem can be easily solved exactly; nevertheless, we present the results with sampling because from that perspective the problem is ill-conditioned (cf. [54]), which means that the approximating solutions $\hat{x}_N$ are likely to vary with replications. That, in turn, ensures that the objective value estimators $\hat{\nu}_N$ do not correspond to the same solution—if they did, the analysis of rate of convergence would reduce to that of pointwise estimation. Thus, we view this case as a good test for the theoretical results presented in the paper.

We adopted the following methodology. We solved the approximating problem (1.3) by using samples generated with standard Monte Carlo methods, LHS, and randomized (t, s)-sequences in base 5—which, as discussed in section 3.2, is a form of RQMC. For each sampling scheme, we solved the problem with sample sizes equal to successive powers of the base, ranging from 5^2 to 5^6. The choice for such sample sizes was driven by two factors: (i) the restriction on the choice of the base in order for a $(0, m, s)$-net to exist ($b \geq s - 1$; cf. the discussion following Theorem 3.4), and (ii) the restriction on sample sizes to be powers of the base in order for a sequence to be a (t, m, s)-net. We feel that the two restrictions together would be rather limiting

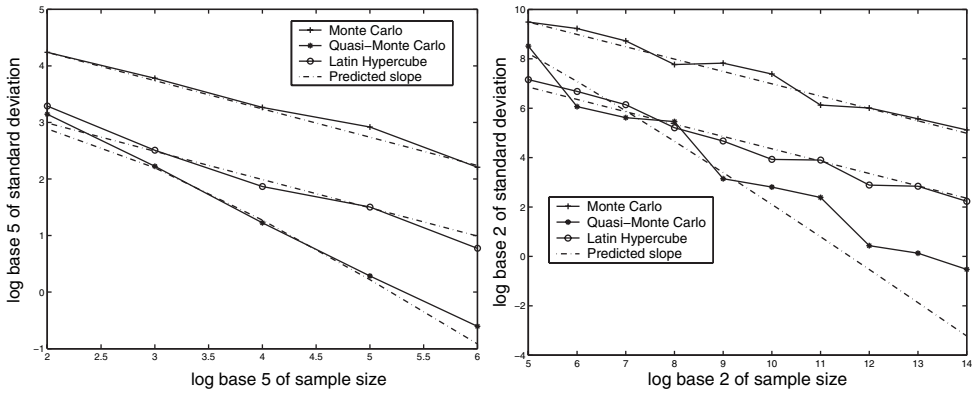


FIG. 5.1. *Rates of convergence for the APL1P problem using base* 5 *(left) and base* 2 *(right).*

in practice; thus, we also present results with base 2, which in addition yields faster generators [13].

For each sample size, twenty-five replications were run, and the standard deviation of the estimators $\hat{\nu}_N$ over these replications was calculated. All simulations used independent random streams. By plotting the logarithm of the standard deviation against the logarithm of the sample size, we can visualize the rate of convergence—for example, with standard Monte Carlo methods one expects to obtain a straight line with slope $-1/2$. We also calculated the *mean-squared error* of the estimators, but, since the results were very similar (in this problem, the bias was much smaller than the standard deviation), we chose not to display them.

The sampling approximation problems were solved in two steps: First, we used the SUTIL library [5] to generate the linear programs corresponding to each sampled problem. SUTIL can construct MPS files for Monte Carlo sampling approximations of two-stage stochastic linear programs; we modified the library slightly to incorporate LHS and randomized (t, s)-sequences, by using the publicly available routines developed in [13]. The resulting MPS files were fed into the software package Xpress–MP$^{\text{TM}}$ from Dash Optimization (available to us under the Academic Partnership Program).

Figure 5.1 shows the results. We can see that both Monte Carlo methods and LHS yield a convergence rate of $N^{-1/2}$, thus corroborating the results of [48] for Monte Carlo methods and of Theorem 3.3 for LHS. The rate for RQMC for both bases appears to be of order N^{-1} (although that is more evident with base 5), which is not as good as the rate in Theorem 3.4; a possible explanation is the absence of the smoothness assumed for that result. It is clear from the figure that the rate obtained with RQMC in either case is better than with both Monte Carlo methods and LHS. Note also that both LHS and RQMC yield estimators with smaller variance than Monte Carlo methods—even though the rate of convergence (in the case of LHS) is the same as that of Monte Carlo methods—and that the variance with RQMC is smaller than with LHS except for very small sample sizes.

LandS. The `LandS` problem has 4 decision variables on the first stage and 12 decision variables on the second stage. Randomness appears only on the right-hand side of the second stage, in the form of demand constraints. There are $s = 3$ independent random variables, each with 100 possible realizations. Thus, the total number of scenarios is 10^6.

The methodology we adopted was the same as in the `APL1P` case, except that used bases 3 and 2. Figure 5.2 shows the results. Again, we see that both Monte

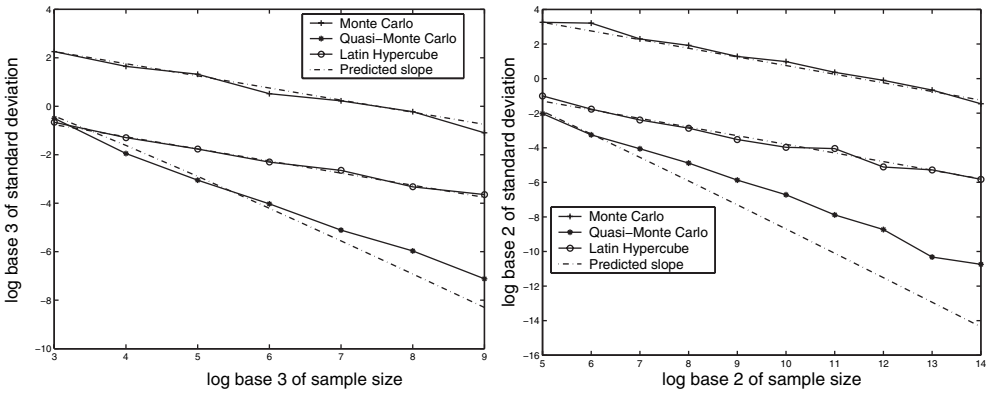


FIG. 5.2. *Rates of convergence for the* `LandS` *problem using base* 3 *(left) and base* 2 *(right).*

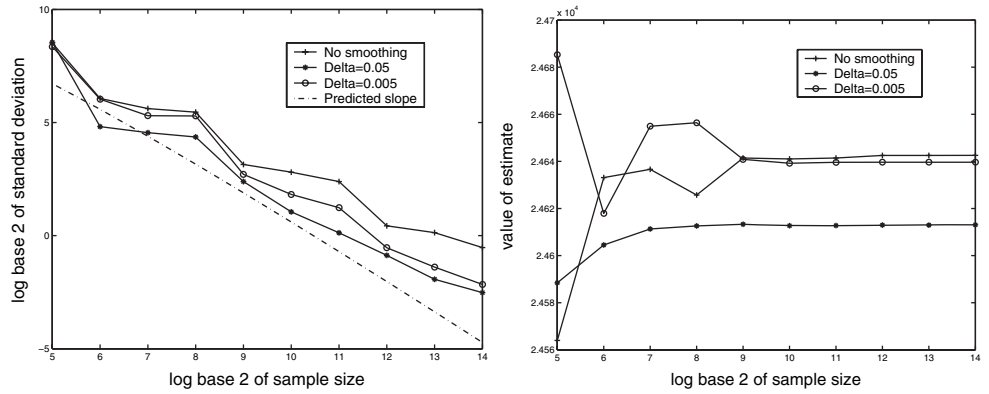


FIG. 5.3. *Rates of convergence (left) and values of the estimate* $\hat{\nu}_N$ *(right) for the APL1P problem under RQMC with smoothing, using base* 2.

Carlo methods and LHS yield a convergence rate of $N^{-1/2}$, and the rate for RQMC appears to be of order N^{-1}. As in the previous example, the RQMC rate is better than the Monte Carlo and LHS rates, both LHS and RQMC yield estimators with smaller variance than Monte Carlo methods, and the variance with RQMC is smaller than with LHS.

A brief study of smoothness. To check the role of the lack of smoothness in the convergence rates, we considered the effect of smoothing the inverse cdfs F_i^{-1}. This was accomplished by replacing F_i^{-1} with a smooth function F_i^{Δ} such that F_i^{-1} and F_i^{Δ} coincide everywhere except on a interval of size 2Δ around each discontinuity point.

Figures 5.3 and 5.4 depict the results. In general, such a smoothing procedure may introduce bias; this is clearly seen in case of the `APL1P` problem, where smoothing does not seem to help much, at least for base 2. With `LandS`, however, smoothing works perfectly—with $\Delta = 0.005$ we obtain the rate predicted by Theorem 3.4 without incurring virtually any bias, even though the theorem is not directly applicable in the absence of Assumption A3 (the smoothed distribution does not have finite support).

As another way to verify the effect of smoothness, we modified the `APL1P` problem by fitting continuous distributions to the discrete data. More precisely, we used Weibull distributions for each of the five random variables in the problem, in such a way that the mean and variance of the random variables were approximately the

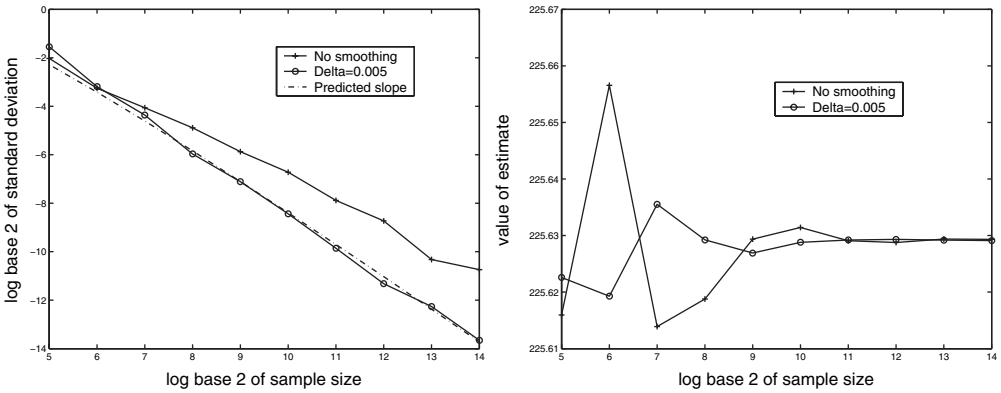


FIG. 5.4. *Rates of convergence (left) and values of the estimate $\hat{\nu}_N$ (right) for the* LandS *problem under RQMC with smoothing, using base* 2.

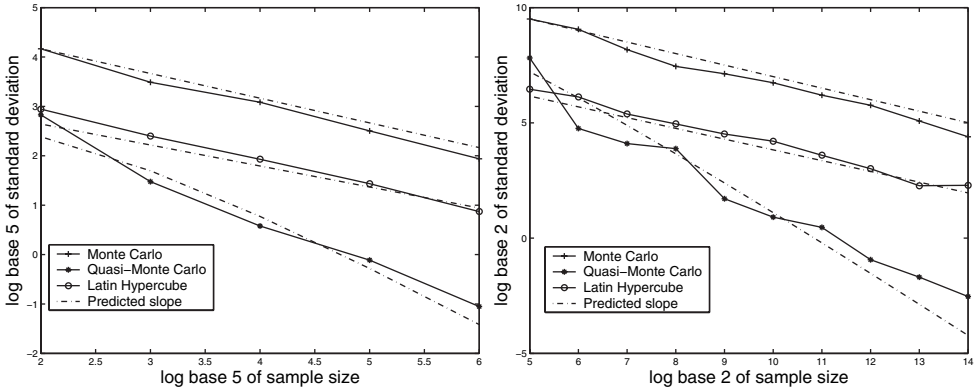


FIG. 5.5. *Rates of convergence for the modified APL1P problem with Weibull distributions, using base* 5 *(left) and base* 2 *(right).*

same as in the original data. Figure 5.5 depicts the results, again for bases 2 and 5. We see that the rates of convergence behave somewhat similarly to the discrete case, although the benefits of using RQMC seem slightly higher in the continuous case.

These results suggest that Theorem 3.4 may be valid under more general conditions than those we have used.

6. Conclusions. The theoretical and numerical results in this paper suggest that alternative sampling methods such as LHS and QMC can be very effective when solving stochastic optimization problems via sample average approximations. The effectiveness is measured in terms of rates of convergence of estimators of optimal solutions and of optimal values as functions of the sample size. The main contribution of the paper is establishing that rates of convergence for pointwise estimators (i.e., estimation of integrals) carry over to estimators of optimal values/solutions, which allows for the use of results for pointwise estimation available in the literature. In particular, the results in the paper show that, under appropriate conditions, it is possible to obtain a rate of convergence of order $[(\log N)^{s-1}/N^3]^{1/2}$ for the approximating optimal values $\hat{\nu}_N$, which asymptotically is much better than the $N^{-1/2}$ obtained with standard Monte Carlo methods.

Such results are very encouraging and at the same time raise some interesting issues for further investigation. One topic concerns the effect of smoothing on the rates of convergence when using RQMC—as discussed earlier, the "ideal" rate $[(\log N)^{s-1}/N^3]^{1/2}$ derived in Theorem 3.4 seems to require smoothness of the inverse cdf and of the objective function. However, our numerical results suggest that such conditions may not be necessary. Moreover, it is important to mention that Theorem 2.10 is valid regardless of any smoothness conditions. That is, if one shows that a CLT holds for RQMC under nonsmooth (or potentially discontinuous) functions with a certain rate, then Theorem 2.10 will ensure that under appropriate conditions the optimal value estimators $\hat{\nu}_N^{\mathrm{RQMC}}$ converge at the same rate. The result in [28] used in the proof of Theorem 3.4 is, however, the only CLT-type result available for RQMC, at least to the best of our knowledge.

On the other hand, our experiments also suggest that Theorem 3.4 is valid even when Assumption A3 does not hold. This is not surprising—as we mentioned earlier, it is possible that a functional version of Assumption B2 holds for the functional space $C(X)$ under RQMC, in which case the conditions of Assumption A3 would not be required; however, we are not aware of the existence of such a result.

It would also be interesting to study the precise effect of having multiple optimal solutions on the rates of convergence of optimal values—the main results we have obtained for that case under LHS and RQMC (Theorems 3.3 and 3.4) require uniqueness of the optimal solution. Such a task, however, is likely to require again a functional or at least multivariate version of Assumption B2 (we note that multivariate CLTs have been proved for LHS but not for RQMC).

Another important issue concerns the dimensionality of the problems. It is well known that the performance of RQMC methods worsens with the number of dimensions—indeed, it is easy to see that, when s is large, the term $[(\log_b N)^{s-1}/N^3]^{1/2}$ becomes smaller than $N^{-1/2}$ only for large N. For example, with $s = 30$ and $b = 2$ one needs $N \geq 2^{16}$ to get the benefits of the RQMC approach. This suggests that RQMC sampling should be used only with some of the random variables involved in the problem; however, determining which ones to select is a nontrivial issue. Research on this topic is underway.

Acknowledgments. We thank Shane Drew for help with the numerical experiments, Jeff Linderoth for assistance with the SUTIL library, and Mihai Anitescu and Alexander Shapiro for discussions on an early version of this paper. We also thank two anonymous referees and the associate editor for their comments.

REFERENCES

[1] T. G. Bailey, P. Jensen, and D. Morton, *Response surface analysis of two-stage stochastic linear programming with recourse*, Naval Res. Logist., 46 (1999), pp. 753–778.
[2] R. Bartle, *The Elements of Real Analysis*, 2nd ed., Wiley, New York, 1987.
[3] P. Billingsley, *Convergence of Probability Measures*, 2nd ed., Wiley, New York, 1999.
[4] J. R. Birge and F. Louveaux, *Introduction to Stochastic Programming*, Springer Ser. Oper. Res., Springer-Verlag, New York, 1997.
[5] J. Czyzyk, J. Linderoth, and J. Shen, *SUTIL: A Utility Library for Handling Stochastic Programs*, 2005. User's Manual. Software available at http://coral.ie.lehigh.edu/sutil.
[6] L. Dai, C. H. Chen, and J. R. Birge, *Convergence properties of two-stage stochastic programming*, J. Optim. Theory Appl., 106 (2000), pp. 489–509.
[7] G. B. Dantzig and P. W. Glynn, *Parallel processors for planning under uncertainty*, Ann. Oper. Res., 22 (1990), pp. 1–21.
[8] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed., Springer-Verlag, New York, 1998.

[9] S. S. Drew and T. Homem-de-Mello, *Some large deviations results for Latin Hypercube Sampling*, Department of Industrial Engineering and Management Sciences, Northwestern University, manuscript.
[10] J. Dupačová and R. J.-B. Wets, *Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems*, Ann. Statist., 16 (1988), pp. 1517–1549.
[11] G. Fishman, *Monte Carlo: Concepts, Algorithms and Applications*, Springer-Verlag, New York, 1997.
[12] B. L. Fox, *Strategies for Quasi-Monte Carlo*, Kluwer Academic, Norwell, MA, 2000.
[13] I. Friedel and A. Keller, *Fast generation of randomized low-discrepancy point sets*, in Monte Carlo and Quasi-Monte Carlo Methods, 2000 (Hong Kong), Springer, Berlin, 2002, pp. 257–273. Software available at http://www.multires.caltech.edu/software/libseq/.
[14] G. Gürkan, A. Y. Özge, and S. M. Robinson, *Sample-path solutions of stochastic variational inequalities*, Math. Program., 84 (1999), pp. 313–334.
[15] H. Heitsch and W. Römisch, *Scenario reduction algorithms in stochastic programming*, Comput. Optim. Appl., 24 (2003), pp. 187–206.
[16] J. L. Higle, *Variance reduction and objective function evaluation in stochastic linear programs*, INFORMS J. Comput., 10 (1998), pp. 236–247.
[17] J. L. Higle and S. Sen, *Stochastic decomposition: An algorithm for two stage linear programs with recourse*, Math. Oper. Res., 16 (1991), pp. 650–669.
[18] J. B. Hiriart-Urruty and C. Lemarechal, *Convex Analysis and Minimization Algorithms*, Vol. I, Springer-Verlag, Berlin, 1993.
[19] G. Infanger, *Monte Carlo (importance) sampling within a Benders decomposition algorithm for stochastic linear programs*, Ann. Oper. Res., 39 (1992), pp. 69–95.
[20] J. Kalagnanam and U. Diwekar, *An efficient sampling technique for off-line quality control*, Technometrics, 39 (1997), pp. 308–319.
[21] Y. M. Kaniovski, A. J. King, and R. J.-B. Wets, *Probabilistic bounds (via large deviations) for the solutions of stochastic programming problems*, Ann. Oper. Res., 56 (1995), pp. 189–208.
[22] A. J. King and R. T. Rockafellar, *Asymptotic theory for solutions in statistical estimation and stochastic programming*, Math. Oper. Res., 18 (1993), pp. 148–162.
[23] A. Kleywegt, A. Shapiro, and T. Homem-de-Mello, *The sample average approximation method for stochastic discrete optimization*, SIAM J. Optim., 12 (2001), pp. 479–502.
[24] M. Koivu, *Variance reduction in sample approximations of stochastic programs*, Math. Program., 103 (2005), pp. 463–485.
[25] A. M. Law and W. D. Kelton, *Simulation Modeling and Analysis*, 3rd ed., McGraw–Hill, New York, 2000.
[26] J. T. Linderoth, A. Shapiro, and S. J. Wright, *The empirical behavior of sampling methods for stochastic programming*, Ann. Oper. Res., 142 (2006), pp. 215–241.
[27] W. Loh, *On Latin hypercube sampling*, Ann. Statist., 24 (1996), pp. 2058–2080.
[28] W.-L. Loh, *On the asymptotic distribution of scrambled net quadrature*, Ann. Statist., 31 (2003), pp. 1282–1324.
[29] F. Louveaux and Y. Smeers, *Optimal investments for electricity generation: A stochastic model and a test problem*, in Numerical Techniques for Stochastic Optimization Problems, Y. Ermoliev and R. J.-B. Wets, eds., Springer-Verlag, Berlin, 1988, pp. 445–452.
[30] R. Lucchetti, G. Salinetti, and R. Wets, *Uniform convergence of probability measures: Topological criteria*, J. Multivariate Anal., 51 (1994), pp. 252–264.
[31] M. D. McKay, R. J. Beckman, and W. J. Conover, *A comparison of three methods for selecting values of input variables in the analysis of output from a computer code*, Technometrics, 21 (1979), pp. 239–245.
[32] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods*, SIAM, Philadelphia, 1992.
[33] A. B. Owen, *A central limit theorem for Latin hypercube sampling*, J. Roy. Statist. Soc. Ser. B, 54 (1992), pp. 541–551.
[34] A. B. Owen, *Randomly permuted* (t, m, s)*-nets and* (t, s)*-sequences*, in Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing (Las Vegas, NV, 1994), Lecture Notes in Statist. 106, Springer-Verlag, New York, 1995, pp. 299–317.
[35] A. B. Owen, *Monte Carlo variance of scrambled net quadrature*, SIAM J. Numer. Anal., 34 (1997), pp. 1884–1910.
[36] A. B. Owen, *Scrambled net variance for integrals of smooth functions*, Ann. Statist., 25 (1997), pp. 1541–1562.
[37] A. B. Owen, *Latin supercube sampling for very high-dimensional simulations*, ACM Trans. Modeling and Computer Simulation, 8 (1998), pp. 71–102.

[38] A. B. Owen, *Monte Carlo, quasi-Monte Carlo, and randomized quasi-Monte Carlo*, in Monte Carlo and Quasi-Monte Carlo Methods 1998, Claremont, CA, Springer-Verlag, Berlin, 2000, pp. 86–97.

[39] T. Pennanen, *Epi-convergent discretizations of multistage stochastic programs*, Math. Oper. Res., 30 (2005), pp. 245–256.

[40] T. Pennanen and M. Koivu, *Epi-convergent discretizations of stochastic programs via integration quadratures*, Numer. Math., 100 (2005), pp. 141–163.

[41] G. C. Pflug, *Scenario tree generation for multiperiod financial optimization by optimal discretization*, Math. Program. Ser. B, 89 (2001), pp. 251–271.

[42] G. C. Pflug, *Scenario Estimation and Generation*, tutorial presented at the X Stochastic Programming Conference, Tucson, AZ, 2004.

[43] E. L. Plambeck, B. R. Fu, S. M. Robinson, and R. Suri, *Sample-path optimization of convex stochastic performance functions*, Math. Program. Ser. B, 75 (1996), pp. 137–176.

[44] S. M. Robinson, *Analysis of sample-path optimization*, Math. Oper. Res., 21 (1996), pp. 513–528.

[45] W. Römisch, *Stability of stochastic programming problems*, in Handbook of Stochastic Optimization, A. Ruszczyński and A. Shapiro, eds., Elsevier Science Publishers B.V., Amsterdam, 2003.

[46] R. Y. Rubinstein and A. Shapiro, *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method*, Wiley, Chichester, England, 1993.

[47] L. Rüschendorf and V. de Valk, *On regression representations of stochastic processes*, Stochastic Process. Appl., 46 (1993), pp. 183–198.

[48] A. Shapiro, *Asymptotic analysis of stochastic programs*, Ann. Oper. Res., 30 (1991), pp. 169–186.

[49] A. Shapiro, *Asymptotic behavior of optimal solutions in stochastic programming*, Math. Oper. Res., 18 (1993), pp. 829–845.

[50] A. Shapiro, *Monte Carlo sampling methods*, in Handbook of Stochastic Optimization, A. Ruszczyński and A. Shapiro, eds., Elsevier Science Publishers B.V., Amsterdam, 2003.

[51] A. Shapiro, *On complexity of multistage stochastic programs*, Oper. Res. Lett., 34 (2006), pp. 1–8.

[52] A. Shapiro and T. Homem-de-Mello, *A simulation-based approach to two-stage stochastic programming with recourse*, Math. Program., 81 (1998), pp. 301–325.

[53] A. Shapiro and T. Homem-de-Mello, *On the rate of convergence of optimal solutions of Monte Carlo approximations of stochastic programs*, SIAM J. Optim., 11 (2000), pp. 70–86.

[54] A. Shapiro, T. Homem-de-Mello, and J. C. Kim, *Conditioning of convex piecewise linear stochastic programs*, Math. Program., 94 (2002), pp. 1–19.

[55] M. L. Stein, *Large sample properties of simulations using Latin hypercube sampling*, Technometrics, 29 (1987), pp. 143–151.

SIAM J. OPTIM.
Vol. 19, No. 2, pp. 552–571

STOCHASTIC PROGRAMS WITH FIRST-ORDER DOMINANCE CONSTRAINTS INDUCED BY MIXED-INTEGER LINEAR RECOURSE*

RALF GOLLMER†, FREDERIKE NEISE†, AND RÜDIGER SCHULTZ†

Abstract. We propose a new class of stochastic integer programs whose special features are dominance constraints induced by mixed-integer linear recourse. For these models, we establish closedness of the constraint set mapping with the underlying probability measure as a parameter. In the case of finite probability spaces, the models are shown to be equivalent to large-scale, block-structured, mixed-integer linear programs. We propose a decomposition algorithm for the latter and discuss computational results.

Key words. stochastic integer programming, stochastic dominance, mixed-integer optimization

AMS subject classifications. 90C15, 90C11, 60E15

DOI. 10.1137/060678051

1. Introduction.

Two-stage stochastic programming models are derived from random optimization problems with information constraints. In the present paper we start out from the following random mixed-integer linear program:

$$\min\{c^\top x + q^\top y \ : \ Tx + Wy = z(\omega),\ x \in X,\ y \in \mathbb{Z}_+^{\bar{m}} \times \mathbb{R}_+^{m'}\} \tag{1}$$

together with the information constraint that x must be selected without anticipation of $z(\omega)$. This leads to a two-stage scheme of alternating decision and observation: The decision on x is followed by observing $z(\omega)$, and then y is taken, thus depending on x and $z(\omega)$. Accordingly, x and y are called first- and second-stage decisions, respectively.

Assume that the ingredients of (1) have conformable dimensions, that W is a rational matrix, and that $X \subseteq \mathbb{R}^m$ is a nonempty polyhedron, possibly involving integer requirements to components of x.

The mentioned two-stage dynamics becomes explicit by the following reformulation of (1):

$$\begin{aligned} &\min_x \left\{ c^\top x + \min_y \{q^\top y \ : \ Wy = z(\omega) - Tx,\ y \in \mathbb{Z}_+^{\bar{m}} \times \mathbb{R}_+^{m'}\} \ : \ x \in X \right\} \\ &\qquad = \min_x \{c^\top x \ + \ \Phi(z(\omega) - Tx) \ : \ x \in X\}, \end{aligned} \tag{2}$$

where

$$\Phi(t) := \min\{q^\top y \ : \ Wy = t,\ y \in \mathbb{Z}_+^{\bar{m}} \times \mathbb{R}_+^{m'}\}. \tag{3}$$

*Received by the editors December 19, 2006; accepted for publication (in revised form) February 11, 2008; published electronically June 11, 2008. The first version of this paper was written while the third author was visiting the Centro de Modelamiento Matematico, Universidad de Chile, Santiago. Partial funding for this research was provided by the German Federal Ministry of Education and Research (BMBF) under grant 03-SCNIVG. We thank these institutions for their support.
http://www.siam.org/journals/siopt/19-2/67805.html

†Department of Mathematics, University of Duisburg-Essen, Campus Duisburg, Forsthausweg 2, D-47048 Duisburg, Germany (gollmer@math.uni-duisburg.de, neise@math.uni-duisburg.de, schultz@math.uni-duisburg.de).

The function Φ, called the value function of the mixed-integer linear program

$$\min\{q^\top y \ : \ Wy = t, \ y \in \mathbb{Z}_+^{\bar{m}} \times \mathbb{R}_+^{m'}\},$$

has been studied in parametric optimization. Under the assumptions

(A1) (complete recourse) $W(\mathbb{Z}_+^{\bar{m}} \times \mathbb{R}_+^{m'}) = \mathbb{R}^s$,

(A2) (sufficiently expensive recourse) $\{u \in \mathbb{R}^s \ : \ W^\top u \leq q\} \neq \emptyset$,

it holds that Φ is real-valued and lower semicontinuous on $\mathbb{R}^s$, i.e., $\liminf_{t_n \to t} \Phi(t_n) \geq \Phi(t)$ for all $t \in \mathbb{R}^s$ [2, 7].

In view of (2), the random optimization problem (1) gives rise to the family of random variables

$$\Big(c^\top x + \Phi(z(\omega) - Tx)\Big)_{x \in X}. \tag{4}$$

Thus every first-stage decision $x \in X$ induces a random variable $f(x,z) := c^\top x + \Phi(z(\omega) - Tx)$. Traditional two-stage stochastic programming aims at optimizing nonanticipative decisions, i.e., finding a best x, or in other words a best member in the family (4) of random variables. For the specification of best, statistical parameters reflecting mean and/or risk are employed. Early approaches in the literature used the expectation, leading to optimization problems

$$\min\{\mathbb{E}[f(x,z)] \ : \ x \in X\}. \tag{5}$$

Employing the weighted sum of $\mathbb{E}$ and some risk measure $\mathcal{R}$ leads to mean-risk models

$$\min\{\mathbb{E}[f(x,z)] + \rho \cdot \mathcal{R}[f(x,z)] \ : \ x \in X\} \qquad (\rho > 0 \text{ fixed}). \tag{6}$$

There is an extensive literature on structural analysis and algorithm design for this class of stochastic programs; see, for instance, [1, 5, 13, 19, 20, 23, 27, 29, 31, 32].

Here we take an alternative view. Rather than heading for best members of (4), we want to identify "acceptable" members and optimize over them. This leads to a new class of stochastic integer programs (see (8) below), whose structural analysis and algorithmic treatment is the aim of the present paper.

Stochastic dominance, an established concept in decision theory [14, 22, 24], provides a possibility to formalize the above-mentioned acceptability. In the present paper we deal with first-order stochastic dominance. A (real-valued) random variable X is said to be stochastically smaller in first order than a random variable Y ($\mathsf{X} \preceq_1 \mathsf{Y}$) iff $\mathbb{E}h(\mathsf{X}) \leq \mathbb{E}h(\mathsf{Y})$ for all nondecreasing functions h for which both expectations exist. An equivalent formulation reads as follows (see, e.g., [24]):

$$\mathsf{X} \preceq_1 \mathsf{Y} \quad \text{iff} \quad \mathbb{P}[\{\omega : \mathsf{X}(\omega) \leq \eta\}] \ \geq \ \mathbb{P}[\{\omega : \mathsf{Y}(\omega) \leq \eta\}] \quad \forall \eta \in \mathbb{R}. \tag{7}$$

Coming back to our two-stage random optimization problem (1) and the related family (4), we assume that some (random) benchmark cost profile $d(\omega)$ is given. We consider only those $x \in X$ acceptable for which the corresponding $f(x,z)$ is stochastically smaller in first order than the benchmark profile $d(\omega)$. Over all acceptable $x \in X$ we optimize some function $g : \mathbb{R}^m \to \mathbb{R}$. This leads to the following stochastic program with first-order dominance constraints induced by mixed-integer linear recourse:

$$\min\{g(x) \ : \ f(x,z) \ \preceq_1 \ d, \ x \in X\}. \tag{8}$$

Stochastic optimization problems with dominance constraints involving general random variables were pioneered in [10, 11, 12, 25]. These papers address structure,

stability, and algorithms for (8) if general random variables, enjoying suitable continuity, smoothness, or linearity properties in x and/or z, are placed instead of $f(x,z)$. The random variables $f(x,z)$, on the one hand, are more specific, since they are essentially given by the mixed-integer value function in (3). On the other hand, the results from [10, 11, 12, 25] are not applicable to our setting due to lacking continuity of Φ; recall the above-mentioned lower semicontinuity.

Our paper is organized as follows. In section 2 we elaborate a basic structural property of the constraint sets of (8) and draw some conclusions, among others, on the stability behavior of (8). Section 3 is devoted to algorithmic aspects. We prove the equivalence of (8) with a structured mixed-integer linear program if the underlying probability spaces are finite. Then we propose a decomposition algorithm for these models. In section 4 we report computational results with this algorithm. The paper concludes with section 5, where a model involving a weaker stochastic order is put into perspective, both theoretically and numerically.

2. Structure and stability. The stochastic program (8) is essentially governed by its constraint set. In what follows we establish some results concerning the basic well posedness of this set.

Let $\mathcal{P}(\mathbb{R}^s)$ and $\mathcal{P}(\mathbb{R})$ be the sets of all Borel probability measures on $\mathbb{R}^s$ and $\mathbb{R}$, respectively. By $\mu \in \mathcal{P}(\mathbb{R}^s)$ and $\nu \in \mathcal{P}(\mathbb{R})$ we denote the probability measures induced by the random variables $z(\omega)$ and $d(\omega)$, respectively. We fix ν and consider the multifunction $C : \mathcal{P}(\mathbb{R}^s) \to 2^{\mathbb{R}^m}$, where

$$C(\mu) \;:=\; \{x \in \mathbb{R}^m \;:\; f(x,z) \preceq_1 d,\, x \in X\}.$$

Moreover, we equip $\mathcal{P}(\mathbb{R}^s)$ with weak convergence of probability measures [3]. A sequence $\{\mu_n\}$ in $\mathcal{P}(\mathbb{R}^s)$ is said to converge weakly to $\mu \in \mathcal{P}(\mathbb{R}^s)$, written $\mu_n \xrightarrow{w} \mu$, if for any bounded continuous function $h : \mathbb{R}^s \to \mathbb{R}$ it holds that $\int_{\mathbb{R}^s} h(z)\mu_n(dz) \to \int_{\mathbb{R}^s} h(z)\mu(dz)$ as $n \to \infty$.

PROPOSITION 2.1. *Assume* (A1) *and* (A2). *Then C is a closed multifunction on $\mathcal{P}(\mathbb{R}^s)$. This means that for arbitrary $\mu \in \mathcal{P}(\mathbb{R}^s)$ and sequences $\mu_n \in \mathcal{P}(\mathbb{R}^s), x_n \in C(\mu_n)$, with $\mu_n \xrightarrow{w} \mu$ and $x_n \to x$, it follows that $x \in C(\mu)$.*

Proof. In view of $x_n \in C(\mu_n)$ and (7) it holds for all n that

$$\nu[d \le \eta] \;\le\; \mu_n[f(x_n,z) \le \eta] \qquad \forall \eta \in \mathbb{R}. \tag{9}$$

(The shorthand notations $d \le \eta$ and $f(x_n,z) \le \eta$ refer to the sets $\{d \in \mathbb{R} : d \le \eta\}$ and $\{z \in \mathbb{R}^s : f(x_n,z) \le \eta\}$, respectively.)

Denote that $M_\eta(x) := \{z \in \mathbb{R}^s \;:\; f(x,z) > \eta\}$. By (A1) and (A2), the function Φ, and hence $f(\cdot,\cdot)$, is lower semicontinuous. Therefore, $M_\eta(x)$ is open for all $\eta \in \mathbb{R}$ and all $x \in \mathbb{R}^m$. With the new notation, (9) says that for all n

$$\nu[d \le \eta] \;+\; \mu_n[M_\eta(x_n)] \;\le\; 1 \qquad \forall \eta \in \mathbb{R}. \tag{10}$$

Since $M_\eta(x)$ is open, the Portmanteau theorem (see [3, Theorem 2.1, pp. 11–12]) implies that

$$\mu[M_\eta(x)] \;\le\; \liminf_n \mu_n[M_\eta(x)] \qquad \forall \eta \in \mathbb{R}. \tag{11}$$

The lower semicontinuity of Φ yields

$$M_\eta(x) \;\subseteq\; \liminf_n M_\eta(x_n) \qquad \forall \eta \in \mathbb{R}. \tag{12}$$

Here "$\liminf_n$" denotes the set-theoretic limits inferior, i.e., the set of all points belonging to all but a finite number of sets $M_\eta(x_n)$. For fixed n, (12) and the lower semicontinuity of the probability measure (see [4, Theorem 4.1, p. 48] now imply that

$$\mu_n[M_\eta(x)] \;\le\; \mu_n[\liminf_k M_\eta(x_k)] \;\le\; \liminf_k \mu_n[M_\eta(x_k)] \qquad \forall \eta \in \mathbb{R}.$$

By taking the limits inferior with respect to n, we obtain

$$\begin{aligned} \liminf_n \mu_n[M_\eta(x)] &\le \liminf_n \liminf_k \mu_n[M_\eta(x_k)] \\ &\le \liminf_n \mu_n[M_\eta(x_n)] \end{aligned} \qquad \forall \eta \in \mathbb{R}. \tag{13}$$

For the last inequality we have picked the diagonal subsequence where $n = k$. By putting together (11) and (13) we arrive at

$$\mu[M_\eta(x)] \;\le\; \liminf_n \mu_n[M_\eta(x_n)] \qquad \forall \eta \in \mathbb{R}. \tag{14}$$

Taking the limits inferior with respect to n in (10) and observing (14) leads to

$$\nu[d \le \eta] \;+\; \mu[M_\eta(x)] \;\le\; \nu[d \le \eta] \;+\; \liminf_n \mu_n[M_\eta(x_n)] \;\le\; 1 \qquad \forall \eta \in \mathbb{R}.$$

This implies (see (10), (9), and (7)) that $f(x, z) \preceq_1 d$. By the closedness of X, $x_n \to x$, and $x_n \in X$ (for all n), we have $x \in X$. Altogether it follows that $x \in C(\mu)$, and the proof is complete. □

Remark 2.2 (about variable ν). Equipping $\mathcal{P}(\mathbb{R})$ with uniform convergence of distribution functions (Kolmogorov–Smirnov convergence), as, for instance, in [10], allows us to extend Proposition 2.1 to the multifunction $\bar{C} : \mathcal{P}(\mathbb{R}^s) \times \mathcal{P}(\mathbb{R}) \to 2^{\mathbb{R}^m}$, where $\bar{C}(\mu, \nu) := \{x \in \mathbb{R}^m : f(x, z) \preceq_1 d,\, x \in X\}$. Indeed, if ν_n converge to ν in the Kolmogorov–Smirnov sense, then $\nu_n[d \le \eta] \to \nu[d \le \eta]$ for all $\eta \in \mathbb{R}$, and the above proof readily extends.

Remark 2.3 (about weak convergence on $\mathcal{P}(\mathbb{R}^s)$). For a different class of random variables, [10] has established a closedness result for a dominance constraint of first order, where convergence on the counterpart space to $\mathcal{P}(\mathbb{R}^s)$ is given by a suitable discrepancy. Compared with [10], Proposition 2.1 applies to a more focused family of random variables (even allowing for discontinuities) with a weaker convergence notion on $\mathcal{P}(\mathbb{R}^s)$, namely, weak convergence of probability measures instead of convergence induced by the discrepancy in [10].

Proposition 2.1 in particular implies that $C(\mu)$ is a closed set for any $\mu \in \mathcal{P}(\mathbb{R}^s)$.

COROLLARY 2.4. *Assume* (A1) *and* (A2). *Then* $C(\mu)$ *is a closed subset of* $\mathbb{R}^m$ *for any* $\mu \in \mathcal{P}(\mathbb{R}^s)$.

The optimization problem (8) thus is well-posed in the sense that for, e.g., lower semicontinuous g, bounded X, and nonempty $C(\mu)$, the infimum is finite and is attained.

It is well known that continuity properties of constraint set mappings, such as the one established in Proposition 2.1, allow for direct conclusions regarding the stability of the related optimization problems. We next turn to such a conclusion.

Consider (8) as a parametric program where the probability distribution μ of the random variable $z(\omega)$ enters as a parameter:

$$P(\mu) \qquad \min\{g(x) \;:\; x \in C(\mu)\}.$$

Studying the stability of stochastic programs with respect to perturbations of the underlying probability distributions is motivated by the incomplete information on these distributions that is often met and by approximation issues in the context of computations; see [28] for a recent overview on stability analysis in stochastic programming.

PROPOSITION 2.5. *Assume* (A1) *and* (A2)*, that X is nonempty and compact, and that g is lower semicontinuous. Let $\bar{\mu} \in \mathcal{P}(\mathbb{R}^s)$ be such that $P(\bar{\mu})$ has an optimal solution. Then the optimal value function $\varphi(\mu) := \inf\{g(x) : x \in C(\mu)\}$ is lower semicontinuous at $\bar{\mu}$.*

Proof. Let $\mu_n \xrightarrow{w} \bar{\mu}$, and assume without loss of generality that $C(\mu_n) \neq \emptyset$ for all n. Otherwise, we would have $\varphi(\mu_n) = +\infty$, which does not interfere with the validity of $\liminf_n \varphi(\mu_n) \geq \varphi(\bar{\mu})$.

Let $\varepsilon > 0$ be arbitrarily fixed. Then there exist $x_n \in C(\mu_n)$ such that $g(x_n) \leq \varphi(\mu_n) + \varepsilon$. By the compactness of X there exists an accumulation point $\bar{x}$ of the x_n. By the closedness of $C(.)$ (Proposition 2.1), it follows that $\bar{x} \in C(\bar{\mu})$. Together with the lower semicontinuity of g, this implies that

$$\varphi(\bar{\mu}) \;\leq\; g(\bar{x}) \;\leq\; \liminf_n g(x_n) \;\leq\; \liminf_n \varphi(\mu_n) + \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, the proof is complete. □

Remark 2.6 (about approximation schemes). It is well known that approximation schemes based on discretization via conditional expectations [6, 18] or on empirical estimation [26, 33] often produce weakly converging sequences of discrete probability measures. Stability results such as Propositions 2.1 and 2.5 thus enable justification of numerical procedures that rely on problem solution for discrete measures, possibly belonging to weakly converging sequences. This links our stability analysis to the solution procedure of the next section. From Proposition 2.1 it follows that accumulation points of feasible solutions to the approximates are feasible solutions to the original problem. Proposition 2.5 yields the fact that limits of optimal values of the approximates never drop below the optimal value of the original.

3. Algorithmic treatment. In the present section we deal with algorithmic possibilities for (8) in case $z(\omega)$ and $d(\omega)$ follow discrete probability distributions with finitely many realizations. We start with establishing an equivalence between (8) and a large-scale, but structured, mixed-integer linear program.

PROPOSITION 3.1. *Let $z(\omega)$ and $d(\omega)$ in* (8) *follow discrete distributions with realizations z_l, $l = 1, \ldots, L$, and d_k, $k = 1, \ldots, K$, as well as probabilities π_l, $l = 1, \ldots, L$, and p_k, $k = 1, \ldots, K$, respectively. Let further $g(x) := g^\top x$ be linear and X be bounded. Assume* (A1) *and* (A2)*. Then there exists a constant* M *such that* (8) *is equivalent to the mixed-integer linear program*

$$\min\Big\{ g^\top x : \begin{array}{lcll} c^\top x + q^\top y_{lk} - d_k & \leq & \mathsf{M}\theta_{lk} & \forall l\ \forall k \\ Tx + W y_{lk} & = & z_l & \forall l\ \forall k \\ \sum_{l=1}^{L} \pi_l \theta_{lk} & \leq & \bar{d}_k & \forall k \\ x \in X,\ y_{lk} \in \mathbb{Z}_+^{\bar{m}} \times \mathbb{R}_+^{m'},\ \theta_{lk} \in \{0,1\} & & & \forall l\ \forall k \end{array} \Big\}, \tag{15}$$

where $\bar{d}_k := 1 - \nu[d \leq d_k]$, $k = 1, \ldots, K$.

Proof. By (7), the constraint $f(x,z) \preceq_1 d$ is equivalent to

$$\nu[d \leq \eta] \;\leq\; \mu[f(x,z) \leq \eta] \qquad \forall \eta \in \mathbb{R}.$$

As shown in [25] this is equivalent to

$$\nu[d \le d_k] \;\le\; \mu[f(x,z) \le d_k] \quad \text{for } k = 1, \dots, K. \tag{16}$$

The asserted constant M is put such that

$$\mathsf{M} \;>\; \sup\big\{c^\top x + \Phi(z_l - Tx) - d_k \;:\; x \in X,\, l \in \{1,\dots,L\},\, k \in \{1,\dots,K\}\big\}.$$

It has to be shown that the right-hand side above is finite. To this end, we employ the following growth property of Φ; see, [2, 7], for instance. Under (A1) and (A2) there exist constants $\alpha > 0, \beta > 0$ such that for all $t_1, t_2 \in \mathbb{R}^s$

$$|\Phi(t_1) - \Phi(t_2)| \;\le\; \alpha\|t_1 - t_2\| + \beta.$$

Moreover, (A2) implies that $\Phi(0) = 0$. This enables the following estimate:

$$\begin{aligned} |c^\top x + \Phi(z_l - Tx) - d_k| &\le |c^\top x| + |\Phi(z_l - Tx) - \Phi(0)| + |d_k| \\ &\le \|c\| \cdot \|x\| + \alpha\|z_l\| + \alpha\|T\| \cdot \|x\| + \beta + |d_k|. \end{aligned}$$

Since X is bounded, this verifies the finiteness of the above supremum.

By considering the complementary event on the right, we rewrite (16) as

$$\mu[f(x,z) > d_k] \;\le\; 1 - \nu[d \le d_k] \;=:\; \bar{d}_k \qquad \text{for } k = 1, \dots, K. \tag{17}$$

For any $k \in \{1, \dots, K\}$ we now consider the following sets:

$$S_1 \;:=\; \{x \in X \;:\; \mu[f(x,z) > d_k] \;\le\; \bar{d}_k\}$$

and

$$S_2 \;:=\; \left\{ x \in X : \begin{array}{ll} \exists \theta_l \in \{0,1\} & \\ \exists y_l \in \mathbb{Z}_+^{\bar{m}} \times \mathbb{R}_+^{m'},\; l = 1, \dots, L, & \\ \text{such that:} & \\ c^\top x + q^\top y_l - d_k & \le\; \mathsf{M}\theta_l \\ Tx + W y_l & =\; z_l \\ \sum_{l=1}^{L} \pi_l \theta_l & \le\; \bar{d}_k \end{array} \right\}.$$

We complete the proof by showing that $S_1 = S_2$ and begin with the inclusion $S_1 \subseteq S_2$.

Let $x \in S_1$, and consider $I := \big\{l \in \{1,\dots,L\} \;:\; c^\top x + \Phi(z_l - Tx) > d_k\big\}$. Then $\sum_{l \in I} \pi_l \le \bar{d}_k$, by the definition of S_1. Put $\theta_l := 1$, for $l \in I$, and $\theta_l := 0$, otherwise. This gives

$$\sum_{l=1}^{L} \pi_l \theta_l \;=\; \sum_{l \in I} \pi_l \;\le\; \bar{d}_k.$$

For $l \notin I$ we have $c^\top x + \Phi(z_l - Tx) \le d_k$. Hence there exists $y_l \in \mathbb{Z}_+^{\bar{m}} \times \mathbb{R}_+^{m'}$ fulfilling

$$c^\top x + q^\top y_l - d_k \;\le\; 0 \;=\; \mathsf{M}\theta_l \qquad \text{and} \qquad Tx + W y_l \;=\; z_l.$$

For $l \in I$ take $y_l \in \mathbb{Z}_+^{\bar{m}} \times \mathbb{R}_+^{m'}$ such that $Tx + Wy_l = z_l$ and $q^\top y_l = \Phi(z_l - Tx)$. By the selection of M we then have

$$c^\top x + q^\top y_l - d_k \;\leq\; \mathsf{M} \;=\; \mathsf{M}\theta_l.$$

This implies that $x \in S_2$.

To show $S_2 \subseteq S_1$ let $x \in S_2$, and consider $I := \big\{l \in \{1, \ldots, L\} \,:\, \theta_l = 0\big\}$. For each $l \in I$, then there exists a $y_l \in \mathbb{Z}_+^{\bar{m}} \times \mathbb{R}_+^{m'}$ such that

$$c^\top x + q^\top y_l - d_k \;\leq\; 0 \qquad \text{and} \qquad Tx + Wy_l \;=\; z_l.$$

Hence $c^\top x + \Phi(z_l - Tx) \;\leq\; d_k$ for all $l \in I$. Therefore

$$\big\{l \in \{1, \ldots, L\} \,:\, c^\top x + \Phi(z_l - Tx) > d_k\big\} \;\;\subseteq\;\; \big\{l \in \{1, \ldots, L\} \,:\, \theta_l = 1\big\}.$$

This yields

$$\mu[c^\top x + \Phi(z - Tx) \;>\; d_k] \;\;\leq\;\; \sum_{l \notin I} \pi_l \theta_l \;\;=\;\; \sum_{l=1}^{L} \pi_l \theta_l \;\;\leq\;\; \bar{d}_k.$$

Thus $x \in S_1$, and the proof is complete. □

As a mixed-integer linear program, the optimization problem from Proposition 3.1 clearly can be tackled by general-purpose mixed-integer linear programming software. With growing numbers L and K of scenarios of the data and the benchmark distributions, however, it can be expected that this approach will come to its limitations.

This motivates us to study decomposition of the model. By having in mind the L-shaped form of the constraint matrix that arises with discrete probability spaces in the traditional stochastic program (5) (see [5, 19, 27, 29]), similarities and differences come to the fore: The constraints

$$\begin{aligned} c^\top x + q^\top y_{lk} - d_k \;&\leq\; \mathsf{M}\theta_{lk} \quad \forall l\, \forall k, \\ Tx + Wy_{lk} \;&=\; z_l \qquad \forall l\, \forall k \end{aligned}$$

correspond to K blocks, each of them in L-shaped form. By the latter we mean that, for fixed k, there are no constraints explicitly interlinking variables y_{lk}, θ_{lk} belonging to different $l \in \{1, \ldots, L\}$. Linkage is established only by the omnipresent x-variables. These variables must not depend on l and must not depend on k. So they couple the K blocks above into a single L-shaped block. The constraints

$$\sum_{l=1}^{L} \pi_l \theta_{lk} \;\;\leq\;\; \bar{d}_k \qquad \forall k \tag{18}$$

provide linkage between variables belonging to different scenarios $l \in \{1, \ldots, L\}$ such that the full model no longer obeys the L-shaped structure.

Our basic algorithmic idea now is to generate lower bounds by a suitable relaxation, to generate upper bounds by a tailored feasibility heuristic, and to embed the two into a branch-and-bound scheme in the spirit of global optimization.

Lower bounds. Relaxation is carried out in a twofold manner: The nonanticipativity of x gets relaxed by introducing copies $x_l, l = 1, \ldots, L$, and the constraints (18) undergo Lagrangean relaxation. This is formalized as follows.

In the objective we put $x = \sum_{l=1}^{L} \pi_l x_l$, and for the constraints (18) we introduce Lagrangean multipliers $\lambda_k \geq 0$, $k = 1, \ldots, K$. The Lagrangean function then reads

$$\begin{aligned}\mathcal{L}(x, \theta, \lambda) &= \sum_{l=1}^{L} \pi_l \cdot g^\top x_l \ + \ \sum_{k=1}^{K} \lambda_k \left(\sum_{l=1}^{L} \pi_l \theta_{lk} - \bar{d}_k \right) \\ &= \sum_{l=1}^{L} \pi_l \cdot g^\top x_l \ + \ \sum_{l=1}^{L} \sum_{k=1}^{K} \lambda_k \cdot (\pi_l \theta_{lk} - \pi_l \bar{d}_k) \\ &= \sum_{l=1}^{L} \mathcal{L}_l(x_l, \theta_l, \lambda),\end{aligned}$$

where

$$\mathcal{L}_l(x_l, \theta_l, \lambda) \ := \ \pi_l \cdot g^\top x_l \ + \ \pi_l \sum_{k=1}^{K} \lambda_k \cdot (\theta_{lk} - \bar{d}_k).$$

This leads to the Lagrangean dual

$$\max\{D(\lambda) \ : \ \lambda \in \mathbb{R}^K_+\},$$

where

$$D(\lambda) \ = \ \min \left\{ \mathcal{L}(x, \theta, \lambda) : \begin{array}{lcll} c^\top x_l + q^\top y_{lk} - d_k & \leq & \mathsf{M}\theta_{lk} & \forall l \, \forall k \\ T x_l + W y_{lk} & = & z_l & \forall l \, \forall k \\ x_l \in X, \ y_{lk} \in \mathbb{Z}^{\bar{m}}_+ \times \mathbb{R}^{m'}_+, \ \theta_{lk} \in \{0,1\} & & & \forall l \, \forall k \end{array} \right\}.$$

The optimization problem behind $D(\lambda)$ now is separable in l, and we obtain

$$D(\lambda) \ = \ \sum_{l=1}^{L} \min \left\{ \mathcal{L}_l(x_l, \theta_l, \lambda) : \begin{array}{lcll} c^\top x_l + q^\top y_{lk} - d_k & \leq & \mathsf{M}\theta_{lk} & \forall k \\ T x_l + W y_{lk} & = & z_l & \forall k \\ x_l \in X, \ y_{lk} \in \mathbb{Z}^{\bar{m}}_+ \times \mathbb{R}^{m'}_+, \ \theta_{lk} \in \{0,1\} & & & \forall k \end{array} \right\}. \tag{19}$$

The Lagrangean dual is a nonsmooth concave maximization (or convex minimization) problem whose optimal value yields a lower bound to the optimal value of the mixed-integer linear program in Proposition 3.1. For solving the Lagrangean dual, bundle-trust algorithms from nonsmooth convex optimization, such as the conic bundle method [17], can be employed. Per iteration, these methods require the function value $D(\lambda)$ and one subgradient from $\partial D(\lambda)$. Here the above separability becomes essential, since it leads to a decomposition of the optimization problem behind $D(\lambda)$ into subproblems corresponding to the individual scenarios z_l, $l = 1, \ldots, L$.

In principle, the above lower bounding procedure can be improved by applying Lagrangean relaxation not only to (18) but also to the nonanticipativity of x that can be expressed by the system of identities $x_1 = x_2 = \cdots = x_L$. This, however, leads to a drastic increase of dimension in the Lagrangean dual, namely, from K to $K + m \cdot (L - 1)$. Recall that L is the number of data scenarios z_l, while K is the number of benchmark scenarios d_k. It is reasonable to assume that L, possibly stemming from past observations, is far bigger than K, possibly stemming from subjective

risk perception. Typically, L can be on the order of several hundreds or even thousands, while K is around 20 or even less. Compared with Lagrangean relaxation of nonanticipativity (see, for instance, [8]), the above dual bounding scheme thus has the advantage that the Lagrangean dual lives in a space of low dimension.

Upper bounds. An upper bound to the optimal value of (15) is computed by the following heuristic that aims at finding a feasible solution to (15). The input of the heuristic consists of the x_l-parts $\tilde{x}_l$ of optimal solutions to the single-scenario problems in (19) for optimal or nearly optimal λ.

ALGORITHM 3.2.

STEP 1:

Understand $\tilde{x}_l$, $l = 1, \ldots, L$, as proposals for x and pick a "reasonable candidate" $\bar{x}$, for instance, one arising most frequently or one with minimal $\mathcal{L}_l(x_l, \theta_l, \lambda)$, or average the $\tilde{x}_l$, $l = 1, \ldots, L$, and round to integers if necessary.

STEP 2:

Check whether the following problems are feasible for $l = 1, \ldots, L$:

$$\min \left\{ g^\top \bar{x} : \begin{array}{lll} c^\top \bar{x} + q^\top y_{lk} - d_k & \leq & \mathsf{M}\theta_{lk} \\ T\bar{x} + W y_{lk} & = & z_l \\ y_{lk} \in \mathbb{Z}_+^{\bar{m}} \times \mathbb{R}_+^{m'},\ \theta_{lk} \in \{0,1\},\ k = 1, \ldots, K & & \end{array} \right\}. \tag{20}$$

As soon as one of them fails to be feasible, $\bar{x}$ cannot be feasible for (15)*, and the heuristic stops with assigning the formal upper bound $+\infty$. Otherwise, go to Step* 3*.*

STEP 3:

Check whether the θ_{lk} found in (20) *fulfill*

$$\sum_{l=1}^{L} \pi_l \theta_{lk} \ \leq \ \bar{d}_k, \qquad k = 1, \ldots, K.$$

If yes, then a feasible solution to (15) *is found. The heuristic stops with the upper bound $g^\top \bar{x}$. Otherwise, go to Step* 4*.*

STEP 4:

Solve for each $l = 1, \ldots, L$:

$$\min \left\{ \sum_{k=1}^{K} \theta_{lk} : \begin{array}{lll} c^\top \bar{x} + q^\top y_{lk} - d_k & \leq & \mathsf{M}\theta_{lk} \\ T\bar{x} + W y_{lk} & = & z_l \\ y_{lk} \in \mathbb{Z}_+^{\bar{m}} \times \mathbb{R}_+^{m'},\ \theta_{lk} \in \{0,1\},\ k = 1, \ldots, K & & \end{array} \right\}.$$

Go to Step 5*.*

STEP 5:

Repeat the test from Step 3 *with the θ_{lk} found in Step* 4*. If the test is positive, then the heuristic stops with the upper bound $g^\top \bar{x}$. Otherwise, the heuristic stops without a feasible solution to* (15) *and assigns the formal upper bound $+\infty$.*

The purpose of Step 4 is to "push down" the θ_{lk} in order to fulfill (18). The implementation is such that Step 4 just continues with the feasible solution found in Step 2. The impact of Step 4 is particularly striking, if in Step 2 the θ_{lk} were "poorly" selected such that the test in Step 3 fails, although $\bar{x}$ is feasible for (15), with different θ_{lk}.

Branch-and-bound scheme. The bounding procedures developed above are integrated into a branch-and-bound scheme where branching is accomplished by partitioning the set X with increasing granularity. Linear inequalities are used for this purpose to maintain the mixed-integer linear description of problems.

This results in the following algorithm. By $\mathbf{P}$ we denote a list of problems, and $\varphi_{LB}(P)$ is a lower bound for the optimal value of $P \in \mathbf{P}$. Moreover, $\bar{\varphi}$ denotes the currently best upper bound to the optimal value of (15), and $X(P)$ is the element in the partition of X belonging to P.

ALGORITHM 3.3.

STEP 1 (INITIALIZATION):

Let $\mathbf{P} := \{(15)\}$ *and* $\bar{\varphi} := +\infty$.

STEP 2 (TERMINATION):

If $\mathbf{P} = \emptyset$, *then the* $\bar{x}$ *that yielded* $\bar{\varphi} = g^\top \bar{x}$ *is optimal.*

STEP 3 (BOUNDING):

Select and delete a problem P *from* $\mathbf{P}$. *Compute a lower bound* $\varphi_{LB}(P)$ *by the bounding procedure developed above, and apply Algorithm* 3.2 *to find a feasible point* $\bar{x}$ *of* P.

STEP 4 (PRUNING):

If $\varphi_{LB}(P) = +\infty$ *(infeasibility of a subproblem in* (19)*) or* $\varphi_{LB}(P) > \bar{\varphi}$ *(inferiority of* P*), then go to Step* 2.

If $\varphi_{LB}(P) = g^\top \bar{x}$ *(optimality for* P*), then check whether* $g^\top \bar{x} < \bar{\varphi}$. *If yes, then* $\bar{\varphi} := g^\top \bar{x}$. *Go to Step* 2.

If $g^\top \bar{x} < \bar{\varphi}$, *then* $\bar{\varphi} := g^\top \bar{x}$.

STEP 5 (BRANCHING):

Create two new subproblems by partitioning the set $X(P)$. *Add these subproblems to* $\mathbf{P}$, *and go to Step* 2.

Generally speaking, the branching in Step 5 is accomplished by applying linear inequalities to maintain representation of subproblems as mixed-integer linear programs. In practice, however, these inequalities usually correspond to ranges of components of variables. For continuous variables, tolerances are used to avoid endless branching with finer and finer granularity.

4. Computations. In the following we report computational results for Algorithm 3.3 applied to test instances from power planning. The first group of instances refers to the optimal management of a dispersed generation (DG) system run by a power utility in Germany; see [16] for a detailed model description. The instances of the second group are inspired by an early stochastic program from the literature, the investment planning problem for electricity generation of [21].

4.1. DG system. The system consists of five engine-based cogeneration (CG) stations, producing power and heat simultaneously, twelve wind turbines, and one hydroelectric power plant. The CG stations include eight gas boilers, nine gas motors, and one gas turbine, and each is equipped with a thermal storage and a cooling device. While the heat is distributed locally, the electricity is fed into the global distribution network. The cost minimal operation of this system with respect to relevant

TABLE 1
Dimensions of mixed-integer linear programming equivalents.

Number of	10 scenarios	20 scenarios	30 scenarios	50 scenarios
Boolean variables	299159	596799	894439	1489719
Continuous variables	283013	564613	846213	1409413
Constraints	742648	1481568	2220488	3698328

technical constraints and fulfillment of heat and power demand can be formulated as a mixed-integer linear program, with on-off decisions for the generation units as the source of integrality. With a planning horizon of 24 hours, divided into quarter-hourly subintervals, this (still deterministic) model has about 17500 variables (9000 Boolean and 8500 continuous) and 22000 constraints.

The optimization problem is influenced by stochasticity on the production side, where the in-feed from renewable resources is not known with certainty, as well as on the consumer side, where demand of electrical and thermal energy are uncertain. The problem turns into a random mixed-integer linear problem, a specification of (1).

Assuming that the uncertainty-prone data are known for the first four hours of the planning horizon leads to a two-stage stochastic program with the decisions belonging to these first four hours as the first stage. For a more detailed description of the arising stochastic program and results on purely expectation-based and mean-risk specifications of (6), see [16, 30].

To derive a benchmark profile d in (8) we first consider $f(\hat{x}, z)$, where $\hat{x}$ denotes an optimal solution to the expectation model (5). With heuristically selected benchmark values, the $f(\hat{x}, z)$ then are clustered around these values, and the probability of each benchmark value arises as the sum of the probabilities of the members in its cluster. Further problem instances were derived by fixing the probabilities and increasing the values of d successively.

A meaningful objective function g is to count the number of start-ups over all units and time steps in the first stage. This number serves as a measure for the abrasion of the DG units. Then the dominance-constrained model minimizes abrasion of units over all generation policies incurring costs that, in a stochastic sense, do not exceed the given benchmark profile.

We report results for instances with $K = 4$ benchmark scenarios and $L = 10$ up to 50 scenarios for heat and power demand. The deterministic equivalents according to Proposition 3.1 then finally are truly large-scale, as seen in Table 1, and can hardly be handled with mixed-integer solvers such as CPLEX [9].

In Tables 2–5 computations for these equivalents with CPLEX are compared to computations made with the implementation ddsip.vSD of Algorithm 3.3 derived in section 3. Problems were solved on a Linux PC with a 3.2 GHz Pentium processor and 2 GB RAM. As a stopping criterion we used a time limit of eight hours.

In all tables the benchmark costs increase successively from instance 1 to instance 5. This means that the dominance constraints get easier to fulfill. As one would expect, this affects the needed numbers of start-ups positively. They decrease with increasing reference values, which is reported in the column "Upper Bound," where the objective value of the current best solution is displayed. The corresponding best lower bound can be found in the column "Lower Bound."

In every table we show the status of the optimization for different points in time. Usually the first two points are the times where either the decomposition method or CPLEX finds the first feasible solution. Also for the time limit of eight hours the

TABLE 2
Results for instances with 10 *data scenarios and* 4 *benchmark scenarios.*

Number of scenarios	Inst.	Benchmarks		Time (sec.)	CPLEX		ddsip.vSD	
		Probability	Benchmark value		Upper bound	Lower bound	Upper bound	Lower bound
10	1	0.12	2895000	430.43	–	29	29	15
		0.21	4851000	899.16	–	29	29	29
		0.52	7789000	15325.75	29	29	29	29
		0.15	10728000					
	2	0.12	2900000	192.48	–	27	28	15
		0.21	4860000	418.90	28	28	28	15
		0.52	7800000	802.94	28	28	28	28
		0.15	10740000					
	3	0.12	3000000	144.63	–	21	21	12
		0.21	5000000	428.61	21	21	21	18
		0.52	8000000	678.79	21	21	21	21
		0.15	11000000					
	4	0.12	3500000	164.34	–	11	13	10
		0.21	5500000	818.26	–	12	13	13
		0.52	8500000	28800.00	13	12	13	13
		0.15	11500000					
	5	0.12	4000000	171.52	–	7	8	8
		0.21	6000000	3304.02	8	8	8	8
		0.52	9000000					
		0.15	12000000					

TABLE 3
Results for instances with 20 *data scenarios and* 4 *benchmark scenarios.*

Number of scenarios	Inst.	Benchmarks		Time (sec.)	CPLEX		ddsip.vSD	
		Probability	Benchmark value		Upper bound	Lower bound	Upper bound	Lower bound
20	1	0.105	2895000	306.89	–	29	29	12
		0.1	4851000	1151.95	–	29	29	29
		0.69	7789000	9484.97	29	29	29	29
		0.105	10728000					
	2	0.105	2900000	703.91	–	27	28	18
		0.1	4860000	1744.75	28	28	28	26
		0.69	7800000	1916.06	28	28	28	28
		0.105	10740000					
	3	0.105	3000000	305.84	–	21	21	10
		0.1	5000000	1682.93	21	21	21	19
		0.69	8000000	2138.94	21	21	21	21
		0.105	11000000					
	4	0.105	3500000	425.98	–	11	13	9
		0.1	5500000	2213.08	–	12	13	13
		0.69	8500000	11236.31	–	12 oom.*	13	13
		0.105	11500000					
	5	0.105	4000000	447.33	–	8	8	8
		0.1	6000000	5599.99	9	8	8	8
		0.69	9000000	7840.09	9	8 oom.*	8	8
		0.105	12000000					

objective values and the best bounds are given for each solver, unless optimality was proven earlier.

For test instances with 20 or 30 scenarios CPLEX sometimes stops before reaching a first feasible solution, because the available memory is exceeded (marked by "oom."). In these cases only the lower bounds already found before the memory error occurred are displayed.

TABLE 4
Results for instances with 30 *data scenarios and* 4 *benchmark scenarios.*

Number of scenarios	Inst.	Benchmarks		Time (sec.)	CPLEX		ddsip.vSD	
		Probability	Benchmark value		Upper bound	Lower bound	Upper bound	Lower bound
30	1	0.085	2895000	473.27	–	28	29	12
		0.14	4851000	1658.02	–	29	29	29
		0.635	7789000	3255.99	–	29 oom.*	29	29
		0.14	10728000					
	2	0.085	2900000	1001.53	–	26	28	18
		0.14	4860000	2694.93	–	27	28	28
		0.635	7800000	3372.24	–	27 oom.*	28	28
		0.14	10740000					
	3	0.085	3000000	469.93	–	17	23	10
		0.14	5000000	3681.15	–	18 oom.*	21	20
		0.635	8000000	28800.00	–	–	21	20
		0.14	11000000					
	4	0.085	3500000	618.21	–	10	14	8
		0.14	5500000	3095.02	–	11 oom.*	14	10
		0.635	8500000	28800.00	–	–	14	13
		0.14	11500000					
	5	0.085	4000000	672.73	–	7	8	8
		0.14	6000000	8504.88	–	8 oom.*	8	8
		0.635	9000000					
		0.14	12000000					

TABLE 5
Results for instances with 50 *data scenarios and* 4 *benchmark scenarios.*

Number of scenarios	Inst.	Benchmarks		Time (sec.)	CPLEX		ddsip.vSD	
		Probability	Benchmark value		Upper bound	Lower bound	Upper bound	Lower bound
50	1	0.09	2895000	745.87	–	–	29	11
		0.135	4851000	2534.21	–	–	29	29
		0.67	7789000					
		0.105	10728000					
	2	0.09	2900000	1549.22	–	–	28	18
		0.135	4860000	4168.89	–	–	28	28
		0.67	7800000					
		0.105	10740000					
	3	0.09	3000000	756.06	–	–	23	10
		0.135	5000000	28800.00	–	–	21	20
		0.67	8000000					
		0.105	11000000					
	4	0.09	3500000	975.20	–	–	15	8
		0.135	5500000	28800.00	–	–	13	12
		0.67	8500000					
		0.105	11500000					
	5	0.09	4000000	1150.95	–	–	8	8
		0.135	6000000					
		0.67	9000000					
		0.105	12000000					

With 50 data scenarios the deterministic equivalents become so large that the available memory is not sufficient to build up the model (lp-) file used by CPLEX, preventing optimization with CPLEX for these instances. Therefore the last table reports only best values and lower bounds calculated with the decomposition method ddsip.vSD.

TABLE 6
Results for instances with 10 *data scenarios,* 4 *benchmark scenarios, and standard branching.*

Number of scenarios	Inst.	Benchmarks		Time (sec.)	CPLEX		ddsip.vSD	
		Probability	Benchmark value		Upper bound	Lower bound	Upper bound	Lower bound
10	1	0.12	2895000	1348.95	–	29	29	18
		0.21	4851000	15325.75	29	29	29	22
		0.52	7789000	28800.00	29	29	29	23
		0.15	10728000					
	2	0.12	2900000	273.78	–	27	28	14
		0.21	4860000	418.90	28	28	28	14
		0.52	7800000	28800.00	28	28	28	22
		0.15	10740000					
	3	0.12	3000000	192.45	–	21	21	12
		0.21	5000000	428.61	21	21	21	12
		0.52	8000000	28800.00	21	21	21	16
		0.15	11000000					
	4	0.12	3500000	227.44	–	11	13	10
		0.21	5500000	2593.35	18	12	13	10
		0.52	8500000	28800.00	13	13	13	11
		0.15	11500000					
	5	0.12	4000000	225.91	–	7	8	8
		0.21	6000000	3304.02	8	8	8	8
		0.52	9000000					
		0.15	12000000					

Our computations show that for all instances the decomposition method reaches the first feasible solution faster then CPLEX does. In most cases this is already an optimal solution. In the computations dealing with 30 and 50 scenarios, the superiority of the decomposition method over general-purpose solvers becomes particularly evident. For 30 scenarios CPLEX can't provide any feasible solution and for 50 scenarios even no lower bound, while ddsip.vSD is able to solve almost all problems.

In reaching these results, proper branching strategies in Step 5 of Algorithm 3.3 turned out to be essential. In the above computations branching priority was given to Boolean variables arising in the objective of (15). As a comparison consider Table 6 reporting computations for the 10-scenario instances with an alternative, rather standard, branching strategy. Here a first-stage variable was selected for branching for which the optimal solutions to the single-scenario problems in (19), with optimal or nearly optimal λ, violated nonanticipativity the most, with violation measured by a suitable entity reflecting dispersion. The performance gains of the first branching strategy over the second are quite remarkable.

4.2. Investment planning. The investment planning problems for electricity generation that form our second group of test instances are inspired by [21]. We consider two-stage versions of the multistage model there and add integrality requirements to the first stage. This leads to a two-stage mixed-integer linear stochastic program where, in the first stage, decisions on capacity expansions for different generation technologies under budget constraints and supply guarantee are made. We assume that these decisions reflect indivisibilities (generation units) and hence are integer-valued. The second stage concerns the minimization of production costs for electricity under the constraints that electricity demand is met and the available capacity is not exceeded.

The electricity demand is captured by a load duration curve assigning to each duration $\tau \in \mathbb{R}_+$ the minimum load to be covered over time spans adding up to

TABLE 7
Dimensions of mixed-integer linear programming equivalents.

Number of	20 scenarios	50 scenarios	100 scenarios	500 scenarios
Boolean and integer variables	404	1004	2004	10004
Continuous variables	38400	96000	192000	960000
Constraints	11622	29022	58022	290022

τ. This is where uncertainty enters, since in practice load durations are typically available only stochastically. The model uses step function approximations for load duration curves. So each data scenario is represented by a (finite) step function.

The aim of the optimization is cost minimization where costs are incurred by the expansion decisions of the first stage and the production levels of the second stage. Together with the random load durations, this leads to a random optimization problem which is a specification of (1).

The benchmarks were constructed in a similar way as in subsection 4.1. With first-stage decisions x fixed to "reasonable" values, the $f(x, z)$ were clustered around heuristically selected benchmark values, whose probabilities were obtained as probabilities of the cluster sets.

As objective function g we considered the capacity expansion of one of the different technologies, possibly one least desired for environmental reasons. The dominance-constrained stochastic program then minimizes expansion of this capacity over all expansion policies whose costs do not exceed the benchmark profile in terms of first-order stochastic dominance.

We report results for instances with $K = 3$ up to 20 benchmark scenarios and $L = 20$ up to 500 scenarios for load duration. Deterministic equivalents according to Proposition 3.1 again become pretty large-scale. Table 7 shows dimensions for $K = 20$ and the different L.

Table 8 summarizes our computations for the investment planning instances. Again, a Linux PC with a 3.2 GHz Pentium processor and 2 GB RAM was used. The time limit was set to one hour.

The first column indicates three principal problem instances marked by their optimal values. (Let us remark that all test instances were constructed in such a way that their optimal values were known in advance.) The next two columns list the numbers K of benchmark and L of data scenarios. The remaining columns list lower and upper bounds obtained when applying CPLEX [9] and our implementation ddsip.vSD of Algorithm 3.3. Time entries deviating from the limit of 1 h indicate that the instance was solved to optimality within this span.

It becomes evident that, at the investment planning instances, Algorithm 3.3 is superior to applying a general-purpose solver such as CPLEX. Although we have experimented with various time limits and parameter settings in CPLEX, such as "emphasize integer feasibility," we were unable to improve the CPLEX results for upper bounds. The instance 0/3/100 (optimal-value/K/L), for example, was solved to optimality by CPLEX after more than three hours only. For the instance 0/10/100, as another example, CPLEX did not find a feasible solution even after four days of computing time.

5. Alternative lower bounding: Increasing convex order. Let X and Y be real-valued random variables. Then X is said to be stochastically smaller than Y in increasing convex order ($\mathsf{X} \preceq_{icx} \mathsf{Y}$) iff $\mathbb{E}h(\mathsf{X}) \leq \mathbb{E}h(\mathsf{Y})$ for all nondecreasing convex functions h for which both expectations exist (cf., e.g., [24]). This gives rise to the

TABLE 8
Results for investment planning instances.

			CPLEX			ddsip.vSD		
Optimal value	K	L	Upper bound	Lower bound	Time	Upper bound	Lower bound	Time
0	3	20	0	0	17 s	0	0	69 s
		50	0	0	2712 s	0	0	138 s
		100	–	0	1 h	0	0	718 s
		500	–	0	1 h	0	0	2162 s
	10	20	0	0	3197 s	0	0	70 s
		50	–	0	1 h	0	0	588 s
		100	–	0	1 h	0	0	2327 s
		500	–	0	1 h	8	0	1 h
	20	20	–	0	1 h	0	0	368 s
		50	–	0	1 h	0	0	2395 s
		100	–	0	1 h	23	0	1 h
		500	–	0	1 h	166	0	1 h
1	3	20	1	1	15.9 s	1	1	659 s
		50	–	0	1 h	1	1	1244 s
		100	–	0	1 h	2	1	1 h
		500	–	0	1 h	3	1	1 h
	10	20	–	0.771	1 h	1	1	1116 s
		50	–	0	1 h	4	1	1 h
		100	–	0	1 h	2	1	1 h
		500	–	0	1 h	8	0	1 h
	20	20	–	0	1 h	1	1	3039 s
		50	–	0	1 h	12	1	1 h
		100	–	0	1 h	2	0	1 h
		500	–	0	1 h	170	0	1 h
100	3	20	100	100	11.31 s	101	72	1 h
		50	–	76.6	1 h	104	38	1 h
		100	–	27	1 h	100	33	1 h
		500	–	0	1 h	111	16	1 h
	10	20	–	99.5	1 h	101	85	1 h
		50	–	40	1 h	102	56	1 h
		100	–	27	1 h	101	44	1 h
		500	–	0	1 h	102	44	1 h
	20	20	–	72	1 h	103	92	1 h
		50	–	40	1 h	207	80	1 h
		100	–	27	1 h	160	67	1 h
		500	–	0	1 h	184	54	1 h

following counterpart model to (8):

$$\min\{g(x) \ : \ f(x,z) \preceq_{icx} \ d, \ x \in X\}. \tag{21}$$

Since $f(x,z) \preceq_1 d$ implies $f(x,z) \preceq_{icx} d$ to hold, but not vice versa, problem (21) is a relaxation to (8), thus providing an alternative means for lower bounding of (8). In what follows we address this issue theoretically and numerically. We begin with reviewing some theoretical facts about (21) that are taken from [15], where structural properties and algorithmic aspects of (21) were explored in detail.

The constraint $f(x,z) \preceq_{icx} d$ can be equivalently expressed as

$$\mathbb{E}_z[f(x,z)-\eta]_+ \ \leq \ \mathbb{E}_d[d-\eta]_+ \quad \forall \eta \in \mathbb{R}. \tag{22}$$

This paves the way for an analogous result to Proposition 2.1: Assuming (A1), (A2), and finite first moments for z and d, the constraint set of (21) defines a closed multifunction in the probability measure underlying z. Moreover, finiteness of $\mathbb{E}_z$ in (22)

provided, (22) defines a convex set if $f(.,z)$ is convex, a situation encountered in two-stage stochastic programs with linear recourse. As soon as integrality requirements on y enter problem (1), however, the convexity of $f(.,z)$ is lost in general, and it has to be expected that (21) has nonconvex constraints. Since the constraint set mapping of (21) is a closed multifunction, a stability result for (21) in the spirit of Proposition 2.5 is valid. This justifies numerical treatment of (21) for finite probability spaces, tacitly assuming that, if necessary, continuous probability distributions of z (and d) are replaced by finite discrete ones that are "sufficiently close" in terms of suitable topologies on spaces of probability measures.

If z and d follow discrete distributions with realizations $z_l,\ l = 1,\dots,L$, and $d_k,\ k = 1,\dots,K$, as well as probabilities $\pi_l,\ l = 1,\dots,L$, and $p_k,\ k = 1,\dots,K$, respectively, and $g(x) := g^\top x$ is linear, then, under (A1) and (A2), problem (21) is equivalent to the mixed-integer linear program

$$\left.\begin{array}{lll} \min\Big\{ g^\top x : & c^\top x + q^\top y_{lk} - d_k \ \le\ v_{lk} & \forall l\ \forall k \\ & Tx + W y_{lk} \ =\ z_l & \forall l\ \forall k \\ & \sum_{l=1}^{L} \pi_l v_{lk} \ \le\ \bar{d}_k & \forall k \\ & x \in X,\ y_{lk} \in \mathbb{Z}_+^{\bar{m}} \times \mathbb{R}_+^{m'},\ v_{lk} \ge 0 & \forall l\ \forall k \end{array}\right\}, \tag{23}$$

where $\bar{d}_k := \mathbb{E}_d[d - d_k]_+,\ k = 1,\dots,K$. Clearly, the optimal value to (23) is a lower bound to the optimal value to (15). Compared to (15), problem (23) looks a little simpler since continuous variables v_{lk} instead of Boolean variables θ_{lk} can be employed for integrating the stochastic order. Nevertheless, (23) remains a mixed-integer model due to the integralities in the y_{lk}. Moreover, (23) bears the same internal coupling as (15): Beside nonanticipativity, the counterparts to (18) link subproblems belonging to the individual scenarios $l \in \{1,\dots,L\}$. Computational tests in [15] have confirmed that this linkage is critical when tackling (23) with mixed-integer linear program solvers such as CPLEX. Thus it seems that, although formally simpler than (15), problem (23) does not offer a "cheap" option for its direct solution.

In [15] a decomposition algorithm for (23) was developed that resembles Algorithm 3.3: Within a branch-and-bound framework, lower bounding is achieved by ignoring nonanticipativity and subjecting the counterparts to (18) to Lagrangean relaxation. Upper bounds are obtained by a feasibility heuristic. This algorithm proves superior to the solver CPLEX.

In our numerical experiments, reported in Tables 9 and 10, we have evaluated the potential of (23) to provide lower bounds to (15) that are tighter than the lower bounds generated via Algorithm 3.3. Table 9 shows the development over time of lower bounds for instance 4 from subsection 4.1 with 20, 30, and 50 data scenarios. Here the model (21) provides preferable lower bounds in early stages of the iteration. However, just the opposite occurs in Table 10 for instance 2 where Algorithm 3.3 yields the preferable lower bounds throughout the iteration.

Acknowledgments. We thank Christoph Helmberg (Technical University of Chemnitz) for giving us access to the implementation of his spectral bundle method. Moreover, we are grateful to Uwe Gotzes (University of Duisburg-Essen) for fruitful discussions and his support in running computational tests. Last but not least, we express our thanks to two anonymous referees and the associate editor for their suggestions improving earlier versions of this paper.

TABLE 9
Lower bounds over time for instance 4.

Number of scenarios	Time (sec.)	Algorithm 3.3	Model (23)
20	386	–	9
	425	9	9
	789	9	10
	1195	9	11 optimal
	2211	13	11
30	500	–	9
	618	8	9
	1016	8	10
	1402	8	11 optimal
	3107	10	11
	3566	11	11
	4557	12	11
	5645	13	11
50	975	8	–
	1026	8	9
	2075	8	10
	2333	9	10
	2869	9	11 optimal
	3456	10	11
	4991	11	11
	6475	12	11

TABLE 10
Lower bounds over time for instance 2.

Number of scenarios	Time (sec.)	Algorithm 3.3	Model (23)
20	377	12	–
	464	12	9
	704	18	9
	977	18	17
	1027	25	17
	1435	25	25
	1444	26	25
	1765	27	25
	1915	28 optimal	25
	2500	28	27 optimal
30	529	11	–
	702	11	9
	1002	18	9
	1433	18	17
	1450	25	17
	2100	25	25
	2497	27	25
	2692	28 optimal	25
	3633	28	27 optimal
50	813	11	–
	1125	11	9
	1549	18	9
	2208	18	17
	2235	25	17
	3198	25	25
	3853	27	25
	4165	28 optimal	25
	5854	28	27 optimal

REFERENCES

[1] S. Ahmed, *Convexity and decomposition of mean-risk stochastic programs*, Math. Program., 106 (2006), pp. 433–446.

[2] B. Bank and R. Mandel, *Parametric Integer Optimization*, Akademie-Verlag, Berlin, 1988.

[3] P. Billingsley, *Convergence of Probability Measures*, Wiley, New York, 1968.

[4] P. Billingsley, *Probability and Measure*, Wiley, New York, 1986.

[5] J. R. Birge and F. Louveaux, *Introduction to Stochastic Programming*, Springer, New York, 1997.

[6] J. R. Birge and R. J.-B. Wets, *Designing approximation schemes for stochastic optimization problems, in particular for stochastic programs with recourse*, Math. Program. Stud., 27 (1986), pp. 54–102.

[7] C. E. Blair and R. G. Jeroslow, *The value function of a mixed integer program:* I., Discrete Math., 19 (1977), pp. 121–138.

[8] C. C. Carøe and R. Schultz, *Dual decomposition in stochastic integer programming*, Oper. Res. Lett., 24 (1999), pp. 37–45.

[9] Cplex Callable Library 9.1.3, ILOG (2005).

[10] D. Dentcheva, R. Henrion, and A. Ruszczyński, *Stability and sensitivity of optimization problems with first order stochastic dominance constraints*, SIAM J. Optim., 18 (2007), pp. 322–337.

[11] D. Dentcheva and R. Ruszczyński, *Optimization with stochastic dominance constraints*, SIAM J. Optim., 14 (2003), pp. 548–566.

[12] D. Dentcheva and A. Ruszczyński, *Optimality and duality theory for stochastic optimization with nonlinear dominance constraints*, Math. Program., 99 (2004), pp. 329–350.

[13] A. Eichhorn and W. Römisch, *Polyhedral risk measures in stochastic programming*, SIAM J. Optim., 16 (2005), pp. 69–95.

[14] P. C. Fishburn, *Utility Theory for Decision Making*, Wiley, New York, 1970.

[15] R. Gollmer, U. Gotzes, and R. Schultz, *Second-Order Stochastic Dominance Constraints Induced by Mixed-Integer Linear Recourse*, Preprint 644/2007, Department of Mathematics, University of Duisburg-Essen, 2007; also available online from www.uni-duisburg.de/FB11/disma/preprints.shtml.

[16] E. Handschin, F. Neise, H. Neumann, and R. Schultz, *Optimal operation of dispersed generation under uncertainty using mathematical programming*, Internat. J. Electrical Power & Energy Systems, 28 (2006), pp. 618–626.

[17] C. Helmberg and K.C. Kiwiel, *A spectral bundle method with bounds*, Math. Program., 93 (2002), pp. 173–194.

[18] P. Kall, A. Ruszczyński, and K. Frauendorfer, *Approximation techniques in stochastic programming*, in Numerical Techniques for Stochastic Optimization, Y. Ermoliev and R.J.-B. Wets, eds., Springer, Berlin, 1988, pp. 33–64.

[19] P. Kall and S. W. Wallace, *Stochastic Programming*, Wiley, Chichester, 1994.

[20] T. K. Kristoffersen, *Deviation measures in two-stage stochastic linear programming*, Math. Methods Oper. Res., 62 (2006), pp. 255–274.

[21] F. V. Louveaux and Y. Smeers, *Optimal investments for electricity generation: A stochastic model and a test problem*, in Numerical Techniques for Stochastic Optimization, Y. Ermoliev and R.J.-B. Wets, eds., Springer, Berlin, 1988, pp. 445–453.

[22] H. B. Mann and D. R. Whitney, *On a test of whether one of two random variables is stochastically larger than the other*, Ann. Math. Stat., 18 (1947), pp. 50–60.

[23] A. Märkert and R. Schultz, *On deviation measures in stochastic integer programming*, Oper. Res. Lett., 33 (2005), pp. 441–449.

[24] A. Müller and D. Stoyan, *Comparison Methods for Stochastic Models and Risks*, Wiley, Chichester, 2002.

[25] N. Noyan, G. Rudolf, and A. Ruszczyński, *Relaxations of linear programming problems with first order stochastic dominance constraints*, Oper. Res. Lett., 34 (2006), pp. 653–659.

[26] D. Pollard, *Convergence of Stochastic Processes*, Springer, New York, 1984.

[27] A. Prékopa, *Stochastic Programming*, Kluwer, Dordrecht, 1995.

[28] W. Römisch, *Stability of stochastic programming problems*, in Stochastic Programming, Handbooks Oper. Res. Management Sci. 10, Elsevier, Amsterdam, 2003, pp. 483–554.

[29] A. Ruszczyński and A. Shapiro, eds., *Stochastic Programming*, Handbooks Oper. Res. Management Sci. 10, Elsevier, Amsterdam, 2003.

[30] R. Schultz and F. Neise, *Algorithms for mean-risk stochastic integer programs in energy*, Investigación Oper., 28 (2007), pp. 4–16.

[31] R. Schultz and S. Tiedemann, *Risk aversion via excess probabilities in stochastic programs with mixed-integer recourse*, SIAM J. Optim., 14 (2003), pp. 115–138.
[32] R. Schultz and S. Tiedemann, *Conditional value-at-risk in stochastic programs with mixed-integer recourse*, Math. Program., 105 (2006), pp. 365–386.
[33] G. R. Shorack and J. A. Wellner, *Empirical Processes with Applications to Statistics*, Wiley, New York, 1986.

SIAM J. OPTIM.
Vol. 19, No. 2, pp. 572–591

THE OPERATOR Ψ FOR THE CHROMATIC NUMBER OF A GRAPH*

NEBOJŠA GVOZDENOVIĆ† AND MONIQUE LAURENT†

Abstract. We investigate hierarchies of semidefinite approximations for the chromatic number $\chi(G)$ of a graph G. We introduce an operator Ψ mapping any graph parameter $\beta(G)$, nested between the stability number $\alpha(G)$ and $\chi(\overline{G})$, to a new graph parameter $\Psi_\beta(G)$, nested between $\alpha(\overline{G})$ and $\chi(G)$; $\Psi_\beta(G)$ is polynomial time computable if $\beta(G)$ is. As an application, there is no polynomial time computable graph parameter nested between the fractional chromatic number $\chi^*(\cdot)$ and $\chi(\cdot)$ unless P = NP. Moreover, based on the Motzkin–Straus formulation for $\alpha(G)$, we give (quadratically constrained) quadratic and copositive programming formulations for $\chi(G)$. Under some mild assumptions, $n/\beta(G) \leq \Psi_\beta(G)$, but, while $n/\beta(G)$ remains below $\chi^*(G)$, $\Psi_\beta(G)$ can reach $\chi(G)$ (e.g., for $\beta(\cdot) = \alpha(\cdot)$). We also define new polynomial time computable lower bounds for $\chi(G)$, improving the classic Lovász theta number (and its strengthenings obtained by adding nonnegativity and triangle inequalities); experimental results on Hamming graphs, Kneser graphs, and DIMACS benchmark graphs will be given in the follow-up paper [N. Gvozdenović and M. Laurent, *SIAM J. Optim.*, 19 (2008), pp. 592–615].

Key words. (fractional) chromatic number, stability number, Lovász theta number, semidefinite programming

AMS subject classifications. 05C15, 90C27, 90C22

DOI. 10.1137/050648237

1. Introduction.

The chromatic number $\chi(G)$ of a graph $G = (V, E)$ is the minimum number of colors needed to color the nodes of G in such a way that adjacent nodes receive distinct colors. Computing $\chi(G)$ is an NP-hard problem [11], and it is also hard to approximate $\chi(G)$ within $|V(G)|^{1/14-\epsilon}$ for any $\epsilon > 0$ [1]. An obvious lower bound for $\chi(G)$ is the clique number $\omega(G)$, defined as the maximum size of a clique (i.e., a set of pairwise adjacent nodes) in G; computing $\omega(G)$ is also hard [11] as well as approximating $\omega(G)$ within $|V(G)|^{1/6-\epsilon}$ for any $\epsilon > 0$ [1]. A well-known stronger lower bound for $\chi(G)$ is $\overline{\vartheta}(G) := \vartheta\left(\overline{G}\right)$, the theta number of the complementary graph, introduced by Lovász [23] (see (2.3)). The theta number satisfies the "sandwich inequality":

$$\omega(G) \leq \overline{\vartheta}(G) \leq \chi(G),$$

and it can be computed to any arbitrary precision in polynomial time since it can be formulated via a semidefinite program. It can also be used for approximately coloring the graph (see [5, 8, 17]). Intensive research has been done for strengthening the bound $\overline{\vartheta}(G)$ towards $\omega(G)$ or, equivalently, $\vartheta(G)$ towards the stability number $\alpha(G)$; see, e.g., [6, 19, 20, 21, 24, 26, 30, 32, 34]. Here $\alpha(G) = \omega(\overline{G})$, the maximum size of a stable set (i.e., a set of pairwise nonadjacent nodes) in G. In particular, hierarchies of semidefinite (or linear) bounds were constructed that find $\alpha(G)$ in $\alpha(G)$ steps [19, 20, 24, 34]. As $\chi(G)$ can be formulated via a 0/1 linear program (see, e.g.,

*Received by the editors December 22, 2005; accepted for publication (in revised form) December 18, 2007; published electronically July 2, 2008. Supported by the Netherlands Organization for Scientific Research grant NWO 639.032.203 and by ADONET, Marie Curie Research Training Network MRTN-CT-2003-504438.

http://www.siam.org/journals/siopt/19-2/64823.html

†Centrum voor Wiskunde en Informatica, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands (N.Gvozdenovic@cwi.nl, M.Laurent@cwi.nl).

[7]), the lift-and-project methods of [19, 24, 34] can in principle be applied to derive hierarchies of semidefinite approximations finding $\chi(G)$ in finitely many steps. To the best of our knowledge such hierarchies have not been investigated in detail so far.

In this paper we propose a systematic investigation of semidefinite approximations for $\chi(G)$. One of our main contributions is a simple construction permitting one to derive from any graph parameter $\beta(G)$ nested between $\alpha(G)$ and $\overline{\chi}(G)$ a new graph parameter $\Psi_\beta(G)$ nested between $\omega(G)$ and $\chi(G)$. For this, given an integer $t \geq 0$, let $K_t \Box G$ denote the Cartesian product of the two graphs G and K_t, with node set

$$V(K_t \Box G) := V(K_t) \times V(G) = \bigcup_{p=1}^{t} V_p, \quad \text{where } V_p := \{pi \mid i \in V(G)\}, \tag{1.1}$$

and having an edge (pi, qj) if $(p \neq q$ and $i = j)$ or if $(p = q$ and $ij \in E(G))$. Chvátal [4] observed the following useful reduction of the chromatic number to the stability number:

$$\chi(G) \leq t \Longleftrightarrow \alpha(K_t \Box G) = |V(G)|. \tag{1.2}$$

(Reverse reductions, from the stability number to the chromatic number, can be found in Poljak [31] and in Schrijver [33].) Given a graph parameter $\beta(\cdot)$ nested between $\alpha(\cdot)$ and $\overline{\chi}(\cdot)$, relation (1.2) motivates the introduction of the new graph parameter $\Psi_\beta(\cdot)$, defining $\Psi_\beta(G)$ as the smallest integer $t \geq 0$ for which $\beta(K_t \Box G) = |V(G)|$. Among other properties, $\Psi_\alpha(G) = \chi(G)$, $\Psi_{\overline{\chi}}(G) = \Psi_{\overline{\chi^*}}(G) = \omega(G)$, $\Psi_\vartheta(G) = \lceil \overline{\vartheta}(G) \rceil$, and $\Psi_{\vartheta'}(G) = \lceil \overline{\vartheta^+}(G) \rceil$. Here χ^* is the fractional chromatic number, and ϑ' and ϑ^+ are variations of ϑ obtained by adding certain nonnegativity conditions; see section 2.1. Moreover, the operator Ψ is monotone nonincreasing and, if $\beta(G)$ is polynomial time computable (resp., given by a semidefinite program), then the same holds for $\Psi_\beta(G)$. A somewhat surprising application is that there does *not* exist a polynomial time computable graph parameter nested between the fractional chromatic number and the chromatic number unless P = NP (see Theorem 2.6). As another application we can give (quadratically constrained) quadratic and copositive programming formulations for $\chi(G)$ based on the Motzkin–Straus formulation for $\alpha(G)$ (see section 2.5).

The operator Ψ permits one to transform any hierarchy of upper bounds for $\alpha(G)$ into a hierarchy of lower bounds for $\chi(G)$. In this paper we study in particular hierarchies of lower bounds for $\chi(G)$ related to the Lasserre hierarchy $\mathrm{las}^{(r)}(G)$ ($r \in \mathbb{N}$) for $\alpha(G)$ [19], which finds $\alpha(G)$ at order $r = \alpha(G)$ and refines several other known hierarchies for $\alpha(G)$. More precisely, we consider two hierarchies $\psi^{(r)}(G)$ and $\Psi_{\mathrm{las}^{(r)}}(G)$ of lower bounds for the chromatic number $\chi(G)$, which satisfy $\psi^{(1)}(G) = \overline{\vartheta}(G)$ and $\psi^{(2)}(G) \geq \overline{\vartheta^{+\triangle}}(G)$ (Meurdesoif strengthening—see section 2.1), and

$$\frac{|V(G)|}{\mathrm{las}^{(r)}(G)} \leq \psi^{(r)}(G) \leq \Psi_{\mathrm{las}^{(r)}}(G) \leq \chi(G).$$

The parameter $\psi^{(r)}(G)$ has the same computational cost as $\mathrm{las}^{(r)}(G)$, but it cannot go beyond the fractional chromatic number; in fact, $\psi^{(r)}(G) = \chi^*(G)$ for $r \geq \alpha(G)$. The parameter $\Psi_{\mathrm{las}^{(r)}}(G)$ has a higher computational cost than $\mathrm{las}^{(r)}(G)$ (one has to evaluate $\mathrm{las}^{(r)}(K_t \Box G)$ for $O(\log n)$ queries on $t \leq n$), but it finds $\chi(G)$ at step $r = n$. Dukanovic and Rendl [9] introduced recently another hierarchy for $\chi(G)$, which is related to the hierarchy of de Klerk and Pasechnik [6] for $\alpha(G)$, both being based on

copositive programming. The hierarchy of Dukanovic and Rendl remains, however, bounded by the fractional chromatic number; see section 3.5 for details.

Although polynomial time computable for any fixed r, the parameters $\psi^{(r)}(G)$ and $\Psi_{\mathrm{las}^{(r)}}(G)$ are yet too costly to compute for large values of n already for order $r = 2$. We propose some variations $\psi(G)$ and $\Psi_\ell(G)$ of the order 2 bounds, which are at least as good as $\overline{\vartheta^+}(G)$. As will be shown in the follow-up paper [14], for vertex-transitive graphs, the computation of $\psi(G)$ involves a semidefinite program with two matrices of sizes $n+1$ and n, while the computation of $\Psi_\ell(G)$ can be reduced to $O(\log n)$ semidefinite programs with matrices of sizes $2n+1, 2n, n$, and n; these formulations are obtained by exploiting symmetries in the structure of the semidefinite programs and symmetries arising from the permutation group $\mathrm{Sym}(t)$ acting on the complete graph K_t.

More details about the results of this paper can also be found in [12].

Contents of the paper. In section 2 we present the operator Ψ and its main properties, we discuss various ways for computing $\Psi_\beta(G)$, and we give (quadratically constrained) quadratic and copositive programming formulations for $\chi(G)$. In section 3 we investigate two hierarchies of lower bounds for $\chi(G)$ related to the hierarchy of Lasserre for $\alpha(G)$ and converging, respectively, to $\chi^*(G)$ and $\chi(G)$. This leads to two bounds $\psi(G)$ and $\Psi_\ell(G)$ formulated via semidefinite programs involving matrices of size $O(n)$. Finally we explore the link between our bounds and the copositive programming-based hierarchies of de Klerk and Pasechnik [6] for $\alpha(G)$ and of Dukanovic and Rendl [9] for $\chi(G)$.

Notation. Given a graph $G = (V, E)$, $\overline{G}$ denotes its complementary graph whose edges are the pairs $uv \notin E(G)$ $(u, v \in V(G),\ u \neq v)$. Throughout we set $V := V(G)$, $n = |V|$, and to avoid trivial technicalities we assume that $G \neq K_n$ and $G \neq \overline{K_n}$, where K_n denotes the complete graph on n nodes. For two graphs G and G', their Cartesian product $G \square G'$ has node set $V(G) \times V(G')$, with two nodes $uu', vv' \in V(G) \times V(G')$ being adjacent in $G \square G'$ if and only if ($u = v$ and $u'v' \in E(G')$) or ($uv \in E(G)$ and $u' = v'$). For an integer $t \geq 1$, we sometimes set $G_t = K_t \square G$ as a shorthand notation for the Cartesian product of G and K_t, whose node set is as in (1.1). Given a graph parameter $\beta(\cdot)$, $\overline{\beta}(\cdot)$ is the graph parameter defined by $\overline{\beta}(G) := \beta(\overline{G})$ for any graph G.

Throughout, the letters $\mathbf{I}, \mathbf{J}$, and e denote, respectively, the identity matrix, the all-ones matrix, and the all-ones vector (of the suitable size); $\mathbb{N}$ is the set of nonnegative integers. For $n \times n$ matrices A, B, $\mathrm{Tr}(A) = \sum_{i=1}^n A_{ii}$ and $\langle A, B \rangle = \mathrm{Tr}(A^T B) = \sum_{i,j=1}^n A_{ij} B_{ij}$. Moreover, the notation $A \succeq 0$ means that A is a symmetric positive semidefinite matrix.

Given a finite set V, $\mathcal{P}(V)$ denotes the collection of all subsets of V. Given an integer r, set $\mathcal{P}_r(V) := \{I \in \mathcal{P}(V) \mid |I| \leq r\}$. $\mathcal{P}_r(V)$ contains the empty subset of V which we will denote as $\mathbf{0}$; thus, for instance, $\mathcal{P}_1(V) = \{\mathbf{0}, \{i\}\ (i \in V)\}$. We sometimes identify $\mathcal{P}_1(V) \setminus \{\mathbf{0}\}$ with V; i.e., we write $\{i\}$ as i and $\{i, j\}$ as ij, and, given a vector $x \in \mathbb{R}^{\mathcal{P}(V)}$ we also set $x_i := x_{\{i\}}$, $x_{ij} := x_{\{i,j\}}$, $x_{ijk} := x_{\{i,j,k\}}$, etc.

Let V be a finite set, and let $\mathcal{G}$ be a subgroup of $\mathrm{Sym}(V)$, the group of permutations of V, also denoted as $\mathrm{Sym}(n)$ if $|V| = n$. $\mathcal{G}$ acts on $\mathcal{P}(V)$ by letting $\sigma(I) := \{\sigma(i) \mid i \in I\}$ for $I \subseteq V$, $\sigma \in \mathcal{G}$. Moreover, $\mathcal{G}$ acts on vectors and matrices indexed by V (and thus on vectors and matrices indexed by $\mathcal{P}(V)$). Namely, for $\sigma \in \mathcal{G}$, $x \in \mathbb{R}^V$, and $M \in \mathbb{R}^{V \times V}$, set $\sigma(x) := (x_{\sigma(i)})_{i \in V}$ and $\sigma(M) := (M_{\sigma(i),\sigma(j)})_{i,j \in V}$. One says that M is invariant under the action of $\mathcal{G}$ if $\sigma(M) = M$ for all $\sigma \in \mathcal{G}$: The matrix $\frac{1}{|\mathcal{G}|!} \sum_{\sigma \in \mathcal{G}} \sigma(M)$, the "symmetrization" of M obtained by applying the Reynolds

operator, is invariant under the action of $\mathcal{G}$. The same holds analogously for vectors. A semidefinite program is said to be invariant under action of $\mathcal{G}$ if, for any feasible matrix X and any $\sigma \in \mathcal{G}$, the matrix $\sigma(X)$ is again feasible with the same objective value; then the optimum value of the program remains unchanged if we restrict to invariant feasible solutions and, in particular, there is an invariant optimal solution.

The automorphism group $\mathrm{Aut}(G)$ of a graph $G = (V, E)$ consists of all $\sigma \in \mathrm{Sym}(V)$ preserving the set of edges. G is said to be vertex-transitive when, given any two nodes $i, j \in V$, there exists $\sigma \in \mathrm{Aut}(G)$, with $\sigma(i) = j$.

2. New parameters and formulations.

2.1. Some known graph parameters. We review here some classic bounds for the stability number $\alpha(G)$ and the chromatic number $\chi(G)$ of a graph $G = (V, E)$. We give some equivalent formulations for the bounds. Some work may be required to derive some of them; for details see, e.g., [22, 33].

• The *fractional clique cover number*, also known as the *fractional chromatic number of* $\overline{G}$:

$$\begin{array}{rlll} \overline{\chi^*}(G) := \max & e^T x & = \min & e^T \lambda \\ \text{s.t.} & \sum_{i \in C} x_i \le 1 \ (C \text{ clique}) & \text{s.t.} & \sum_{C \text{ clique}} \lambda_C \chi^C = e, \\ & x \in \mathbb{R}^V_+, & & \lambda \ge 0. \end{array} \tag{2.1}$$

It is well known (and easy to verify) that $\alpha(G) \le \overline{\chi^*}(G) \le \overline{\chi}(G)$, and

$$\omega(G)\overline{\chi^*}(G) \ge |V(G)|, \quad \text{with equality when } G \text{ is vertex-transitive.} \tag{2.2}$$

It is hard to compute the fractional chromatic number, and, for some $\epsilon > 0$, there is no polynomial time algorithm to approximate $\chi^*(G)$ within $|V(G)|^\epsilon$ unless P = NP [25].

• *Lovász's theta number* (introduced in [23]):

$$\begin{array}{rlll} \vartheta(G) := \max & \langle \mathbf{J}, X \rangle & = \min & t \\ \text{s.t.} & \mathrm{Tr}(X) = 1 & \text{s.t.} & U_{ii} = 1 \ (i \in V), \\ & X_{ij} = 0 \ \ (ij \in E(G)), & & U_{ij} = -\dfrac{1}{t-1} \ (ij \in E(\overline{G})), \\ & X \succeq 0, & & U \succeq 0, \ t \ge 2, \end{array} \tag{2.3}$$

where X and U are symmetric matrices indexed by V. The minimization program in the above definition of $\vartheta(G)$ is used, e.g., in [17] for constructing a vector k-coloring. We will also use the following equivalent formulation:

$$\begin{array}{rll} \vartheta(G) = \max & \sum_{i \in V} X_{ii} & \\ \text{s.t.} & X_{\mathbf{00}} = 1, \ X_{ij} = 0 \ (ij \in E), & \\ & X_{ii} = X_{\mathbf{0}i} \ (i \in V), \ X \succeq 0, & \end{array} \tag{2.4}$$

where the matrix variable X is indexed by the set $\mathcal{P}_1(V)$. Lovász [23] proved the following analogue of (2.2) for the pair $(\vartheta, \overline{\vartheta})$:

$$\vartheta(G)\overline{\vartheta}(G) \ge |V(G)|, \quad \text{with equality when } G \text{ is vertex-transitive.} \tag{2.5}$$

• The *strengthening of the theta number* of [26, 32]:

(2.6)
$$\begin{array}{rlll} \vartheta'(G) := \max & \langle \mathbf{J}, X\rangle & = \min & t \\ \text{s.t.} & \mathrm{Tr}(X) = 1 & \text{s.t.} & U_{ii} = 1 \ (i \in V), \\ & X_{ij} = 0 \ \ (ij \in E(G)), & & U_{ij} \le -\dfrac{1}{t-1} \ \ \left(ij \in E\left(\overline{G}\right)\right), \\ & X \succeq 0, \ X \ge 0, & & U \succeq 0, \ t \ge 2. \end{array}$$

• *Szegedy's number* [36]:

(2.7)
$$\begin{array}{rlll} \vartheta^+(G) := \max & \langle \mathbf{J}, X\rangle & = \min & t \\ \text{s.t.} & \mathrm{Tr}(X) = 1 & \text{s.t.} & U_{ii} = 1 \ (i \in V), \\ & X_{ij} \le 0 \ \ (ij \in E(G)), & & U_{ij} = -\dfrac{1}{t-1} \ \ \left(ij \in E\left(\overline{G}\right)\right), \\ & X \succeq 0, & & U_{ij} \ge -\dfrac{1}{t-1} \ \ (ij \in E(G)), \\ & & & U \succeq 0, \ t \ge 2. \end{array}$$

Szegedy [36] showed that the analogue of (2.2) and (2.5) also holds for the pair $(\vartheta', \overline{\vartheta^+})$:

$$\vartheta'(G)\overline{\vartheta^+}(G) \ge |V(G)|, \quad \text{with equality when } G \text{ is vertex-transitive.} \tag{2.8}$$

Thus one may see the pairs (α, χ^*), $(\vartheta, \overline{\vartheta})$, and $(\vartheta', \overline{\vartheta^+})$ as "reciprocal" pairs of graph parameters. We will see later in this paper (see Theorem 3.1(e)) that they are in fact part of a more general hierarchy of reciprocal pairs.

• Meurdesoif [27] defines *the bound* $\vartheta^{+\triangle}(G)$ obtained by adding the "triangle inequalities" $U_{ij} + U_{jk} - U_{ik} \le 1$ (for $ij, jk \in E$) to the minimization program defining $\vartheta^+(G)$ in (2.7).

The above parameters satisfy

$$\alpha(G) \le \vartheta'(G) \le \vartheta(G) \le \vartheta^+(G) \le \vartheta^{+\triangle}(G) \le \overline{\chi^*}(G) \le \overline{\chi}(G).$$

The inequality $\vartheta^{+\triangle}(G) \le \overline{\chi^*}(G)$ will follow from Theorem 3.1(c) and (d), and the other inequalities follow directly by using the definitions.

2.2. The operator Ψ. By using relation (1.2), we see that the chromatic number of a graph G can be defined as the optimum solution of the following program:

$$\chi(G) = \min_{t \in \mathbb{N}} \ t \ \text{ s.t. } \alpha(K_t \square G) = |V(G)|. \tag{2.9}$$

This fact motivates the following definition.

DEFINITION 2.1. *Given a graph parameter $\beta(\cdot)$ satisfying*

$$\min\left(\alpha(\cdot), \frac{|V(\cdot)|}{\omega(\cdot)}\right) \le \beta(\cdot) \le \overline{\chi}(\cdot), \tag{2.10}$$

define the graph parameter $\Psi_\beta(\cdot)$ by

$$\Psi_\beta(G) := \min_{t \in \mathbb{N}} \ t \ \ \textit{s.t. } \beta(K_t \square G) = |V(G)|. \tag{2.11}$$

Note added in proof. The operator Ψ applies in fact to the larger range of graph parameters $\beta(\cdot)$ satisfying $\frac{|V(\cdot)|}{\chi(\cdot)} \le \beta(\cdot) \le \bar{\chi}(\cdot)$, thus including graph parameters satisfying relation (2.10). See [12] for the details.

LEMMA 2.2.

(a) *The graph parameter* $\Psi_\beta(G)$ *is well defined if* $\beta(\cdot)$ *satisfies* (2.10).

(b) *The operator* Ψ *is monotone nonincreasing; that is,* $\Psi_{\beta_2}(\cdot) \le \Psi_{\beta_1}(\cdot)$ *if* $\beta_1(\cdot)$ *and* $\beta_2(.)$ *satisfy* (2.10) *and* $\beta_1(\cdot) \le \beta_2(\cdot)$.

(c) $\Psi_\alpha(G) = \chi(G)$.

(d) $\Psi_\beta(G) = \omega(G)$ *for* $\beta(\cdot) := \frac{|V(\cdot)|}{\omega(\cdot)}$.

(e) $\Psi_{\bar{\chi}}(G) = \omega(G)$.

(f) $\Psi_\beta(G) = \chi(G)$ *for* $\beta(\cdot) := \min(\alpha(\cdot), \frac{|V(\cdot)|}{\omega(\cdot)})$.

(g) *If* $\beta(\cdot)$ *satisfies* (2.10), *then*

$$\omega(\cdot) \le \Psi_\beta(\cdot) \le \chi(\cdot). \tag{2.12}$$

Proof. (a) Assume that $\beta(\cdot)$ satisfies (2.10), and let $1 \le t \le n := |V(G)|$. As $\omega(K_t \Box G) = \max(t, \omega(G))$, we have $\frac{|V(K_t \Box G)|}{\omega(K_t \Box G)} \ge t$; together with $\alpha(K_t \Box G) \ge t$, this implies that $\beta(K_t \Box G) \ge t$. On the other hand, $\beta(K_t \Box G) \le \overline{\chi}(K_t \Box G) \le n$. Therefore, $\beta(K_n \Box G) = n$, thus showing that $\Psi_\beta(G)$ is well-defined.

(b) If $\beta_1(\cdot) \le \beta_2(\cdot)$ satisfies (2.10), then $\beta_1(K_t \Box G) = n$ implies that $\beta_2(K_t \Box G) = n$, which gives $\Psi_{\beta_2}(G) \le \Psi_{\beta_1}(G)$.

(c) The identity $\Psi_\alpha(G) = \chi(G)$ follows directly from (2.9).

(d) For $\beta(\cdot) := \frac{|V(\cdot)|}{\omega(\cdot)}$, the identity $\Psi_\beta(G) = \omega(G)$ follows from the fact that $\omega(K_t \Box G) = \max(t, \omega(G))$.

(e) We verify that $\Psi_{\bar{\chi}}(G) = \omega(G)$. As $\overline{\chi}(\cdot) \ge \frac{|V(\cdot)|}{\omega(\cdot)}$, we deduce by using (b) and (d) that $\Psi_{\overline{\chi}}(G) \le \Psi_{|V|/\omega}(G) = \omega(G)$. To show the reverse inequality, consider a clique C in G of size $\omega(G)$, and let C_t be the subset of $V(K_t \Box G)$ consisting of all of the copies of the nodes in C. Thus C_t is covered by t cliques of $K_t \Box G$. As the remaining nodes of $K_t \Box G$ can be covered by $n - |C|$ cliques, we have $\overline{\chi}(K_t \Box G) \le t + n - |C|$. Therefore $\overline{\chi}(K_t \Box G) = n$ implies that $t \ge |C| = \omega(G)$, which shows that $\Psi_{\bar{\chi}}(G) \ge \omega(G)$.

(f) Consider now the parameter $\beta(\cdot) := \min(\alpha(\cdot), \frac{|V(\cdot)|}{\omega(\cdot)})$. As $\beta(\cdot) \le \alpha(\cdot)$, we deduce by using (b) that $\Psi_\beta(G) \ge \Psi_\alpha(G) = \chi(G)$, and equality holds since one can easily verify that $\beta(K_t \Box G) = n$ for $t := \chi(G)$.

(g) Relation (2.12) now follows directly by using again (b). □

COROLLARY 2.3. *If* $\beta(\cdot)$ *is a graph parameter satisfying* $\frac{|V(\cdot)|}{\omega(\cdot)} \le \beta(\cdot) \le \bar{\chi}(\cdot)$, *then* $\Psi_\beta = \omega$. *In particular,* $\Psi_{\overline{\chi^*}} = \omega$.

Proof. The proof follows directly from Lemma 2.2(b), (d), and (e) and (2.2). □

Therefore, the operator Ψ takes a graph parameter $\beta(G)$ (nested, e.g., between $\alpha(G)$ and $\overline{\chi}(G)$) and produces the integer lower bound $\Psi_\beta(G)$ (nested between $\omega(G)$ and $\chi(G)$) for the chromatic number $\chi(G)$; Figure 2.1 illustrates how the operator Ψ acts on various parameters. As $\alpha(G)\chi^*(G) \ge |V(G)|$,

$$\beta(G) \ge \alpha(G) \Longrightarrow \chi(G) \ge \chi^*(G) \ge \frac{|V(G)|}{\beta(G)}.$$

The next lemma shows that, under the mild assumption (2.13), $\Psi_\beta(G)$ is at least as good as the obvious lower bound $|V(G)|/\beta(G)$ for $\chi(G)$. However, $\Psi_\beta(G)$ may be equal to $\chi(G)$, while $\frac{|V(G)|}{\beta(G)}$ always remains below the fractional chromatic number $\chi^*(G)$. One can easily verify that condition (2.13) holds for the graph parameters

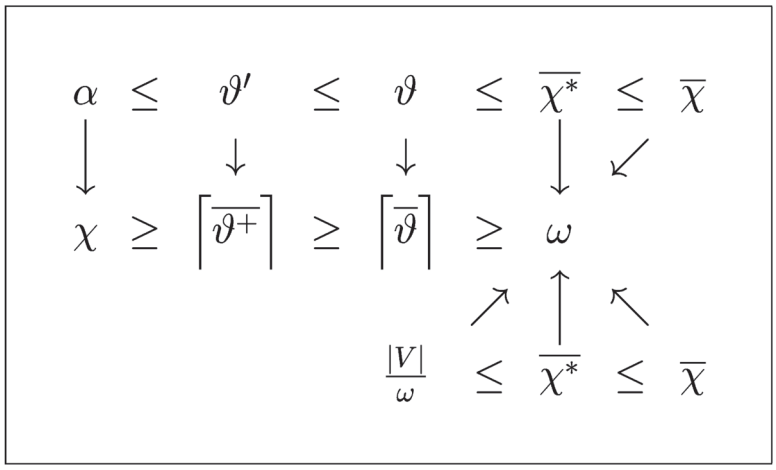


FIG. 2.1. *Converting graph parameters by the operator* Ψ.

considered in this paper, e.g., for $\beta(\cdot) = \alpha(\cdot)$, $\chi(\cdot)$, $\chi^*(\cdot)$, $\vartheta(\cdot)$, $\vartheta'(\cdot)$ and the parameter $\mathrm{las}^{(r)}(\cdot)$ defined later in (3.1) (see [12] for details).

LEMMA 2.4. *Assume that the graph parameter* $\beta(\cdot)$ *satisfies* $\alpha(\cdot) \le \beta(\cdot) \le \overline{\chi}(\cdot)$ *and*

$$\beta(K_t \Box G) \le t\beta(G) \quad \textit{for all } t \in \mathbb{N}. \tag{2.13}$$

Then $\Psi_\beta(G) \ge \frac{|V(G)|}{\beta(G)}$.

Proof. If $\beta(K_t \Box G) = |V(G)|$, then $|V(G)| \le t\beta(G)$, i.e., $t \ge \frac{|V(G)|}{\beta(G)}$. □

Remark 2.5. If $\beta(\cdot) \in [\alpha(\cdot), \overline{\chi}(\cdot)]$, then $\Psi_\beta(G) - \frac{|V(G)|}{\beta(G)} \le \chi(G) - \frac{|V(G)|}{\overline{\chi}(G)}$, with equality, e.g., when G is a perfect graph (since then $\alpha(G) = \overline{\chi}(G) = \beta(G)$ and $\omega(G) = \chi(G) = \Psi_\beta(G)$). Hence the gap $\Psi_\beta(G) - \frac{|V(G)|}{\beta(G)}$ can be made arbitrarily large. For instance, this gap is equal to $n - \frac{2n}{n+1} = n\frac{n-1}{n+1}$ when G is the disjoint union of a clique of size n and n isolated points.

We will investigate in the next section how the operator Ψ applies to the theta number $\vartheta(\cdot)$ and its strengthening $\vartheta'(\cdot)$. We now present an easy but quite surprising consequence of Lemma 2.2 concerning the complexity of graph parameters nested between the fractional chromatic and chromatic numbers or, more generally, in the interval $[|V(\cdot)|/\omega(\cdot), \overline{\chi}(\cdot)]$. The key observation is that the operator Ψ maps the whole interval $[|V(\cdot)|/\omega(\cdot), \overline{\chi}(\cdot)]$ to a single graph parameter (namely, the clique number $\omega(\cdot)$), which is hard to compute.

THEOREM 2.6. *If* $\beta(\cdot)$ *is a graph parameter satisfying* $\frac{|V(\cdot)|}{\omega(\cdot)} \le \beta(\cdot) \le \overline{\chi}(\cdot)$, *then there is no algorithm permitting one to compute* $\beta(G)$ *in time polynomial in* $|V(G)|$ *unless* $P = NP$. *As* $\frac{|V(\cdot)|}{\omega(\cdot)} \le \overline{\chi^*}(\cdot) \le \overline{\chi}(\cdot)$, *the same conclusion holds if* $\overline{\chi^*}(\cdot) \le \beta(\cdot) \le \overline{\chi}(\cdot)$.

Proof. By applying Lemma 2.2, we find that $\Psi_\beta(\cdot) = \omega(\cdot)$. Suppose that one can compute $\beta(G)$ in time $f(n)$, where f is a polynomial in $n = |V(G)|$. Then one can compute $\Psi_\beta(G) = \omega(G)$ in time $\sum_{l=1}^{n} f(ln)$, thus polynomial in n. As computing the clique number is an NP-hard problem [11], this implies that P = NP. □

Let us mention a few graph parameters that are known to lie within the "hard" interval $[\chi^*, \chi]$. Hence none of them can be computed in polynomial time unless

P = NP; such a result was known already, e.g., for the circular chromatic number $\chi_c(G)$ [3].

The *circular graph chromatic number* (or *star chromatic number*) $\chi_c(G)$, introduced by Vince [37] and further studied, e.g., in [3, 39], is defined as follows. Given $r \in \mathbb{R}$, $r \geq 2$, a function $f : V(G) \to [0, r)$ is said to be a r-coloring if $1 \leq |f(u)-f(v)| \leq r-1$ for all edges $uv \in E(G)$. Then $\chi_c(G)$ is defined as the infimum of all r for which G has a r-coloring. The following hold: $\chi(G) - 1 < \chi_c(G) \leq \chi(G)$ and $\chi^*(G) \leq \chi_c(G) \leq \chi(G)$ (see, e.g., [39]).

Another graph parameter lying in the hard interval $[\chi^*, \chi]$ is the *local chromatic number* $\chi_{\mathrm{loc}}(G)$, introduced in [10] as the minimum over all proper colorings of G of the largest number of colors used to color the neighborhood $N_G(v) = \{w \in V(G) \mid vw \in E(G)\}$ of any vertex $v \in V(G)$. Obviously, $\chi_{\mathrm{loc}}(G) \leq \chi(G)$ (the gap between the two parameters can in fact be arbitrarily large [10]), and Körner, Pilotto, and Simonyi [18] show that $\chi^*(G) \leq \chi_{\mathrm{loc}}(G)$.

The *independence ratio* of a graph G is $i(G) := \frac{\alpha(G)}{|V(G)|}$, and its *Hall ratio* is $\rho(G) := \max_{H \subseteq G} \frac{|V(H)|}{\alpha(H)}$, where the maximum is taken over all subgraphs of G. For an integer $k \geq 1$, let $G^{\Box k}$ denote the graph obtained by taking the Cartesian product of k copies of G. Then the *ultimate independence ratio* $I(G)$ and the *ultimate Hall ratio* $h_\Box(G)$ are defined, respectively, as $I(G) := \lim_{k\to\infty} i(G^{\Box k})$ and $h_\Box(G) := \lim_{k\to\infty} \rho(G^{\Box k})$. These graph parameters are studied, e.g., in [15, 16, 35]. In particular, the following relations with fractional and circular chromatic numbers are shown there:

$$\chi^*(G) \leq \frac{1}{I(G)} = h_\Box(G) \leq \chi_c(G) \leq \chi(G)$$

(see [39] for the inequality $1 \leq I(G)\chi_c(G)$).

2.3. Action of the operator Ψ on the theta number. The next theorem shows that the operator Ψ maps the theta number $\vartheta(\cdot)$ to $\lceil \overline{\vartheta}(\cdot) \rceil$ and its strengthening $\vartheta'(\cdot)$ to $\lceil \overline{\vartheta^+}(\cdot) \rceil$. De Klerk, Pasechnik, and Warners [5] consider a graph parameter closely related to Ψ_ϑ for which they can also show that it coincides with $\lceil \overline{\vartheta}(\cdot) \rceil$.

THEOREM 2.7. *For any graph G the following hold:*

(i) $\Psi_\vartheta(G) = \lceil \overline{\vartheta}(G) \rceil$,

(ii) $\Psi_{\vartheta'}(G) = \lceil \overline{\vartheta^+}(G) \rceil$.

We first state two lemmas that we need for the proof of Theorem 2.7.

LEMMA 2.8. *Let X be a $t \times t$ block matrix, having an $n \times n$ matrix A as its diagonal blocks and an $n \times n$ matrix B as nondiagonal blocks, i.e.,*

$$X = \underbrace{\begin{pmatrix} A & B & \dots & B \\ B & A & \dots & B \\ \vdots & \vdots & \ddots & \vdots \\ B & B & \dots & A \end{pmatrix}}_{t \text{ blocks}}. \tag{2.14}$$

Then $X \succeq 0 \iff A - B \succeq 0$ and $A + (t-1)B \succeq 0$.

Proof. We define a $t \times t$ block matrix U_t having the same block structure as the matrix X. For $p, q = 1, \dots, t$, let U_t^{pq} denote the (p, q)th block of U_t, defined by

$$U_t^{pq} := \begin{cases} \frac{1}{\sqrt{t}}\mathbf{I} & \text{if } p = 1 \text{ or } q = 1, \\ \left(\frac{1}{\sqrt{t}+t} - 1\right)\mathbf{I} & \text{if } p = q \geq 2, \\ \frac{1}{\sqrt{t}+t}\mathbf{I} & \text{otherwise.} \end{cases} \tag{2.15}$$

Here $\mathbf{I}$ stands for the identity matrix of order n. Notice that U_t is symmetric and orthogonal, i.e., $U_t(U_t)^T = \mathbf{I}$. Let $Y := (U_t)^T X U_t$. Then $Y \succeq 0$ if and only if $X \succeq 0$, and a simple calculation gives

$$Y = \begin{bmatrix} A+(t-1)B & 0 & \dots & 0 \\ 0 & A-B & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A-B \end{bmatrix}, \tag{2.16}$$

which shows the lemma. □

LEMMA 2.9. *For a positive semidefinite $n \times n$ matrix X, $n\mathrm{Tr}(X) \geq \langle \mathbf{J}, X \rangle$, with equality if and only if $X = c\mathbf{J}$ for some nonnegative scalar c.*

Proof. As $X \succeq 0$, its entries satisfy $X_{ii} + X_{jj} \geq 2X_{ij}$ for all $i, j \in \{1, \dots, n\}$. Thus, $n \sum_{i=1}^n X_{ii} \geq \sum_{i,j=1}^n X_{ij}$. Equality holds if and only if $X_{ii} + X_{jj} = 2X_{ij}$ for all i, j, which gives $X_{ii} = X_{jj} = X_{ij}$ for all i, j. □

Proof of Theorem 2.7. (i) As G has at least one edge, $\vartheta(G) < n$ and thus $\Psi_\vartheta(G) \geq 2$. Let (t, X) be a feasible solution for the program defining $\Psi_\vartheta(G)$; that is,

$$X \succeq 0,\ X_{uv} = 0\ (uv \in E(K_t \square G)),\ \mathrm{Tr}(X) = 1,\ \langle \mathbf{J}, X \rangle = n. \tag{2.17}$$

Here the matrix X is indexed by $V(K_t \square G) = \cup_{p=1}^t V_p$ (recall (1.1)) and $t \in \mathbb{N}$, $t \geq 2$. As the program (2.17) is invariant under action of the group $\mathrm{Sym}(t)$, one may assume that X is invariant under action of $\mathrm{Sym}(t)$. Then X has the block form (2.14). By using Lemma 2.8, (2.17) can be rewritten as

$$\begin{gathered} A - B \succeq 0,\ A + (t-1)B \succeq 0,\ A_{ij} = 0\ (ij \in E(G)),\ \mathrm{diag}(B) = 0, \\ \mathrm{Tr}(A) = \frac{1}{t},\ \langle \mathbf{J}, A + (t-1)B \rangle = \frac{n}{t}. \end{gathered} \tag{2.18}$$

Lemma 2.9 implies that $A + (t-1)B = \frac{1}{nt}\mathbf{J}$. By setting $U := nt(A - B)$, we find that

$$U = \frac{1}{t-1}(nt^2 A - \mathbf{J}). \tag{2.19}$$

One can verify that (t, U) is feasible for the program

$$\min\ t\ \text{ s.t. } \mathrm{diag}(U) = e,\ U_{ij} = -\frac{1}{t-1}\ (ij \in E(G)),\ U \succeq 0,\ t \geq 2 \tag{2.20}$$

defining the parameter $\overline{\vartheta}(G)$ (see (2.3)). As $t \in \mathbb{N}$, this implies that $\Psi_\vartheta(G) \geq \lceil \overline{\vartheta}(G) \rceil$. Conversely, let (t, U) be feasible for (2.20), with t an integer. Define the matrices A and B via the equations

$$A - B = \frac{1}{nt}U \ \text{ and } A + (t-1)B = \frac{1}{nt}\mathbf{J}, \tag{2.21}$$

and let X be the corresponding block matrix as in (2.14). One can verify that (2.18) holds and thus (2.17) holds, too. That is, (t, X) is feasible for (2.17). Thus we have shown that

$$\Psi_\vartheta(G) = \min_{t \in \mathbb{N}}\ t\ \text{ s.t. } \mathrm{diag}(U) = e,\ U_{ij} = -\frac{1}{t-1}\ (ij \in E(G)),\ U \succeq 0,\ t \geq 2. \tag{2.22}$$

We now show that $\Psi_\vartheta(G) \le \lceil \overline{\vartheta}(G) \rceil$. For this, set $t := \overline{\vartheta}(G)$, and take an optimal solution U to the program (2.20). Then, by setting $Y := \frac{t-1}{\lceil t \rceil - 1} U + \frac{\lceil t \rceil - t}{\lceil t \rceil - 1} \mathbf{I}$, the pair $(\lceil t \rceil, Y)$ is feasible for (2.22) with objective value $\lceil t \rceil$, which implies that $\lceil t \rceil \ge \Psi_\vartheta(G)$. Thus equality $\lceil \overline{\vartheta}(G) \rceil = \Psi_\vartheta(G)$ holds.

The proof of (ii) is analogous to that of (i). Simply note that adding the condition $X \ge 0$ to (2.17) amounts to adding the condition $A, B \ge 0$ to (2.18) and thus, in view of (2.19), to adding the condition $U_{ij} \ge -\frac{1}{t-1}$ $(i, j \in V)$ to (2.22). □

2.4. Semidefinite programming formulation for the new bounds. We consider here issues related to the computation of $\Psi_\beta(G)$. We assume throughout that $\beta(\cdot)$ satisfies (2.10). There is an obvious way to find $\Psi_\beta(G)$, namely, by computing $\beta(K_t \Box G)$ for each $t = 1, \dots, n$. We now observe that, when $\beta(\cdot)$ is monotone nondecreasing (with respect to taking induced subgraphs), one can use binary search, and it suffices to compute $\beta(K_t \Box G)$ for $O(\log n)$ instances of t.

LEMMA 2.10. *Assume that*

$$\beta(K_t \Box G) \le \beta(K_{t+1} \Box G) \text{ for all } t \in \mathbb{N}. \tag{2.23}$$

Then $\beta(K_t \Box G) = n \Longleftrightarrow \Psi_\beta(G) \le t$.

Proof. The "only if" part follows from the definition of $\Psi_\beta(G)$. For the "if" part assume that $t_0 := \Psi_\beta(G) \le t$. Then $\beta(K_{t_0} \Box G) = n \le \beta(K_t \Box G)$ implies that $\beta(K_t \Box G) = n$, since $\beta(K_t \Box G) \le \overline{\chi}(G) \le n$. □

Under assumption (2.23) one can use binary search for computing $\Psi_\beta(G)$. Namely, given $t_0 \in [1, n]$, compute $\beta(K_{t_0} \Box G)$. There are two cases:

- $\beta(K_{t_0} \Box G) < n$. Then $\Psi_\beta(G) \ge t_0 + 1$ (by the above lemma), and we can now restrict the search to $t \in [t_0 + 1, n]$.
- Or $\beta(K_{t_0} \Box G) = n$. Then $\Psi_\beta(G) \le t_0$, and we can restrict the search to $t \in [1, t_0]$.

Therefore, one can find $\Psi_\beta(G)$ by computing $\beta(K_t \Box G)$ for $O(\log n)$ queries of t.

Observe that one may restrict the range of search for t. Suppose that we know a lower bound t_1 and an upper bound t_2 on $\chi(G)$; that is, $t_1 \le \chi(G) \le t_2$. Then we may assume that $t \le t_2$ in the definition of $\Psi_\beta(G)$, and if we add the condition $t \ge t_1$, then one still obtains a lower bound for $\chi(G)$. Therefore, we may restrict the binary search to $t \in [t_1, t_2]$. For instance, one can choose $t_1 = 3$ if G is not bipartite, or $t_1 = \omega(G)$, and $t_2 = \Delta(G) + 1$ (or even $\Delta(G)$ by Brook's theorem (see [33]) if G is not a clique or an odd circuit), $\Delta(G)$ being the maximum degree of G.

Next we show that $\Psi_\beta(G)$ can be formulated via a single semidefinite program when $\beta(\cdot)$ is given by a semidefinite program satisfying certain assumptions. Namely, our construction applies to the case when the semidefinite program defining $\beta(\cdot)$ involves at least one equality constraint of the form $\langle A, X \rangle = 1$, with $A \succeq 0$. Then one may assume without loss of generality (w.l.o.g.) that all other (in)equality constraints in the program are homogeneous, i.e., of the form $\langle B, X \rangle \ge 0$. (Write any equation $\langle B, X \rangle = 0$ as two opposite inequalities $\langle -B, X \rangle \ge 0$ and $\langle B, X \rangle \ge 0$.) So let us assume that, for an arbitrary graph H, we can express $\beta(H)$ as

$$\begin{aligned} \beta(H) = \max \langle C(H), X(H) \rangle \quad \text{s.t.} \quad & \langle A(H), X(H) \rangle = 1, \\ & \mathcal{B}(H)(X(H)) \ge 0, \\ & X(H) \succeq 0, \end{aligned} \tag{2.24}$$

where $C(H)$ and $A(H)$ are constant symmetric $n \times n$ matrices, $\mathcal{B}(H) : S_n \to \mathbb{R}^{d(H)}$ is a linear operator, and $X(H)$ is the matrix variable. Note that $d(\cdot)$ depends on H,

e.g., $d(H) = 2|E(H)|$ in the formulation of $\vartheta(H)$. Moreover we assume that

$$A(H) \succeq 0, \tag{2.25}$$

$$\langle A(H), X(H)\rangle = 0 \Longrightarrow \langle C(H), X(H)\rangle = 0. \tag{2.26}$$

Note that assumptions (2.23), (2.24), (2.25), and (2.26) hold, e.g., for $\vartheta(\cdot)$ or for the Lasserre hierarchy considered in section 3.1. Recall that our operator Ψ maps $\beta(\cdot)$ in the following way:

$$\begin{array}{rlll} \Psi_\beta(G) := \min t & = \min & t & \\ \text{s.t. } \beta(G_t) = n & \text{s.t.} & \langle C(G_t), X(G_t)\rangle = n, & \\ & & \langle A(G_t), X(G_t)\rangle = 1, & \\ & & \mathcal{B}(G_t)(X(G_t)) \geq 0, & \\ & & X(G_t) \succeq 0. & \end{array} \tag{2.27}$$

Here we use the more concise notation $G_t := K_t \square G$. Let us define

$$\begin{array}{rlll} \Phi_\beta(G) := \min \displaystyle\sum_{t=1}^n t\langle A(G_t), X(G_t)\rangle & \text{s.t.} & \displaystyle\sum_{t=1}^n \langle C(G_t), X(G_t)\rangle = n, & \\ & & \displaystyle\sum_{t=1}^n \langle A(G_t), X(G_t)\rangle = 1, & \\ & & \mathcal{B}(G_t)(X(G_t)) \geq 0 \ (t = 1, \ldots, n), & \\ & & X(G_t) \succeq 0 \ (t = 1, \ldots, n). & \end{array} \tag{2.28}$$

THEOREM 2.11. *Under assumptions* (2.24), (2.25), *and* (2.26), $\Phi_\beta(G) = \Psi_\beta(G)$.

Proof. Take a feasible solution $(t, X(G_t))$ for the program (2.27), and for $k \neq t$ set $X(G_k) := 0$. In this way one obtains a feasible solution for (2.28) with the same objective value as (2.27), which shows that $\Phi_\beta(G) \leq \Psi_\beta(G)$. Conversely, let $X(G_t)$ $(t = 1, \ldots, n)$ be a feasible solution for (2.28), and set $a_t := \langle A(G_t), X(G_t)\rangle$. Thus $a_t \geq 0$ since $A(G_t) \succeq 0$ (by assumption (2.25)) and $\sum_t a_t = 1$. Consider t for which $a_t > 0$. As $\langle A(G_t), \frac{X(G_t)}{a_t}\rangle = 1$, $\frac{X(G_t)}{a_t}$ is feasible for (2.24) (with $H = G_t$), which implies that $\langle C(G_t), \frac{X(G_t)}{a_t}\rangle \leq \beta(G_t) \leq n$; moreover, equality $\langle C(G_t), \frac{X(G_t)}{a_t}\rangle = n$ implies that $\beta(G_t) = n$ and thus $\Psi_\beta(G) \leq t$. Now we have

$$n = \sum_t \langle C(G_t), X(G_t)\rangle = \sum_{t|a_t>0} a_t \left\langle C(G_t), \frac{X(G_t)}{a_t}\right\rangle \leq \left(\sum_{t|a_t>0} a_t\right) n = n.$$

(Here we used assumption (2.26) for the second equality.) Therefore, equality holds throughout, which implies that $\Psi_\beta(G) \leq t$ whenever $a_t > 0$. Hence, $\sum_t t a_t = \sum_{t|a_t>0} t a_t \geq \Psi_\beta(G)(\sum_{t|a_t>0} a_t) = \Psi_\beta(G)$, which gives $\Phi_\beta(G) \geq \Psi_\beta(G)$. □

Hence, under the assumptions (2.24), (2.25), and (2.26), the parameter $\Psi_\beta(G)$ can be formulated via the semidefinite program (2.28), which involves a block-diagonal matrix with diagonal blocks $X(G_1), \ldots, X(G_n)$, each $X(G_t)$ being the matrix variable involved in the program (2.24) for the graph $H = G_t$. For instance, if (2.24) involves a matrix variable of order $f(V(H))$, then (2.28) involves a block-diagonal matrix with block sizes $f(n), f(2n), \ldots, f(n^2)$. As explained above one can reduce the size of the program (2.28) by restricting the range of t in program (2.28) to $t \in [t_1, t_2]$, where $t_1 \leq \chi(G) \leq t_2$.

2.5. Copositive programming formulation for the chromatic number. The technique used in section 2.4 can also be applied to derive (quadratically constrained) quadratic and copositive programming formulations for the chromatic number. Recall that a matrix X is *copositive* if $x^T X x \geq 0$ for all $x \geq 0$. A matrix X is *completely positive* if it belongs to the dual of the cone of copositive matrices, i.e., if it can be written as $X = \sum_i x_i x_i^T$ for some $x_i \geq 0$.

Our starting point is the theorem of Motzkin and Straus [28], which, for a graph G with adjacency matrix A_G, gives the following formulation for its stability number:

$$\frac{1}{\alpha(G)} = \min\ x^T(\mathbf{I} + A_G)x \ \text{ s.t. } x \in \mathbb{R}_+^{V(G)},\ e^T x = 1, \tag{2.29}$$

or, equivalently (see [6]),

$$\alpha(G) = \min\ t \ \text{ s.t. } t(\mathbf{I} + A_G) - \mathbf{J} \text{ is copositive.} \tag{2.30}$$

By using (2.29), we can rewrite the program (2.9) as

$$\chi(G) = \min\ t \text{ s.t. } x_t^T(\mathbf{I} + A_{G_t})x_t = \frac{1}{n},\ e_t^T x_t = 1,\ x_t \in \mathbb{R}_+^{V(G_t)}. \tag{2.31}$$

Here and below e_t denotes the all-ones vector in $\mathbb{R}^{V(G_t)}$. By using the idea from section 2.4 let us define

$$\begin{aligned} \Phi_1(G) := \min & \sum_{t=1}^n t(e_t^T x_t)^2 \\ \text{s.t. } & \sum_{t=1}^n (e_t^T x_t)^2 = 1, \\ & \sum_{t=1}^n x_t^T(\mathbf{I} + A_{G_t})x_t = \frac{1}{n}, \\ & x_t \in \mathbb{R}_+^{V(G_t)}\ (t = 1, \ldots, n). \end{aligned} \tag{2.32}$$

PROPOSITION 2.12. $\Phi_1(G) = \chi(G)$.

Proof. By taking a feasible solution (t, x_t) for the program (2.31) and setting $x_k = 0$ for $k \neq t$, we obtain a feasible solution for (2.32) with objective value t. Thus, $\Phi_1(G) \leq \chi(G)$. Conversely, let x_t $(t = 1, \ldots, n)$ be feasible for (2.32). Then

$$\frac{1}{n} = \sum_t x_t^T(\mathbf{I} + A_{G_t})x_t = \sum_{t|x_t \neq 0} \frac{x_t^T}{e_t^T x_t}(\mathbf{I} + A_{G_t})\frac{x_t}{e_t^T x_t}(e_t^T x_t)^2 \geq \frac{1}{n}\sum_{t|x_t \neq 0}(e_t^T x_t)^2 = \frac{1}{n}.$$

We have used $\frac{x_t^T}{e_t^T x_t}(\mathbf{I} + A_{G_t})\frac{x_t}{e_t^T x_t} \geq \frac{1}{\alpha(G_t)} \geq \frac{1}{n}$. Hence equality holds throughout, which implies that $\alpha(G_t) = n$ if $x_t \neq 0$ and thus $\chi(G) \leq t$ if $x_t \neq 0$. Therefore,

$$\sum_t t(e_t^T x_t)^2 = \sum_{t|x_t \neq 0} t(e_t^T x_t)^2 \geq \chi(G) \sum_{t|x_t \neq 0}(e_t^T x_t)^2 = \chi(G).$$

This shows that $\Phi_1(G) \geq \chi(G)$. ∎

Up to rescaling, we obtain the following formulation for $\chi(G)$ involving only quadratic constraints:

$$\begin{aligned} \chi(G) = \min \quad & \frac{1}{n^2}\sum_{t=1}^{n} t(e_t^T x_t)^2 \\ \text{s.t.} \quad & \sum_{t=1}^{n}(e_t^T x_t)^2 = n^2, \\ & \sum_{t=1}^{n} x_t^T(\mathbf{I}+A_{G_t})x_t = n, \\ & x_t \in \mathbb{R}_+^{V(G_t)} \ (t=1,\dots,n). \end{aligned} \tag{2.33}$$

It is not difficult to verify that the above program remains a formulation of $\chi(G)$ if we replace the condition $x_t \geq 0$ (for all t) by the condition that x_t is 0/1 valued (for all t). Therefore this gives a 0/1 (quadratically constrained) quadratic programming formulation for the chromatic number involving $O(n^3)$ variables.

By starting from (2.33), we can now derive a copositive programming formulation for $\chi(G)$. Namely, consider the program

$$\begin{aligned} \Phi_2(G) := \min \quad & \frac{1}{n^2}\sum_{t=1}^{n} t\langle \mathbf{J}, X_t\rangle \\ \text{s.t.} \quad & \sum_{t=1}^{n}\langle \mathbf{J}, X_t\rangle = n^2, \\ & \sum_{t=1}^{n}\langle \mathbf{I}+A_{G_t}, X_t\rangle = n, \\ & X_t \text{ completely positive } \ (t=1,\dots,n). \end{aligned} \tag{2.34}$$

PROPOSITION 2.13. $\Phi_2(G) = \chi(G)$.

Proof. The formulation (2.33) for $\chi(G)$ implies directly that $\Phi_2(G) \leq \chi(G)$. Conversely, let X_t $(1 \leq t \leq n)$ be a feasible solution for (2.34). Consider t for which $X_t \neq 0$. Say, $X_t = \sum_{i_t} x_{i_t} x_{i_t}^T$ where $x_{i_t} \geq 0$, $x_{i_t} \neq 0$ for all i_t. Thus $\lambda_{i_t} := \sqrt{\langle \mathbf{J}, x_{i_t} x_{i_t}^T\rangle} = e_t^T x_{i_t} > 0$. Set $y_{i_t} := \frac{x_{i_t}}{\lambda_{i_t}}$. By assumption, we have $\sum_t \langle n(\mathbf{I}+A_{G_t})-\mathbf{J}, X_t\rangle = 0$. By (2.30), each matrix $n(\mathbf{I}+A_{G_t})-\mathbf{J}$ is copositive, since $n \geq \alpha(G_t)$. This implies that $\langle n(\mathbf{I}+A_{G_t})-\mathbf{J}, X_t\rangle = 0$ and thus $\langle n(\mathbf{I}+A_{G_t})-\mathbf{J}, x_{i_t}x_{i_t}^T\rangle = 0$ for all i_t. From this follows that $\langle \mathbf{I}+A_{G_t}, y_{i_t}y_{i_t}^T\rangle = \frac{1}{n}$ for all i_t. As $e_t^T y_{i_t} = 1$, y_{i_t} is feasible for the program (2.31), implying that $\chi(G) \leq t$ whenever $X_t \neq 0$. Now $(1/n^2)\sum_t t\langle \mathbf{J}, X_t\rangle \geq (1/n^2)\chi(G)\sum_t \langle \mathbf{J}, X_t\rangle = \chi(G)$, giving $\Phi_2(G) \geq \chi(G)$. □

By rewriting the condition $\sum_t \langle \mathbf{I}+A_{G_t}, X_t\rangle = n$ as $\sum_t \langle n(\mathbf{I}+A_{G_t})-\mathbf{J}, X_t\rangle = 0$, the dual conic program of (2.34) reads:

$$\max_{y,z} \ y \text{ s.t. } \frac{1}{n^2}(t-y)\mathbf{J} + z(n(\mathbf{I}+A_{G_t})-\mathbf{J}) \text{ copositive for } 1 \leq t \leq n. \tag{2.35}$$

There is no duality gap since the program (2.35) is strictly feasible. Thus (2.35) is yet another formulation of $\chi(G)$. This opens the road to another type of hierarchy of relaxations for $\chi(G)$, obtained by approximating the copositive cone by tractable subcones as suggested by Parrilo [29]. This type of approach based on copositive programming has been studied, e.g., in [2] for standard quadratic optimization problems, in [6, 13, 30] for the stable set problem, and recently in [9] for the coloring problem. We will come back to it in section 3.5.

3. Semidefinite hierarchies for (fractional) chromatic numbers. We have seen in the previous section how to construct semidefinite programming lower bounds for the chromatic number of a graph from semidefinite programming upper bounds on the stability number. Several hierarchies of such upper bounds for the stability number have been proposed in the literature, in particular, in [6, 19, 24, 30, 34]. These hierarchies were further studied and compared, e.g., in [13, 20]. It turns out that Lasserre's hierarchy, proposed in [19], gives the tightest bounds. For this reason we focus in this section on this hierarchy, and we show how it can be used and transformed to produce hierarchies of lower bounds for the (fractional) chromatic number. We will also discuss the link with another hierarchy recently proposed by Dukanovic and Rendl [9] based on copositive programming.

3.1. Lasserre's hierarchy towards the stability number. For a subset $S \subseteq V$ and an integer $r \geq 1$, define the vectors $\chi^S \in \{0,1\}^V$ with ith entry 1 if and only if $i \in S$ (for $i \in V$) and $\chi^{S,r} \in \{0,1\}^{\mathcal{P}_r(V)}$ with Ith entry 1 if and only if $I \subseteq S$ (for $I \in \mathcal{P}_r(V)$). Given a vector $x = (x_I)_{I \in \mathcal{P}_{2r}(V)}$, consider the matrix:

$$M_r(x) := (x_{I \cup J})_{I,J \in \mathcal{P}_r(V)}$$

indexed by $\mathcal{P}_r(V)$, known as the *(combinatorial) moment matrix* of x of order r. Consider the program:[1]

$$\text{las}^{(r)}(G) := \max \sum_{i \in V} x_i \ \text{ s.t. } M_r(x) \succeq 0,\ x_{\mathbf{0}} = 1,\ x_{ij} = 0\ (ij \in E), \tag{3.1}$$

with variable $x \in \mathbb{R}^{\mathcal{P}_{2r}(V)}$. As the feasible region is bounded, the maximum is indeed attained in program (3.1). Obviously, $\text{las}^{(r+1)}(G) \leq \text{las}^{(r)}(G)$ (since $M_r(x)$ is a principal submatrix of $M_{r+1}(x)$) and, in view of (2.4), $\text{las}^{(1)}(G) = \vartheta(G)$. In this way one obtains a hierarchy of semidefinite programming bounds for the stability number, known as Lasserre's hierarchy [19, 20]. Indeed, if S is a stable set, the vector $x := \chi^{S,2r}$ is feasible for (3.1) with objective value $|S|$, showing that $\alpha(G) \leq \text{las}^{(r)}(G)$. For fixed r, the parameter $\text{las}^{(r)}(G)$ can be computed in polynomial time (to an arbitrary precision) since the semidefinite program (3.1) involves matrices of size $O(n^r)$ with $O(n^{2r})$ variables (see, e.g., [38] for details on semidefinite programming). It is shown in [20] that, for $r \geq \alpha(G)$,

$$x \text{ is feasible for (3.1)} \iff x = \sum_{S \text{ stable}} \lambda_S \chi^{S,2r},\ \text{for some } \lambda \geq 0,\ \sum_{S \text{ stable}} \lambda_S = 1. \tag{3.2}$$

This implies that

$$\alpha(G) = \text{las}^{(r)}(G) \text{ for } r \geq \alpha(G). \tag{3.3}$$

3.2. An analogous semidefinite programming hierarchy towards the fractional chromatic number. For an integer $r \geq 1$, define the parameter

$$\psi^{(r)}(G) := \min t \ \text{ s.t. } M_r(x) \succeq 0,\ x_{\mathbf{0}} = t,\ x_i = 1\ (i \in V),\ x_{ij} = 0\ (ij \in E), \tag{3.4}$$

where the variable x is indexed by $\mathcal{P}_{2r}(V)$. Note that one can avoid the variable t simply by replacing t by $x_{\mathbf{0}}$ in the objective function. We choose this formulation in

[1] One can easily verify that, under the condition $M_r(x) \succeq 0$, the edge condition $x_{ij} = 0$ for $ij \in E$ implies that $x_I = 0$ for any $I \in \mathcal{P}_{2r}(V)$ containing an edge.

order to have a unified presentation of the various bounds; compare, e.g., with (2.9), (2.11), (3.9), (3.12), and (3.14). Again the minimum is attained in program (3.4), and, for fixed r, one can compute $\psi^{(r)}(G)$ to any arbitrary precision in polynomial time.

THEOREM 3.1. *The parameters $\psi^{(r)}(G)$ satisfy:*

(a) $\psi^{(r)}(G) \le \psi^{(r+1)}(G)$,

(b) $\psi^{(1)}(G) = \overline{\vartheta}(G)$,

(c) $\overline{\vartheta^{+\triangle}}(G) \le \psi^{(2)}(G)$,

(d) $\psi^{(r)}(G) \le \chi^*(G)$, *with equality if $r \ge \alpha(G)$,*

(e) $\psi^{(r)}(G)\mathrm{las}^{(r)}(G) \ge |V(G)|$, *with equality if G is vertex-transitive.*

Proof. (a) is obvious. For (b), let $M_1(x) = \begin{pmatrix} t & e^T \\ e & M \end{pmatrix}$ be a matrix optimal for (3.4) with $r = 1$. Then $\psi^{(1)}(G) = t \ge 2$ (as G has an edge) and $M_1(x) \succeq 0$ or, equivalently, $M - \frac{1}{t}ee^T \succeq 0$. After setting $U := \frac{t}{t-1}(M - \frac{1}{t}ee^T) = \frac{t}{t-1}M - \frac{1}{t-1}ee^T$, we can rewrite the program for $\psi^{(1)}(G)$ in the following way:

$$\begin{aligned} \psi^{(1)}(G) = \min\ t\ \text{ s.t. }\quad & U_{ii} = 1, \\ & U_{ij} = -\frac{1}{t-1}\ (ij \in E), \\ & U \succeq 0,\ t \ge 2. \end{aligned}$$

Thus, in view of (2.3), $\psi^{(1)}(G) = \overline{\vartheta}(G)$.

(c) Assume that (t, x) is feasible for the program defining $\psi^{(2)}(G)$. Consider the principal submatrix X of $M_2(x)$ indexed by $\{k, ij, ik, jk\}$, where i, j, k are distinct elements of V and the vector $w := (1, 1, -1, -1)^T$. Then $w^T X w \ge 0$ gives $x_{ik} + x_{jk} - x_{ij} \le 1$. By setting $U := \frac{t}{t-1}((x_{ij})_{i,j=1}^n - \frac{1}{t}\mathbf{J})$, one can now verify that (t, U) is feasible for the program defining $\overline{\vartheta^{+\triangle}}(G)$, which shows the result.

(d) Let λ be an optimum solution for the minimization program defining $\chi^*(G)$ (recall (2.1)). That is, $e^T\lambda = \chi^*(G)$, $\sum_{S \text{ stable}} \lambda_S \chi^S = e$, and $\lambda \ge 0$. For $r \in \mathbb{N}$, the vector $x := \sum_{S \text{ stable}} \lambda_S \chi^{S,r}$ is feasible for (3.4) with objective value $\chi^*(G)$, which shows that $\psi^{(r)}(G) \le \chi^*(G)$. Assume now that $r \ge \alpha(G)$, and consider an optimum solution $M_r(x)$ for (3.4). By setting $y := \frac{1}{\psi^{(r)}(G)}x$, we have $M_r(y) \succeq 0$, $y_{\mathbf{0}} = 1$, and $y_{ij} = 0$ $(ij \in E)$. By using (3.2) we derive $y = \sum_{S \text{ stable}} \lambda_S \chi^{S,2r}$ for some $\lambda_S \ge 0$, with $\sum_S \lambda_S = 1$. By rescaling and taking the projection onto the subspace $\mathbb{R}^V$, we find a decomposition $e = \psi^{(r)}(G)\sum_{S \text{ stable}} \lambda_S \chi^S$, with $\sum_S \lambda_S \psi^{(r)}(G) = \psi^{(r)}(G)$, which shows that $\chi^*(G) \le \psi^{(r)}(G)$.

(e) Take again an optimum solution $M_r(x)$ for (3.4), and let $n = |V(G)|$. Since $M_r(\frac{1}{\psi^{(r)}(G)}x)$ is feasible for (3.1) with objective value $\frac{n}{\psi^{(r)}(G)}$, we get $\mathrm{las}^{(r)}(G) \ge \frac{n}{\psi^{(r)}(G)}$. Assume that G is vertex-transitive. Then there exists an optimum solution x for (3.1) which is invariant under the action of the automorphism group of G. In particular, $x_i = x_j$ for all $i, j \in V$ and thus $x_i = \frac{\mathrm{las}^{(r)}(G)}{n}$ for all $i \in V$. Then the matrix $\frac{n}{\mathrm{las}^{(r)}(G)}M_r(x)$ is feasible for (3.4), yielding $\psi^{(r)}(G) \le \frac{n}{\mathrm{las}^{(r)}(G)}$. ∎

Theorem 3.1 shows that the reciprocity relations (2.5) and (2.2) for the pairs $(\vartheta, \overline{\vartheta}) = (\mathrm{las}^{(1)}, \psi^{(1)})$ and $(\alpha, \chi^*) = (\mathrm{las}^{(r)}, \psi^{(r)})$ (for r large, $r \ge \alpha(G)$) extend to any order r pair $(\mathrm{las}^{(r)}, \psi^{(r)})$ in the hierarchy.

3.3. The hierarchy $\Psi_{\mathrm{las}^{(r)}}(G)$ $(r \ge 0)$ towards the chromatic number. By applying the operator Ψ to the hierarchy $\mathrm{las}^{(r)}(\cdot)$ introduced in section 3.1, we

obtain the following hierarchy of lower bounds for $\chi(G)$:

$$\begin{aligned}\Psi_{\mathrm{las}^{(r)}}(G) &= \min\ t \ \text{ s.t. } \quad \mathrm{las}^{(r)}(G_t) = n \\ &= \min\ t \ \text{ s.t. } \quad y_{\mathbf{0}} = 1,\ \sum_{u\in V(G_t)} y_u = n, \\ &\qquad\qquad\quad y_{uv} = 0\ (uv \in E(G_t)),\ M_r(y) \succeq 0,\end{aligned} \tag{3.5}$$

where the variable y is indexed by $\mathcal{P}_{2r}(V(G_t))$. As $\alpha(G_t) \le n$, we deduce by using (3.3) that $\mathrm{las}^{(n)}(G_t) = \alpha(G_t)$ for all $t \in \mathbb{N}$. Therefore, (1.2) implies the following.

PROPOSITION 3.2. $\Psi_{\mathrm{las}^{(n)}}(G) = \chi(G)$.

In fact, this new hierarchy $\Psi_{\mathrm{las}^{(r)}}$ refines the hierarchy $\psi^{(r)}$.

PROPOSITION 3.3. *For any integer $r \ge 1$, $\psi^{(r)}(G) \le \Psi_{\mathrm{las}^{(r)}}(G)$.*

Proof. Let (t, y) be feasible for the program defining the parameter $\Psi_{\mathrm{las}^{(r)}}(G)$; that is, $y \in \mathbb{R}^{\mathcal{P}_{2r}(V(G_t))}$ satisfies $y_{\mathbf{0}} = 1$, $y_{uv} = 0$ $(uv \in E(G_t))$, $\sum_{u\in V(G_t)} y_u = n$, and $M_r(y) \succeq 0$. We may assume w.l.o.g. that y is invariant under the action of the symmetric group $\mathrm{Sym}(t)$. The next claim determines y_u for $u \in V(G_t)$.

CLAIM 3.4. *$y_u = \frac{1}{t}$ for all $u \in V(G_t)$.*

Proof. Let X denote the principal submatrix of $M_r(y)$ indexed by $\mathcal{P}_1(V(G_t))$. With respect to the partition of $\mathcal{P}_1(V(G_t)) \sim \{\mathbf{0}\} \cup V(G_t)$ into $\{\mathbf{0}\} \cup V_1 \cup \cdots \cup V_t$ (recall (1.1)), the matrix X has the block form

$$\begin{pmatrix} 1 & a^T & a^T & \dots & a^T \\ a & A & B & \dots & B \\ a & B & A & \dots & B \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a & \underbrace{B}_{t \text{ blocks}} & B & \dots & A \end{pmatrix}, \tag{3.6}$$

where $a = \mathrm{diag}(A)$, $\mathrm{diag}(B) = 0$, $A_{ij} = 0$ for $ij \in E(G)$, and $e^T a = \frac{n}{t}$. By taking the Schur complement with respect to the left upper corner and using Lemma 2.8, we have $A + (t-1)B - taa^T \succeq 0$. This implies that $\langle \mathbf{J}, A + (t-1)B\rangle \ge t(e^T a)^2 = \frac{n^2}{t}$. On the other hand, by Lemma 2.9, $\langle \mathbf{J}, A + (t-1)B\rangle \le n\mathrm{Tr}(A + (t-1)B) = n\mathrm{Tr}(A) = \frac{n^2}{t}$. Hence equality holds, implying that $A + (t-1)B = \frac{1}{t}\mathbf{J}$ and thus $a = \frac{1}{t}e$. This shows that $y_u = \frac{1}{t}$ for all $u \in V(G_t)$. ∎

Define the vector $x \in \mathbb{R}^{\mathcal{P}_{2r}(V)}$ with Ith entry $x_I := t y_{\{pi \mid i\in I\}}$ for $I \in \mathcal{P}_{2r}(V) \setminus \{\mathbf{0}\}$ (where p is any fixed integer in $\{1, \dots, t\}$) and $x_{\mathbf{0}} = t$. Then $M_r(x) \succeq 0$, since it coincides with the principal submatrix of $M_r(ty)$ indexed by $\{\mathbf{0}\} \cup \{\{pi \mid i \in I\} \mid I \in \mathcal{P}_r(V)\setminus\{\mathbf{0}\}\}$. Moreover, $x_{\mathbf{0}} = t$ and $x_i = 1$ for $i \in V$. Thus, (t, x) is feasible for the program (3.4), which implies that $\psi^{(r)}(G) \le \Psi_{\mathrm{las}^{(r)}}(G)$. ∎

In summary, we have shown the following relations among the graph parameters $\mathrm{las}^{(r)}(G)$, $\psi^{(r)}(G)$, and $\Psi_{\mathrm{las}^{(r)}}(G)$:

$$\frac{|V(G)|}{\mathrm{las}^{(r)}(G)} \le \psi^{(r)}(G) \le \Psi_{\mathrm{las}^{(r)}}(G) \le \chi(G). \tag{3.7}$$

Let us point out again that, while $\psi^{(r)}(G)$ remains below the fractional chromatic number $\chi^*(G)$, $\Psi_{\mathrm{las}^{(r)}}(G)$ may reach the chromatic number $\chi(G)$.

3.4. Variations of the second order bounds. As observed in Theorem 3.1 and Proposition 3.3, we have

$$\overline{\vartheta^+}(G) \le \overline{\vartheta^{+\triangle}}(G) \le \psi^{(2)}(G) \le \Psi_{\mathrm{las}^{(2)}}(G).$$

To compute $\psi^{(2)}(G)$ one needs to solve a semidefinite program with matrix size $O(n^2)$ and with $O(n^4)$ variables. We now introduce some variations of the parameters $\psi^{(2)}(G)$ and $\Psi_{\mathrm{las}^{(2)}}(G)$ which are less costly to compute but still at least as good as $\overline{\vartheta^+}(G)$. The idea is to consider, instead of the full moment matrix of order 2, some principal submatrix of it. Namely, given $h \in V$, let $M_2(h;x)$ denote the principal submatrix of $M_2(x)$ indexed by the subset $\mathcal{P}_1(V) \cup \{\{h,i\} \mid i \in V\}$ of $\mathcal{P}_2(V)$. Thus in order to define the matrices $M_2(h;x)$ for all $h \in V$, one needs only the components of x indexed by $\mathcal{P}_3(V)$. By following [21], define the following upper bound for the stability number $\alpha(G)$:

$$\ell(G) := \max \sum_{i\in V} x_i \ \text{ s.t. } M_2(h;x) \succeq 0 \ (h \in V),\ x_{\mathbf{0}} = 1,\ x_{ij} = 0 \ (ij \in E(G)), \tag{3.8}$$

with variable $x \in \mathbb{R}^{\mathcal{P}_3(V)}$. Obviously,

$$\mathrm{las}^{(2)}(G) \le \ell(G) \le \mathrm{las}^{(1)}(G) = \vartheta(G).$$

Next, define the graph parameter

$$\begin{aligned} \psi(G) := \min\ t \ \text{ s.t. } \quad & M_2(h;x) \succeq 0 \ (h \in V),\ x_{ij} = 0 \ (ij \in E(G)), \\ & x_{\mathbf{0}} = t,\ x_i = 1 \ (i \in V), \end{aligned} \tag{3.9}$$

where the variable x is indexed by $\mathcal{P}_3(V)$. Again one can avoid variable t by replacing t by $x_{\mathbf{0}}$ in the objective function. We first observe that the pair (ℓ, ψ) satisfies the analogue of the reciprocity relation from Theorem 3.1(e) for the pairs $(\mathrm{las}^{(r)}, \psi^{(r)})$.

PROPOSITION 3.5. *We have*

$$\ell(G)\psi(G) \ge |V(G)|, \ \textit{with equality if } G \textit{ is vertex-transitive,} \tag{3.10}$$

$$\overline{\vartheta^+}(G) \le \psi(G) \le \psi^{(2)}(G). \tag{3.11}$$

Proof. The proof for (3.10) is analogous to that of Theorem 3.1(e), and the right inequality in (3.11) is obvious. For the left inequality, let (t,x) be feasible for (3.9). Observe first that $x_{hi} \ge 0$ for all $h,i \in V$, since x_{hi} is the diagonal entry of $M_2(h;x)$ at the $\{h,i\}$th position and $M_2(h;x) \succeq 0$. Let A denote the principal submatrix of $M_2(h;x)$ indexed by V. Then $A = (x_{ij})_{i,j\in V} \ge 0$ and $A - \frac{1}{t}\mathbf{J} \succeq 0$, which implies that $U := \frac{t}{t-1}(A - \frac{1}{t}\mathbf{J})$ is feasible for the program defining $\overline{\vartheta^+}(G)$ (recall (2.7)). □

By applying the operator Ψ to the parameter $\ell(\cdot)$ (introduced in (3.8)), one obtains the lower bound $\Psi_\ell(G)$ for $\chi(G)$, defined as

$$\begin{aligned} \Psi_\ell(G) &= \min_{t\in\mathbb{N}}\ t \ \text{ s.t. } \quad \ell(K_t \square G) = n \\ &= \min_{t\in\mathbb{N}}\ t \ \text{ s.t. } \quad \sum_{u\in V(G_t)} y_u = n,\ y_{uv} = 0 \ (uv \in E(G_t)), \\ &\qquad\qquad\qquad\quad y_{\mathbf{0}} = 1,\ M_2(u;y) \succeq 0 \ (u \in V(G_t)), \end{aligned} \tag{3.12}$$

where the variable y is indexed by $\mathcal{P}_3(V(G_t))$. (Recall that $G_t = K_t \square G$.)

PROPOSITION 3.6. $\psi(G) \le \Psi_\ell(G) \le \Psi_{las^{(2)}}(G)$.

Proof. The right inequality follows from Lemma 2.2(b), and the proof for the left inequality is analogous to that of Proposition 3.3. □

In summary, we have the following analogue of (3.7) about $\ell(G)$, $\psi(G)$, and $\Psi_\ell(G)$:

$$\frac{|V(G)|}{\ell(G)} \le \psi(G) \le \Psi_\ell(G) \le \chi(G). \tag{3.13}$$

Again, $\psi(G) \leq \chi^*(G)$ since $\psi^{(2)}(G) \leq \chi^*(G)$, but $\Psi_\ell(G)$ may sometimes reach $\chi(G)$. The bound $\Psi_\ell(G)$ can be especially useful when the gap between $\chi^*(G)$ and $\chi(G)$ is large, e.g., when $\chi^*(G) \sim \omega(G) < \chi(G)$. We refer to the follow-up paper [14], where such graph instances will be considered (e.g., Kneser graphs) with experimental results. One can easily verify that the graph parameter $\ell(\cdot)$ is monotone nondecreasing with respect to induced subgraphs. Therefore, as explained in section 2.4, one can compute $\Psi_\ell(G)$ by evaluating $\ell(G_t)$ for $O(\log n)$ queries of t. We will show in the follow-up paper [14] how to give a more compact reformulation for the program (3.12) when G is a vertex-transitive graph. Namely, we will show there that each $\ell(G_t)$ can be computed via a semidefinite program involving four matrices of size $2n+1$, $2n$, n, and n, respectively.

3.5. Link with copositive programming-based hierarchies. We have just seen one possible construction for hierarchies of bounds towards $\alpha(G)$ and $\chi^*(G)$, based on the method of Lasserre. As mentioned earlier in this section there are several other possible constructions for approximating the stable set problem. However, to the best of our knowledge, such constructions were much less investigated for the coloring problem. Recently Dukanovic and Rendl [9] investigated a hierarchy of lower bounds for $\chi^*(G)$, which is closely related to the hierarchy of de Klerk and Pasechnik [6] for $\alpha(G)$; both are based on copositive programming and some of its tractable relaxations in terms of sums of squares of polynomials, proposed by Parrilo [29]. Let $\mathcal{C}_n$ denote the cone of $n \times n$ copositive matrices and $\mathcal{C}_n^*$ its dual cone, consisting of the completely positive matrices. Thus $M \in \mathcal{C}_n$ if and only if $p_M(x) := \sum_{i,j=1}^n M_{ij} x_i^2 x_j^2 \geq 0$ for all $x \in \mathbb{R}^n$. Obviously if, for some $r \in \mathbb{N}$, the polynomial $p_M(x)(\sum_{i=1}^n x_i^2)^r$ can be written as a sum of squares of polynomials (s.o.s. for short), then $M \in \mathcal{C}_n$. By following Parrilo [29], for an integer $r \geq 0$, define the cone

$$K_n^{(r)} := \left\{ M \in \mathbb{R}^{n \times n} \mid p_M(x) \left(\sum_{i=1}^n x_i^2 \right)^r \text{ is s.o.s.} \right\}.$$

Thus, $K_n^{(r)} \subseteq K_n^{(r+1)} \subseteq \mathcal{C}_n$. By following [6], define the graph parameter

$$\vartheta^{(r)}(G) := \min\ t \ \text{ s.t. } t(\mathbf{I} + A_G) - \mathbf{J} \in K_n^{(r)}.$$

In view of (2.30), $\alpha(G) \leq \vartheta^{(r)}(G)$. Moreover, it is proved in [6] that $\vartheta^{(0)}(G) = \vartheta'(G)$ (defined in (2.6)) and $\lfloor \vartheta^{(r)}(G) \rfloor = \alpha(G)$ for $r \geq (\alpha(G))^2$. Dukanovic and Rendl [9] propose an analogous hierarchy toward the fractional chromatic number. To start with, they show the following copositive programming formulation for $\chi^*(G)$:

$$\chi^*(G) = \min\ t \ \text{ s.t. } \begin{array}{l} X_{ii} = t\ (i \in V),\ X_{ij} = 0\ (ij \in E(G)), \\ X \in \mathcal{C}_n^*,\ X - \mathbf{J} \succeq 0. \end{array} \tag{3.14}$$

For an integer $r \geq 0$, let $\kappa^{(r)}(G)$ denote the graph parameter obtained by replacing the cone $\mathcal{C}_n$ by its subcone $K_n^{(r)}$ in (3.14). Thus, $\kappa^{(r)}(G) \leq \kappa^{(r+1)}(G) \leq \chi^*(G)$. Moreover, it is proved in [9] that $\kappa^{(0)}(G) = \overline{\vartheta^+}(G)$ (defined in (2.7)) and that the pair $(\vartheta^{(r)}, \kappa^{(r)})$ satisfies the reciprocity relation:

$$\vartheta^{(r)}(G)\kappa^{(r)}(G) \geq |V(G)|, \text{ with equality if } G \text{ is vertex-transitive,} \tag{3.15}$$

thus extending (2.8) for the case $r = 0$.

Now one may wonder what the link is between the two hierarchies $\mathrm{las}^{(r)}$ and $\vartheta^{(r)}$ for α and between the two hierarchies $\psi^{(r)}$ and $\kappa^{(r)}$ for χ^*. Here is what we can say about this. In order to be able to compare the various bounds we have to add nonnegativity to the definition of $\mathrm{las}^{(r)}$ and $\psi^{(r)}$; namely, let $\mathrm{las}^{(r)}_{\geq 0}(G)$ (resp., $\psi^{(r)}_{\geq 0}(G)$, $\ell_{\geq 0}(G)$, and $\psi_{\geq 0}(G)$) denote the parameter obtained by adding the condition $x \geq 0$ to program (3.1) (resp., to (3.4), (3.8), and (3.9)). The analogue of Theorem 3.1(e) holds for the pairs $(\mathrm{las}^{(r)}_{\geq 0}, \psi^{(r)}_{\geq 0})$ and $(\ell_{\geq 0}, \psi_{\geq 0})$ as well, and we have $\mathrm{las}^{(1)}_{\geq 0}(G) = \vartheta'(G) = \vartheta^{(0)}(G)$ and $\psi^{(1)}_{\geq 0}(G) = \overline{\vartheta}^{+}(G) = \kappa^{(0)}(G)$. It is shown in [13] that, for any graph G,

$$\mathrm{las}^{(r)}_{\geq 0}(G) \leq \vartheta^{(r-1)}(G) \quad \text{for all } r \geq 1,$$

and the same proof technique also shows that $\ell_{\geq 0}(G) \leq \vartheta^{(1)}(G)$ (see [12] for details). In view of the reciprocity relations for the pairs $(\ell_{\geq 0}, \psi_{\geq 0})$, $(\mathrm{las}^{(r)}_{\geq 0}, \psi^{(r)}_{\geq 0})$, and $(\vartheta^{(r)}, \kappa^{(r)})$, this implies that

$$\kappa^{(1)}(G) \leq \psi_{\geq 0}(G),\ \kappa^{(r-1)}(G) \leq \psi^{(r)}_{\geq 0}(G)\ (r \geq 1), \quad \text{when } G \text{ is vertex-transitive.}$$

It is an open question to determine whether the above inequalities remain valid when G is not vertex-transitive. See [9, 14] for instances of Hamming graphs (which are indeed vertex-transitive) having a substantial gap between the two bounds $\kappa^{(1)}(G)$ and $\psi_{\geq 0}(G)$.

Acknowledgments. We are very grateful to two referees for their careful reading and for their useful suggestions, which helped improve the presentation of the paper. We also thank A. Schrijver for bringing the reduction relation (1.2) to our attention.

REFERENCES

[1] M. Bellare and M. Sudan, *Improved non-approximability results*, in Proceedings of the 26th Annual ACM Symposium on Theory of Computing, 1994, pp. 184–193.

[2] I. M. Bomze, M. Dür, E. de Klerk, A. Quist, C. Roos, and T. Terlaky, *On copositive programming and standard quadratic optimization problems*, J. Global Optim., 18 (2000), pp. 301–320.

[3] J. A. Bondy and P. Hell, *A note on the star chromatic number*, J. Graph Theory, 14 (1990), pp. 479–482.

[4] V. Chvátal, *Edmonds polytopes and a hierarchy of combinatorial problems*, Discrete Math., 4 (1973), pp. 305–337.

[5] E. de Klerk, D. V. Pasechnik, and J. P. Warners, *On approximate graph colouring and MAX-k-CUT algorithms based on the ϑ-function*, J. Comb. Optim., 8 (2004), pp. 267–294.

[6] E. de Klerk and D. V. Pasechnik, *Approximation of the stability number of a graph via copositive programming*, SIAM J. Optim., 12 (2002), pp. 875–892.

[7] I. M. Diaz and P. Zabala, *A branch-and-cut algorithm for graph coloring*, Discrete Appl. Math., 154 (2006), pp. 826–847.

[8] I. Dukanovic and F. Rendl, *A semidefinite programming-based heuristic for graph coloring*, Discrete Appl. Math., 156 (2008), pp. 180–189.

[9] I. Dukanovic and F. Rendl, *Copositive programming motivated bounds on the clique and the chromatic number*, Optimization Online (2006), available online from http://www.optimization-online.org/DB_HTML/2006/05/1403.html.

[10] P. Erdös, Z. Füredi, A. Hajnal, P. Komjáth, V. Rödl, and Á. Seress, *Coloring graphs with locally few colors*, Discrete Math., 59 (1986), pp. 21–34.

[11] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, Freeman, San Francisco, 1979.

[12] N. Gvozdenović, *Approximating the Stabiliy Number and the Chromatic Number of Graph Via Semidefinite Programming*, Ph.D. thesis, in preparation.

[13] N. Gvozdenović and M. Laurent, *Semidefinite bounds for the stability number of a graph via sums of squares of polynomials*, Math. Program., 110 (2007), pp. 145–173 (extended abstract in Lecture Notes in Comput. Sci. 3509, M. Jünger and V. Kaibel, eds., Springer, Berlin, pp. 136–151).

[14] N. Gvozdenović and M. Laurent, *Computing semidefinite programming lower bounds for the (fractional) chromatic number via block-diagonalization*, SIAM J. Optim., 19 (2008), pp. 592–615.

[15] G. Hahn, P. Hell, and S. Poljak, *On the ultimate independence ratio of a graph*, European J. Combin., 16 (1995), pp. 253–261.

[16] P. Hell, X. Yu, and H. Zhou, *Independence ratios of graph powers*, Discrete Math., 127 (1994), pp. 213–220.

[17] D. Karger, R. Motwani, and M. Sudan, *Approximate graph coloring by semidefinite programming*, J. ACM, 45 (1998), pp. 246–265.

[18] J. Körner, C. Pilotto, and G. Simonyi, *Local chromatic number and Sperner capacity*, J. Combin. Theory Ser. B, 95 (2005), pp. 101–117.

[19] J. B. Lasserre, *An explicit exact SDP relaxation for nonlinear* $0-1$ *programs*, Lecture Notes in Comput. Sci. 2081, K. Aardal and A.M.H. Gerards, eds., 2001, Springer, Berlin, pp. 293–303.

[20] M. Laurent, *A comparison of the Sherali-Adams, Lovász-Schrijver and Lasserre relaxations for* $0-1$ *programming*, Math. Oper. Res., 28 (2003), pp. 470–496.

[21] M. Laurent, *Strengthened semidefinite programming bounds for codes*, Math. Program., 109 (2007), pp. 239–261.

[22] M. Laurent and F. Rendl, *Semidefinite programming and integer programming*, in Handbook on Discrete Optimization, K. Aardal, G. Nemhauser, and R. Weismantel, eds., Elsevier B. V., Amsterdam, 2005, pp. 393–514.

[23] L. Lovász, *On the Shannon capacity of a graph*, IEEE Trans. Inform. Theory, 25 (1979), pp. 1–7.

[24] L. Lovász and A. Schrijver, *Cones of matrices and set-functions and* $0-1$ *optimization*, SIAM J. Optim., 1 (1991), pp. 166–190.

[25] Y. Lund and M. Yannakakis, *On the hardness of approximating minimization problems*, J. Assoc. Comput. Mach., 41 (1994), pp. 960–981.

[26] R. J. McEliece, E. R. Rodemich, and H. C. Rumsey, *The Lovász' bound and some generalizations*, J. Combin. Inform. System Sci., 3 (1978), pp. 134–152.

[27] P. Meurdesoif, *Strengthening the Lovász* $\theta(\overline{G})$ *bound for graph coloring*, Math. Program., 102 (2005), pp. 577–588.

[28] T. S. Motzkin and E. G. Straus, *Maxima for graphs and a new proof of a theorem of Túran*, Canad. J. Math., 17 (1965), pp. 533–540.

[29] P. Parrilo, *Structured Semidefinite Programs and Semialgebraic Geometry Methods in Robustness and Optimization*, Ph.D. thesis, California Institute of Technology, 2000.

[30] J. Peña, J. Vera, and L. Zuluaga, *Computing the stability number of a graph via linear and semidefinite programming*, SIAM J. Optim., 18 (2007), pp. 87–105.

[31] S. Poljak, *A note on stable sets and colorings of graphs*, Comment. Math. Univ. Carolin., 15 (1974), pp. 307–309.

[32] A. Schrijver, *A comparison of the Delsarte and Lovász bounds*, IEEE Trans. Inform. Theory, 25 (1979), pp. 425–429.

[33] A. Schrijver, *Combinatorial Optimization - Polyhedra and Efficiency*, Springer, Berlin, 2003.

[34] H. D. Sherali and W. P. Adams, *A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems*, SIAM J. Discrete Math., 3 (1990), pp. 411–430.

[35] G. Simonyi, *Asymptotic values of the Hall-ratio for graph powers*, Discrete Math., 306 (2006), pp. 2593–2601.

[36] M. Szegedy, *A note on the theta number of Lovász and the generalized Delsarte bound*, in Proceedings of the 35th IEEE Annual Symposium on Foundations of Computer Science, 1994, pp. 36–39.

[37] A. Vince, *Star chromatic number*, J. Graph Theory, 12 (1988), pp. 551–559.

[38] H. Wolkowicz, R. Saigal, and L. Vandenberghe, *Handbook of Semidefinite Programming*, Kluwer Academic, Boston, 2000.

[39] X. Zhu, *Circular chromatic number: A survey*, Discrete Math., 229 (2001), pp. 371–410.

SIAM J. OPTIM.
Vol. 19, No. 2, pp. 592–615

COMPUTING SEMIDEFINITE PROGRAMMING LOWER BOUNDS FOR THE (FRACTIONAL) CHROMATIC NUMBER VIA BLOCK-DIAGONALIZATION*

NEBOJŠA GVOZDENOVIĆ† AND MONIQUE LAURENT†

Abstract. Recently we investigated in [*SIAM J. Optim.*, 19 (2008), pp. 572–591] hierarchies of semidefinite approximations for the chromatic number $\chi(G)$ of a graph G. In particular, we introduced two hierarchies of lower bounds: the "ψ"-hierarchy converging to the fractional chromatic number and the "Ψ"-hierarchy converging to the chromatic number of a graph. In both hierarchies the first order bounds are related to the Lovász theta number, while the second order bounds would already be too costly to compute for large graphs. As an alternative, relaxations of the second order bounds are proposed in [*SIAM J. Optim.*, 19 (2008), pp. 572–591]. We present here our experimental results with these relaxed bounds for Hamming graphs, Kneser graphs, and DIMACS benchmark graphs. Symmetry reduction plays a crucial role as it permits us to compute the bounds by using more compact semidefinite programs. In particular, for Hamming and Kneser graphs, we use the explicit block-diagonalization of the Terwilliger algebra given by Schrijver [*IEEE Trans. Inform. Theory*, 51 (2005), pp. 2859–2866]. Our numerical results indicate that the new bounds can be much stronger than the Lovász theta number. For some of the DIMACS instances we improve the best known lower bounds significantly.

Key words. chromatic number, Lovász theta number, semidefinite programming, Terwilliger algebra, Hamming graph, Kneser graph

AMS subject classifications. 05C15, 90C27, 90C22

DOI. 10.1137/070683520

1. Introduction.

The chromatic number $\chi(G)$ of a graph G is the smallest number of colors needed to color the vertices of G so that no two adjacent vertices share the same color. Determining $\chi(G)$ is an NP-hard problem [14], and it is hard to approximate $\chi(G)$ within $|V(G)|^{1/14-\epsilon}$ for any $\epsilon > 0$ [1]. Finding a proper vertex coloring with a small number of colors is essential in many real-world applications. A lot of work has been done in order to develop efficient heuristics for this problem (see, e.g., [5]). Nevertheless, these methods can provide us only with upper bounds on the chromatic number. Lower bounds were mainly obtained by using linear programming [26, 27], critical subgraphs [8], and semidefinite programming (SDP) [9, 10, 11, 18, 28, 32]. The semidefinite approaches are based on computing (variations of) the well-known lower bound $\overline{\vartheta}(G) := \vartheta\left(\overline{G}\right)$, the theta number of the complementary graph, introduced by Lovász [24]. The theta number satisfies the "sandwich inequality":

$$\omega(G) \leq \overline{\vartheta}(G) \leq \chi(G),$$

and it can be computed to any arbitrary precision in polynomial time since it can be formulated via a semidefinite program of size $|V(G)|$. Here $\omega(G)$ is the clique number of G, defined as the maximum size of a clique (i.e., a set of pairwise adjacent nodes) in G, the stability number $\alpha(G) := \omega(\overline{G})$ of G being the maximum size of

*Received by the editors February 23, 2007; accepted for publication (in revised form) January 22, 2008; published electronically July 2, 2008. Supported by the Netherlands Organization for Scientific Research grant NWO 639.032.203 and by ADONET, Marie Curie Research Training Network MRTN-CT-2003-504438.

http://www.siam.org/journals/siopt/19-2/68352.html

†Centrum voor Wiskunde en Informatica, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands (N.Gvozdenovic@cwi.nl, M.Laurent@cwi.nl).

a stable set (i.e., a set of pairwise nonadjacent nodes) in G. The theta number has been strengthened towards the chromatic number by using nonnegativity [32], triangle inequalities [28], or some lift-and-project methods [11]. Computational results were reported in [9, 10, 11]. A common feature shared by all of these bounds is that they remain below the fractional chromatic number $\chi^*(G)$. Thus they are of little use when $\chi^*(G)$ is close to the clique number $\omega(G)$. In [17] the authors investigated another type of lift-and-project approach leading to a hierarchy of bounds converging to the chromatic number $\chi(G)$. We explore in the present follow-up paper the behavior of these bounds through experimental results on several classes of graphs.

The approach in [17] is based on the following reduction of Chvátal [6] of the chromatic number to the stability number:

$$\chi(G) \leq t \Longleftrightarrow \alpha(K_t \Box G) = |V(G)|, \tag{1.1}$$

where $K_t \Box G$ denotes the Cartesian product of K_t, the complete graph on t nodes, and the graph G. For a given graph parameter $\beta(\cdot)$ satisfying $\alpha(\cdot) \leq \beta(\cdot) \leq \overline{\chi}(\cdot)$, define the new graph parameter $\Psi_\beta(\cdot)$ by

$$\Psi_\beta(G) := \min_{t \in \mathbb{N}} t \ \text{ s.t. } \beta(K_t \Box G) = |V(G)|. \tag{1.2}$$

As shown in [17], the operator Ψ is monotone nonincreasing and satisfies

$$\omega(G) = \Psi_{\overline{\chi}}(G) \leq \Psi_\beta(G) \leq \Psi_\alpha(G) = \chi(G) \text{ and } \Psi_\vartheta(G) = \left\lceil \overline{\vartheta}(G) \right\rceil. \tag{1.3}$$

In other words the operator Ψ transforms upper bounds for the stability number into lower bounds for the chromatic number. An interesting bound for $\alpha(\cdot)$ from the computational point of view is the graph parameter $\ell(\cdot)$ introduced by Laurent [21] as a relaxation of the second order bound in Lasserre's hierarchy for $\alpha(\cdot)$ (see [19, 21]). Two hierarchies for the chromatic number, related to Lasserre's hierarchy for $\alpha(\cdot)$, are studied in [17], as well as two bounds $\psi(\cdot)$ and $\Psi_\ell(\cdot)$ related to the parameter $\ell(\cdot)$. See section 2.2 for the precise definition of the parameters ℓ, ψ, and Ψ_ℓ.

In the present paper we investigate how to compute the bounds $\psi(\cdot)$ and $\Psi_\ell(\cdot)$ for Hamming graphs and for Kneser graphs. Coloring Hamming graphs is of interest, e.g., to the Borsuk problem (see [33]), and the chromatic number of Kneser graphs was computed in the celebrated paper [23] of Lovász by using topological methods; see, e.g., [25] for a study of topological lower bounds for the chromatic number. The Hamming graph $G = H(n, \mathcal{D})$ has node set $V(G) = \{0,1\}^n$, with an edge uv if the Hamming distance between u and v lies in the given set $\mathcal{D} \subseteq \{1, \ldots, n\}$. For $n \geq 2r$, the Kneser graph $K(n, r)$ is the subgraph of $H(n, \{2r\})$ induced by the set of words $u \in \{0,1\}^n$ with weight $\sum_{i=1}^n u_i = r$. The Hamming graph has a large automorphism group which enables us to block-diagonalize and reformulate the programs for $\psi(G)$ and $\Psi_\ell(G)$ in such a way that they involve $O(n)$ matrices of size $O(n)$ (instead of $2^n = |V(G)|$). As a crucial ingredient we use the block-diagonalization for the Terwilliger algebra given by Schrijver [31]. We also use this technique, which was extended to constant-weight codes in [31], for computing the bound $\Psi_\ell(\cdot)$ for Kneser graphs. For Kneser graphs, the bound $\psi(\cdot)$ coincides with the fractional chromatic number (see section 4), but, as will be seen in Table 2, $\Psi_\ell(K(n, r))$ can go beyond the fractional chromatic number. We report experimental results for Hamming and Kneser graphs in Tables 1 and 2. For some instances, the parameter $\psi(G)$ improves substantially the theta number $\bar{\vartheta}(G)$, and adding nonnegativity may also help; moreover, while $\Psi_\ell(G)$ hardly improves upon $\psi(G)$ for Hamming graphs, it does give an improvement for Kneser graphs.

Finally we introduce a further variation $\psi_K(G)$ of our bounds (where K is a clique in G), which can be especially useful for graphs without apparent symmetries. By using a simple block-diagonalization argument, $\psi_K(G)$ can be formulated via a semidefinite program involving $|K|$ matrices of size $|V(G)|$ and one matrix of size $|V(G)|+1$. The bound $\psi_K(G)$ is bounded above by the fractional chromatic number $\chi^*(G)$. We report experimental results on some DIMACS benchmark instances. To the best of our knowledge, our lower bound improves the best known lower bound in the literature for several instances of DSJC and DSJR graphs, sometimes substantially. Moreover, for the two instances $G =$ DSJC125.9 and DSJR500.1c, we can determine the exact value of the chromatic number $\chi(G)$, since our lower bound matches the known upper bound for $\chi(G)$. This indicates that the bound ψ_K can be quite strong for random graphs, despite the fact that it remains below the fractional chromatic number. Moreover, we observed experimentally that adding nonnegativity constraints to the formulation of ψ_K does not help for the DSJC instances, which is similar to the observation made in [9] that strengthening the theta number with nonnegativity does not help for random graphs.

More details about the results of this paper can also be found in [16].

Contents of the paper. In section 2 we recall the definitions of the graph parameters $\ell(\cdot)$, $\psi(\cdot)$, and $\Psi_\ell(\cdot)$ and their main properties; we show how symmetry in the semidefinite programming formulations and in the graph can be exploited to (sometimes dramatically) reduce the sizes of the semidefinite programs defining these bounds. Section 3 is devoted to the computation of the bounds for Hamming graphs; we describe how to block-diagonalize the matrices in the semidefinite programs and report computational experiments. In section 4 we focus on the graph parameter $\Psi_\ell(\cdot)$ for Kneser graphs; we present the block-diagonalization of the matrices and conclude the section with computational results. We describe in section 5 the new lower bound $\psi_K(\cdot)$, which we test on some DIMACS benchmark graphs.

Notation. Given a graph $G = (V, E)$, $\overline{G}$ denotes its complementary graph whose edges are the pairs $uv \notin E(G)$ $(u, v \in V(G),\ u \neq v)$. Given a graph parameter $\beta(\cdot)$, $\overline{\beta}(\cdot)$ is the graph parameter defined by $\overline{\beta}(G) := \beta(\overline{G})$ for any graph G. For two graphs G and G', their Cartesian product $G\square G'$ has node set $V(G) \times V(G')$, with two nodes uu', $vv' \in V(G) \times V(G')$ being adjacent in $G\square G'$ if and only if ($u = v$ and $u'v' \in E(G')$) or ($uv \in E(G)$ and $u' = v'$). For an integer $t \geq 1$, K_t is the complete graph on t nodes. We also set $G_t = K_t\square G$ as a shorthand notation for the Cartesian product of G and K_t.

Throughout, the letters $\mathbf{I}$, $\mathbf{J}$, and e denote, respectively, the identity matrix, the all-ones matrix, and the all-ones vector (of suitable size); $\mathbb{N}$ is the set of nonnegative integers. For matrices A and A' indexed, respectively, by $I \times J$ and $I' \times J'$, their tensor product $A \otimes A'$ is the matrix indexed by $(I \times I') \times (J \times J')$, with $(A \otimes A')_{(i,i'),(j,j')} := A_{i,j}B_{i',j'}$. Moreover, the notation $A \succeq 0$ means that A is a symmetric positive semidefinite matrix.

Given a finite set V, $\mathcal{P}(V)$ denotes the collection of all subsets of V. Given an integer r, set $\mathcal{P}_r(V) := \{I \in \mathcal{P}(V) \mid |I| \leq r\}$; in particular, $\mathcal{P}_1(V) = \{\emptyset, \{i\}\ (i \in V)\}$. Sometimes (e.g., when dealing with Hamming graphs) we deal with the collection $\mathcal{P}_1(V)$, where $V = \mathcal{P}(N)$, and $N = \{1, \ldots, n\}$; then $\mathcal{P}_1(V)$ contains $\emptyset$ (the empty subset of V) and $\{\emptyset\}$ (the singleton subset of V consisting of the empty subset of N). To avoid confusion we use the symbol $\mathbf{0}$ to denote the empty subset of V, so that $\mathcal{P}_1(V) = \{\mathbf{0}, \{i\}\ (i \in V)\}$. We sometimes identify $\mathcal{P}_1(V) \setminus \{\mathbf{0}\}$ with V; i.e., we

write $\{i\}$ as i and $\{i,j\}$ as ij, and, given a vector $x \in \mathbb{R}^{\mathcal{P}(V)}$, we also set $x_i := x_{\{i\}}$, $x_{ij} := x_{\{i,j\}}$, $x_{ijk} := x_{\{i,j,k\}}$ (for $i,j,k \in V$), etc.

Let V be a finite set, and let $\mathcal{G}$ be a subgroup of $\mathrm{Sym}(V)$, the group of permutations of V, also denoted as $\mathrm{Sym}(n)$ if $|V| = n$. Then $\mathcal{G}$ acts on $\mathcal{P}(V)$ by letting $\sigma(I) := \{\sigma(i) \mid i \in I\}$ for $I \subseteq V$, $\sigma \in \mathcal{G}$. Moreover, $\mathcal{G}$ acts on vectors and matrices indexed by $\mathcal{P}_r(V)$, by letting $\sigma(x) := (x_{\sigma(I)})_{I \in \mathcal{P}_r(V)}$, $\sigma(M) := (M_{\sigma(I),\sigma(J)})_{I,J \in \mathcal{P}_r(V)}$ for $x \in \mathbb{R}^{\mathcal{P}_r(V)}$, $M \in \mathbb{R}^{\mathcal{P}_r(V) \times \mathcal{P}_r(V)}$, and $\sigma \in \mathcal{G}$. One says that M is invariant under the action of $\mathcal{G}$ if $\sigma(M) = M$ for all $\sigma \in \mathcal{G}$; then the matrix $\frac{1}{|\mathcal{G}|!} \sum_{\sigma \in \mathcal{G}} \sigma(M)$, the "symmetrization" of M obtained by applying the Reynolds operator, is invariant under the action of $\mathcal{G}$. The analogue statement holds for vectors. A semidefinite program is said to be invariant under the action of $\mathcal{G}$ if, for any feasible matrix X and any $\sigma \in \mathcal{G}$, the matrix $\sigma(X)$ is again feasible with the same objective value; then the optimum value of the program remains unchanged if we restrict to invariant feasible solutions, and, in particular, there is an invariant optimal solution.

The automorphism group $\mathrm{Aut}(G)$ of a graph $G = (V,E)$ consists of all $\sigma \in \mathrm{Sym}(V)$ preserving the set of edges. G is said to be vertex-transitive when, given any two nodes $i,j \in V$, there exists $\sigma \in \mathrm{Aut}(G)$, for which $\sigma(i) = j$. For instance, for the graph $G_t = K_t \square G$, $\mathrm{Sym}(t) \times \mathrm{Aut}(G) \subseteq \mathrm{Aut}(G_t)$, where $(\tau,\sigma) \in \mathrm{Sym}(t) \times \mathrm{Aut}(G)$ acts on $V(G_t)$ (and thus on $\mathcal{P}_r(V(G_t))$ for $r \in \mathbb{N}$) by $(\tau,\sigma)(p,i) = (\tau(p),\sigma(i))$ for $(p,i) \in V(K_t) \times V(G)$. We will deal in this paper with semidefinite programs involving matrices indexed by $\mathcal{P}_r(V(G_t))$, which are invariant under this action of $\mathrm{Sym}(t) \times \mathrm{Aut}(G)$.

2. Graph parameters.

2.1. Classic bounds. We recall here some classic bounds for the chromatic number $\chi(G)$ of a graph $G = (V,E)$. Throughout section 2, $V = V(G)$ is the node set of graph G and $n := |V(G)|$. (For details see, e.g., [17, 22, 30].)

- The *fractional chromatic number of* G:

$$\chi^*(G) := \max\ e^T x \ \text{ s.t. } \sum_{i \in S} x_i \le 1 \ (S \text{ stable}), \quad x \in \mathbb{R}^V_+. \tag{2.1}$$

It is well known (and easy to verify) that $\omega(G) \le \chi^*(G) \le \chi(G)$, and

$$\alpha(G)\chi^*(G) \ge |V(G)| \ \text{ with equality when } G \text{ is vertex-transitive.} \tag{2.2}$$

- *Lovász's theta number* (introduced in [24]):

$$\begin{aligned} \overline{\vartheta}(G) = \vartheta(\overline{G}) := \quad & \max \quad e^T Y e \\ & \text{s.t.} \quad \sum_{i \in V} Y_{ii} = 1, \\ & \qquad Y_{ij} = 0 \ \ (ij \in E(\overline{G})), \\ & \qquad Y \succeq 0, \end{aligned} \tag{2.3}$$

where Y is a symmetric matrix indexed by V. For a later purpose we recall the following equivalent formulation (cf., e.g., [15, Theorem 9.3.12]):

$$\begin{aligned} \overline{\vartheta}(G) = \quad & \min \quad X_{\mathbf{00}} \\ & \text{s.t.} \quad X_{ii} = X_{\mathbf{0}i} \ (i \in V), \\ & \qquad X_{ij} = 0 \ (ij \in E(G)), \\ & \qquad X \succeq 0, \end{aligned} \tag{2.4}$$

where the matrix variable X is indexed by the set $\mathcal{P}_1(V) = V \cup \{\mathbf{0}\}$. Lovász [24] proved the following analogue of (2.2) for the pair $(\vartheta, \overline{\vartheta})$:

$$\vartheta(G)\overline{\vartheta}(G) \geq |V(G)| \text{ with equality when } G \text{ is vertex-transitive.} \tag{2.5}$$

• *Szegedy's number* was first defined in [32]. We present the following equivalent formulation from [17]:

$$\begin{array}{rlll} \overline{\vartheta^+}(G) = \vartheta^+(\overline{G}) = & \min & X_{\mathbf{00}} \\ & \text{s.t.} & X_{ii} = X_{\mathbf{0}i} \ (i \in V), \\ & & X_{ij} = 0 \ (ij \in E(G)), \\ & & X \geq 0, \ X \succeq 0. \end{array} \tag{2.6}$$

The above parameters satisfy

$$\omega(G) \leq \overline{\vartheta}(G) \leq \overline{\vartheta^+}(G) \leq \chi^*(G) \leq \chi(G).$$

2.2. The bounds ℓ, ψ, and Ψ_ℓ. We review here the graph parameters $\ell(\cdot)$ proposed in [21] and $\psi(\cdot)$ and $\Psi_\ell(\cdot)$ proposed in [17]; for details see also [16]. For a subset $S \subseteq V$ and an integer $r \geq 1$, define the vectors $\chi^S \in \{0,1\}^V$, with ith entry 1 if and only if $i \in S$ (for $i \in V$), and $\chi^{S,r} \in \{0,1\}^{\mathcal{P}_r(V)}$, with Ith entry 1 if and only if $I \subseteq S$ (for $I \in \mathcal{P}_r(V)$). Given a vector $x = (x_I)_{I \in \mathcal{P}_{2r}(V)}$, consider the matrix:

$$M_r(x) := (x_{I \cup J})_{I,J \in \mathcal{P}_r(V)}$$

known as the *(combinatorial) moment matrix* of x of order r. Consider the programs:

$$\text{las}^{(r)}(G) := \max \sum_{i \in V} x_i \ \text{ s.t. } M_r(x) \succeq 0, \ x_{\mathbf{0}} = 1, \ x_{ij} = 0 \ (ij \in E), \tag{2.7}$$

$$\psi^{(r)}(G) := \min \ t \ \text{ s.t. } M_r(x) \succeq 0, \ x_{\mathbf{0}} = t, \ x_i = 1 \, (i \in V), \ x_{ij} = 0 \ (ij \in E), \tag{2.8}$$

where the variable x is indexed by $\mathcal{P}_{2r}(V)$. Note that the variable t can be avoided in (2.8) by replacing t by $x_{\mathbf{0}}$ in the objective function. We choose this formulation to emphasize the analogy with the formulations (2.13), (2.17), and (5.1) below. The above two programs were studied, respectively, in [19, 20] and in [17]. In particular, the following holds:

$$\alpha(G) = \text{las}^{(\alpha(G))} \leq \cdots \leq \text{las}^{(r+1)}(G) \leq \text{las}^{(r)}(G) \leq \cdots \leq \text{las}^{(1)}(G) = \vartheta(G), \tag{2.9}$$

$$\vartheta(\overline{G}) = \psi^{(1)}(G) \leq \cdots \leq \psi^{(r)}(G) \leq \psi^{(r+1)}(G) \leq \cdots \leq \psi^{(\alpha(G))}(G) = \chi^*(G), \tag{2.10}$$

$$\psi^{(r)}(G)\text{las}^{(r)}(G) \geq |V(G)| \text{ with equality if } G \text{ is vertex-transitive.} \tag{2.11}$$

Thus the parameters $\text{las}^{(r)}(G)$ (for $r = 1, \ldots, \alpha(G)$) create a hierarchy of upper bounds for the stability number, while the parameters $\psi^{(r)}(G)$ create a hierarchy of lower bounds for the fractional coloring number. Theoretically, the parameters $\text{las}^{(r)}(G)$ and $\psi^{(r)}(G)$ can be computed to any precision in polynomial time for fixed r, since the semidefinite programs (2.7) and (2.8) involve matrices of size $O(n^r)$. On the other hand, in practice, we are not able to compute $\text{las}^{(2)}(G)$ or $\psi^{(2)}(G)$ for "interesting" graphs, that is, for graphs of reasonably large size. For this reason some variations of the parameters $\text{las}^{(2)}(G)$ and $\psi^{(2)}(G)$ were proposed in [17, 21]. The idea is to consider, instead of the full moment matrix of order 2, a number of principal submatrices of

it. Given $h \in V$, let $M_2(h;x)$ denote the principal submatrix of $M_2(x)$ indexed by the subset $\mathcal{P}_1(V) \cup \{\{h,i\} \mid i \in V\}$ of $\mathcal{P}_2(V)$. Thus in order to define the matrices $M_2(h;x)$ for all $h \in V$, one needs only the components of x indexed by $\mathcal{P}_3(V)$. Following [17, 21], define the upper bound for the stability number $\alpha(G)$:

$$\ell(G) := \max \sum_{i \in V} x_i \ \text{ s.t. } M_2(h;x) \succeq 0 \ (h \in V),\ x_{\mathbf{0}} = 1,\ x_{ij} = 0 \ (ij \in E(G)), \tag{2.12}$$

and the lower bound for the fractional coloring number $\chi^*(G)$:

$$\begin{aligned} \psi(G) := \min\ t \ \text{ s.t. } \quad & M_2(h;x) \succeq 0 \ (h \in V),\ x_{ij} = 0 \ (ij \in E(G)), \\ & x_{\mathbf{0}} = t,\ x_i = 1 \ (i \in V), \end{aligned} \tag{2.13}$$

where the variable x is indexed by $\mathcal{P}_3(V)$. For the parameter $\ell(G)$ we have (see [21])

$$\alpha(G) \le \mathrm{las}^{(2)}(G) \le \ell(G) \le \mathrm{las}^{(1)}(G) = \vartheta(G) \le \overline{\chi}(G), \tag{2.14}$$

while $\psi(G)$ satisfies (see [17])

$$\overline{\vartheta^+}(G) \le \psi(G) \le \psi^{(2)}(G). \tag{2.15}$$

They also satisfy an inequality similar to (2.11), namely,

$$\psi(G)\ell(G) \ge |V(G)| \text{ with equality if } G \text{ is vertex-transitive.} \tag{2.16}$$

As $\alpha(\cdot) \le \ell(\cdot) \le \overline{\chi}(\cdot)$ (by (2.14)), we can apply the operator Ψ from (1.2) to $\ell(\cdot)$ and obtain the lower bound $\Psi_\ell(G)$ for $\chi(G)$, defined as

$$\Psi_\ell(G) = \min_{t \in \mathbb{N}} t \ \text{ s.t. } \ell(G_t) = n. \tag{2.17}$$

The parameter $\ell(G_t)$ is defined via the program

$$\begin{aligned} \ell(G_t) = \max \sum_{u \in V(G_t)} y_u \ \text{ s.t. } \quad & M_2(u;y) \succeq 0 \ (u \in V(G_t)), \\ & y_{\mathbf{0}} = 1,\ y_{uv} = 0 \ (uv \in E(G_t)), \end{aligned} \tag{2.18}$$

where the variable y is indexed by $\mathcal{P}_3(V(G_t))$. (Recall that $G_t = K_t \square G$.) Finally, the two parameters $\psi(G)$ and $\Psi_\ell(G)$ were compared in [17], where the following relation is shown:

$$\overline{\vartheta}(G) \le \psi(G) \le \Psi_\ell(G) \le \chi(G). \tag{2.19}$$

Let us finally note that one can easily strengthen the bounds $\ell(G)$, $\psi(G)$, and $\Psi_\ell(G)$, e.g., by requiring nonnegativity[1] of the variables. Let $\ell_{\ge 0}(G)$ (resp., $\psi_{\ge 0}(G)$) denote the variation of $\ell(G)$ (resp., $\psi(G)$) obtained by adding the condition $x \ge 0$ to

[1]Note, however, that the condition $x_{ij} \ge 0 \ \ \forall i,j \in V$ already automatically holds in (2.12) and (2.13), since it is implied by $M_2(h;x) \succeq 0 \ \forall h \in V$ (as x_{hi} occurs as a diagonal entry of $M_2(h;x)$). Analogously, $y_{uv} \ge 0 \ \ \forall u,v \in V(G_t)$ automatically holds in (2.18).

(2.12) (resp., (2.13)); we have again $\psi_{\geq 0}(G)\ell_{\geq 0}(G) = |V(G)|$ when G is vertex-transitive. Define accordingly $\Psi_{\ell_{\geq 0}}(G)$, which amounts to requiring $y \geq 0$ in (2.18).

2.3. Exploiting symmetry to compute the bounds ℓ, ψ, and Ψ_ℓ. We group here some observations about the complexity of computing the graph parameters $\ell(\cdot)$, $\psi(\cdot)$, and $\Psi_\ell(\cdot)$. We show how one can exploit symmetry, present in the structure of the matrices involved in the programs defining the parameters or in the graph instance, in order to reduce the size of the programs. This symmetry reduction is a crucial step as it allows reformulating the parameters via more compact programs. In this way we will be able to compute the graph parameters for certain large graphs (with as many as 2^{20} nodes for certain Hamming graphs), a task that would obviously be out of reach without applying this symmetry reduction.

We begin with observing that the matrix $M_2(h;x)$, used in definitions (2.12) and (2.13), has a special block structure, whose symmetry can be exploited to "block-diagonalize" it. Recall that $M_2(h;x)$ is indexed by the set $\mathcal{P}_1(V) \cup \{\{h,i\} \mid i \in V\} = \{\mathbf{0}\} \cup \{\{i\} \mid i \in V\} \cup \{\{h,i\} \mid i \in V\}$. Here we keep the two occurrences of the singleton $\{h\}$ in the index set, occurring first as $\{i\}$ for $i = h$ and second as $\{i,h\}$ for $i = h$. Thus, the index set of $M_2(h;x)$ is partitioned into $\{\mathbf{0}\}$ and two copies of V.

LEMMA 2.1. *With respect to this partition of its index set, the matrix $M_2(h;x)$ has the block form:*

$$M_2(h;x) = \begin{pmatrix} a & c^T & d^T \\ c & C & D \\ d & D & D \end{pmatrix}, \tag{2.20}$$

where $a = x_{\mathbf{0}}$, $c_i = x_i$, $d_i = x_{hi}$ $(i \in V)$, $C_{ij} = x_{ij}$, and $D_{ij} = x_{hij}$ $(i,j \in V)$. Then

$$M_2(h;x) \succeq 0 \Longleftrightarrow \begin{pmatrix} a - c_h & c^T - d^T \\ c - d & C - D \end{pmatrix} \succeq 0 \quad \text{and} \quad D \succeq 0. \tag{2.21}$$

Proof. The form (2.20) follows directly from the definition of $M_2(h;x)$. To show (2.21), observe that the row of $M_2(h;x)$ indexed by $\{h\}$ has the form (c_h, d^T, d^T). Indeed, for $i,j \in V$, $C_{ij} = x_{\{i,j\}}$, $D_{ij} = x_{\{h,i,j\}}$, $c_j = x_j$, and $d_j = x_{\{h,j\}}$, implying that $C_{hj} = D_{hj} = d_j$. As in [21], we perform some row/column manipulation on $M_2(h;x)$ to show (2.21). Say the second row/column of $M_2(h;x)$ is indexed by $\{h\}$, i.e., h comes first when listing the elements of V. Then

$$U_1^T M_2(h;x) U_1 = \begin{pmatrix} a - c_h & c^T - d^T & 0 \\ c - d & C & D \\ 0 & D & D \end{pmatrix}, \quad \text{setting } U_1 := \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & \mathbf{I} \end{pmatrix},$$

where $\mathbf{I}$ is the identity matrix of order $2n-1$ ($n = |V|$). Next,

$$U_2^T (U_1^T M_2(h;x) U_1) U_2 = \begin{pmatrix} a - c_h & c^T - d^T & 0 \\ c - d & C - D & 0 \\ 0 & 0 & D \end{pmatrix}, \quad \text{setting } U_2 := \begin{pmatrix} 1 & 0 & 0 \\ 0 & \mathbf{I} & 0 \\ 0 & -\mathbf{I} & \mathbf{I} \end{pmatrix},$$

where $\mathbf{I}$ has order n. □

Hence, in (2.12) and (2.13), we may replace each constraint $M_2(h;x) \succeq 0$ (which involves a matrix of size $2n+1$) by two constraints involving matrices of sizes $n+1$ and n.

We now consider symmetries present in the graph instance G. Observe that the program (2.12) (or (2.13)) is invariant under the action of $\mathrm{Aut}(G)$. Hence one may assume that the variable x is invariant under the action of $\mathrm{Aut}(G)$. Therefore, when G is vertex-transitive, it suffices to require the condition $M_2(h;x) \succeq 0$ for *one* choice of $h \in V$ (instead of for *all* $h \in V$), and thus $\ell(G)$ and $\psi(G)$ can be computed via a semidefinite program involving two linear matrix inequality (LMIs) matrices of sizes $n+1$, n and with $O(n^2)$ variables.

We now turn to the graph parameter $\Psi_\ell(G)$. In order to determine $\Psi_\ell(G)$, we need to compute the parameter $\ell(G_t) = \ell(K_t \square G)$ from (2.18) (for several queries of $t \in \mathbb{N}$). As was just observed above, the program defining $\ell(G_t)$ is invariant under the action of $\mathrm{Aut}(G_t)$ thus in particular under the action of $\mathrm{Sym}(t) \times \mathrm{Aut}(G)$ or simply of $\mathrm{Sym}(t)$. In particular, in program (2.18), one may assume that y is invariant under the action of $\mathrm{Sym}(t)$. Moreover, it suffices to require the condition $M_2(u;y) \succeq 0$ for all $u \in V_1$ instead of for all $u \in V(G_t)$; here $V_1 = \{1i \mid i \in V\}$ denotes the "first layer" of the node set $V(G_t) = \{pi \mid p = 1, \ldots, t,\ i \in V\}$ of G_t. Furthermore, when G is vertex-transitive, it suffices to require $M_2(u;y) \succeq 0$ for *one* choice of $u \in V_1$ instead of for *all* $u \in V_1$.

We now show, by using the invariance of y under the action of $\mathrm{Sym}(t)$, that the matrix $M_2(u;y)$ has a special block structure, whose symmetry can be used to block-diagonalize it. To begin with, with respect to the partition $\{\mathbf{0}\} \cup \{\{v\} \mid v \in V(G_t)\} \cup \{\{u,v\} \mid v \in V(G_t)\}$ of its index set, the matrix $M_2(u;y)$ has the block form shown in (2.20) with a, c, d, C, and D being now defined in terms of y (instead of x). In view of (2.21), we have

$$M_2(u;y) \succeq 0 \Longleftrightarrow \begin{pmatrix} y_{\mathbf{0}} - y_u & c^T - d^T \\ c - d & C - D \end{pmatrix} \succeq 0 \ \text{ and } \ D \succeq 0. \tag{2.22}$$

Next we observe that the invariance of y under $\mathrm{Sym}(t)$ implies a special block structure for the matrices C and D.

LEMMA 2.2. *Consider the partition $V(G_t) = V_1 \cup \cdots \cup V_t$ of the node set of graph G_t, where $V_p := \{pi \mid i \in V\}$ for $p = 1, \ldots, t$. With respect to this partition, the matrices C and D have the block form:*

$$C = \begin{pmatrix} A^1 & A^2 & \cdots & A^2 \\ A^2 & A^1 & \cdots & A^2 \\ \vdots & \vdots & \ddots & \vdots \\ A^2 & \cdots & \cdots & A^1 \end{pmatrix}, \qquad D = \begin{pmatrix} B^1 & B^2 & B^2 & \cdots & B^2 \\ (B^2)^T & B^3 & B^4 & \cdots & B^4 \\ (B^2)^T & B^4 & B^3 & \cdots & B^4 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ (B^2)^T & B^4 & \cdots & \cdots & B^3 \end{pmatrix}, \tag{2.23}$$

where[2] *$A^1, \ldots, B^4 \in \mathbb{R}^{n \times n}$. Moreover, by setting $a_1 := \mathrm{diag}(A^1)$, $b_1 := \mathrm{diag}(B^1)$, and $b_3 := \mathrm{diag}(B^3)$, we have $c = [a_1^T, \ldots, a_1^T]^T$ and $d = [b_1^T\, b_3^T\, b_3^T, \ldots, b_3^T]^T$.*

Proof. Consider $i, j \in V$ and $p, q, p', q' \in \{1, \ldots, t\}$, with $p = q$ if and only if $p' = q'$. Then $C_{pi,qj} = y_{\{pi,qj\}} = y_{\{p'i,q'j\}} = C_{p'i,q'j}$; indeed, as there exists $\sigma \in \mathrm{Sym}(t)$ mapping $\{p,q\}$ to $\{p',q'\}$, the equality $y_{\{pi,qj\}} = y_{\{p'i,q'j\}}$ follows from the fact that y is invariant under the action of $\mathrm{Sym}(t)$. This shows that C has the form indicated in (2.23); the argument is analogous for matrix D. ∎

[2]Here A^i or B^i should not be interpreted as powers of A or B, as i is not an exponent but just an upper index.

To fix ideas, set $u = 1h \in V_1$ (where $h \in V$ is a given node of G). Then the entries of $A^1, \ldots, B^4$ are given by

$$\begin{aligned} A^1_{ij} = y_{\{1i,1j\}},\ A^2_{ij} = y_{\{1i,2j\}},\ B^1_{ij} = y_{\{1i,1h,1j\}},\\ B^2_{ij} = y_{\{1i,1h,2j\}},\ B^3_{ij} = y_{\{2i,1h,2j\}},\ B^4_{ij} = y_{\{2i,1h,3j\}} \end{aligned} \tag{2.24}$$

for $i, j \in V$. (Recall that $y_{\{1i,1j\}} = y_{\{pi,pj\}}$, $y_{\{1i,2j\}} = y_{\{pi,qj\}}$, and $y_{\{1i,2j,3h\}} = y_{\{pi,qj,rh\}}$ for any distinct $p, q, r \in \{1, \ldots, t\}$ since y is invariant under the action of $\mathrm{Sym}(t)$.) Moreover, the edge constraints $y_{uv} = 0$ (for $uv \in E(G_t)$) in (2.18) can be reformulated as

$$\begin{aligned} &A^1_{ij} = 0 \ \text{ if } \ ij \in E(G),\\ &B^1_{ij} = 0 \ \text{ if } \ \{i,j,h\} \text{ contains an edge of } G,\\ &B^2_{ij} = 0 \ \text{ if } \ hi \in E(G) \text{ or } j \in \{i,h\},\\ &B^3_{ij} = 0 \ \text{ if } \ ij \in E(G) \text{ or if } h \in \{i,j\},\\ &B^4_{ij} = 0 \ \text{ if } \ h \in \{i,j\},\\ &\mathrm{diag}(A^2) = \mathrm{diag}(B^2) = \mathrm{diag}(B^4) = 0 \end{aligned} \tag{2.25}$$

for distinct $i, j \in V$.

The next lemma indicates how one can further block-diagonalize the two matrices appearing at the right-hand side of the equivalence in (2.22).

LEMMA 2.3. *We have*

$$D \succeq 0 \Longleftrightarrow \begin{pmatrix} B^1 & (t-1)B^2 \\ (t-1)(B^2)^T & (t-1)B^3 + (t-1)(t-2)B^4 \end{pmatrix},\ B^3 - B^4 \succeq 0.$$

Moreover,

$$\begin{pmatrix} y_{\mathbf{0}} - y_u & c^T - d^T \\ c - d & C - D \end{pmatrix} \succeq 0 \Longleftrightarrow\ A^1 - B^3 - A^2 + B^4 \succeq 0 \ \ \textit{and}$$

$$\begin{pmatrix} y_{\mathbf{0}} - y_u & a_1^T - b_1^T & (t-1)(a_1^T - b_3^T) \\ & A^1 - B^1 & (t-1)(A^2 - B^2) \\ & & (t-1)(A^1 - B^3) + (t-1)(t-2)(A^2 - B^4) \end{pmatrix} \succeq 0.$$

(We wrote only the upper triangular part in the above (symmetric) matrix.)

Proof. Consider the orthogonal matrices

$$M := \begin{pmatrix} \mathbf{I} & 0 \\ 0 & U_{t-1} \end{pmatrix}, \qquad N := \begin{pmatrix} 1 & 0 \\ 0 & M \end{pmatrix},$$

where $\mathbf{I}$ is the identity matrix of order n and U_{t-1} is defined as follows. U_{t-1} is a $(t-1) \times (t-1)$ block matrix where, for $p, q = 1, \ldots, t-1$, its (p,q)th block U^{pq}_{t-1} is the $n \times n$ matrix defined as

$$U^{pq}_{t-1} := \begin{cases} \frac{1}{\sqrt{t-1}}\mathbf{I} & \text{if } p = 1 \text{ or } q = 1,\\ \left(\frac{1}{\sqrt{t-1}+t-1} - 1\right)\mathbf{I} & \text{if } p = q \geq 2,\\ \frac{1}{\sqrt{t-1}+t-1}\mathbf{I} & \text{otherwise.} \end{cases} \tag{2.26}$$

Notice that U_{t-1} is symmetric and orthogonal, i.e., $U_{t-1}(U_{t-1})^T = \mathbf{I}$. A simple calculation shows that

$$MDM = \begin{pmatrix} B^1 & \sqrt{t-1}B^2 & 0 & \dots & 0 \\ \sqrt{t-1}(B_2)^T & B^3 + (t-2)B^4 & 0 & \dots & 0 \\ 0 & 0 & B^3 - B^4 & & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & B^3 - B^4 \end{pmatrix}.$$

The first assertion of the lemma now follows after multiplying the second row/column block by $\sqrt{t-1}$. Next we have

$$N \begin{pmatrix} y_{\mathbf{0}} - y_u & c^T - d^T \\ c - d & C - D \end{pmatrix} N = \begin{pmatrix} y_{\mathbf{0}} - y_u & (c-d)^T M \\ M(c-d) & M(C-D)M \end{pmatrix}.$$

As the matrix $C - D$ has the same type of block shape as D, we deduce from the above that $M(C - D)M$ is block-diagonal. More precisely, the first diagonal block has the form

$$\begin{pmatrix} A^1 - B^1 & \sqrt{t-1}(A^2 - B^2) \\ \sqrt{t-1}(A^2 - B^2)^T & (A^1 - B^3) + (t-2)(A^2 - B^4) \end{pmatrix},$$

and the remaining $t - 2$ diagonal blocks are all equal to $A^1 - B^3 - A^2 + B^4$. One can moreover verify that $(c-d)^T M = \big(a_1^T - b_1^T, \sqrt{t-1}(a_1^T - b_3^T), 0 \dots 0\big)$. From this follows the second assertion of the lemma. □

In summary, we have obtained the following more compact semidefinite program for the parameter $\ell(G_t)$:

(2.27)

$$\begin{aligned} \ell(G_t) = \max\ & te^T a_1 \ \text{ s.t. } a_1 = \operatorname{diag}(A^1),\ b_1 = \operatorname{diag}(B^1),\ b_3 = \operatorname{diag}(B^3) \in \mathbb{R}^n, \\ & A^1, A^2, B^1, B^2, B^3, B^4 \in \mathbb{R}^{n \times n} \text{ satisfy (2.25) and} \\ & \begin{pmatrix} 1 - (a_1)_h & a_1^T - b_1^T & (t-1)(a_1^T - b_3^T) \\ & A^1 - B^1 & (t-1)(A^2 - B^2) \\ & & (t-1)(A^1 - B^3) + (t-1)(t-2)(A^2 - B^4) \end{pmatrix} \succeq 0, \\ & \begin{pmatrix} B^1 & (t-1)B^2 \\ & (t-1)B^3 + (t-1)(t-2)B^4 \end{pmatrix} \succeq 0, \\ & A^1 - A^2 - B^3 + B^4 \succeq 0, \\ & B^3 - B^4 \succeq 0. \end{aligned}$$

This formulation applies when G is vertex-transitive; here h is any fixed node of G. Hence $\Psi_\ell(G)$ can be obtained by computing $\ell(G_t)$ for $O(\log n)$ queries of the parameter t (see [17]) and the computation of each $\ell(G_t)$ is via an SDP involving four LMIs matrices of size $2n + 1$, $2n$, n, and n, respectively. The above reductions obviously apply to the stronger bound $\Psi_{\ell_{\geq 0}}$ obtained by adding nonnegativity, i.e., by adding the constraints $A^1, \dots, B^4 \geq 0$ in (2.27).

3. Bounds for Hamming graphs. We indicate here how to compute the parameters $\psi(G)$ and $\Psi_\ell(G)$ when G is a Hamming graph. Given an integer $n \geq 1$ and $\mathcal{D} \subseteq N := \{1, \dots, n\}$, G is the graph $H(n, \mathcal{D})$ with node set $V(G) := \mathcal{P}(N)$ and with an edge (I, J) if $|I \triangle J| \in \mathcal{D}$ (for $I, J \in \mathcal{P}(N)$). Thus we now have $|V(G)| = 2^n$.

As G is vertex-transitive, we can use the program (2.27). As the program (2.27) involves matrices of size $O(2^n)$, it cannot be solved directly for interesting values of n. However, one can use the fact that the Hamming graph $G = H(n, \mathcal{D})$ has a large automorphism group for reducing the size of the matrices $A^1, \dots, B^4$ involved in the program (2.27). Each permutation $\sigma \in \mathrm{Sym}(n)$ induces an automorphism of G by letting $\sigma(I) := \{\sigma(i) \mid i \in I\}$ for $I \in \mathcal{P}(N)$, and, for any $K \in \mathcal{P}(N)$, the *switching mapping* s_K defined by $s_K(I) := I \triangle K$ (for $I \in \mathcal{P}(N)$) is also an automorphism of G. Then $\mathrm{Aut}(G) = \{\sigma s_K \mid \sigma \in \mathrm{Sym}(n),\ K \in \mathcal{P}(N)\}$ and $|\mathrm{Aut}(G)| = n!2^n$.

It turns out that the matrices $A^1, \dots, B^4$ appearing in (2.27) belong to the Terwilliger algebra of the Hamming graph. By using the explicit block-diagonalization of the Terwilliger algebra, presented by Schrijver [31], we are able to block-diagonalize the matrices in (2.27) which enables the computation of $\Psi_\ell(G)$ for $G = H(n, \mathcal{D})$ for n up to 20. We recall the details needed for our treatment in the next subsection.

3.1. The Terwilliger algebra. For $i, j, p = 0, \dots, n$, let $M^{p,n}_{i,j}$ denote the 0/1 matrix indexed by $\mathcal{P}(N)$ whose (I, J)th entry is 1 if $|I| = i$, $|J| = j$, and $|I \cap J| = p$ and equal to 0 otherwise. The set

$$\mathcal{A}_n := \left\{ \sum_{i,j,p=0}^{n} x^p_{i,j} M^{p,n}_{i,j} \mid x^p_{i,j} \in \mathbb{R} \right\}$$

is an algebra, known as the *Terwilliger algebra* of the Hamming graph. For $k = 0, \dots, n$, let M^n_k be the matrix indexed by $\mathcal{P}(N)$ whose (I, J)th entry is 1 if $|I \triangle J| = k$ and 0 otherwise. The set

$$\mathcal{B}_n := \left\{ \sum_{k=0}^{n} x_k M^n_k \mid x_k \in \mathbb{R} \right\}$$

is an algebra, known as the *Bose–Mesner algebra* of the Hamming graph. Obviously, $\mathcal{B}_n \subseteq \mathcal{A}_n$, since $M^n_k = \sum_{i,j,p \mid i+j-2p=k} M^{p,n}_{i,j}$. As is well known, $\mathcal{B}_n$ is a commutative algebra, and thus all matrices in $\mathcal{B}_n$ can be simultaneously diagonalized (cf. Delsarte [7]). The Terwilliger algebra is not commutative, and thus it cannot be diagonalized; however, it can be block-diagonalized, as explained in [31]. We recall the main result below.

Given integers $i, j, k, p = 0, \dots, n$, set

$$\beta^{p,n}_{i,j,k} := \sum_{u=0}^{n} (-1)^{p-u} \binom{u}{p} \binom{n-2k}{n-k-u} \binom{n-k-u}{i-u} \binom{n-k-u}{j-u}, \tag{3.1}$$

$$\alpha^{p,n}_{i,j,k} := \beta^{p,n}_{i,j,k} \binom{n-2k}{i-k}^{-\frac{1}{2}} \binom{n-2k}{j-k}^{-\frac{1}{2}}. \tag{3.2}$$

THEOREM 3.1 (see [31]). *For a matrix $M = \sum_{i,j,p} M^{p,n}_{i,j} x^p_{i,j}$ in the Terwilliger algebra,*

$$M \succeq 0 \Longleftrightarrow M_k := \left(\sum_p \alpha^{p,n}_{i,j,k} x^p_{i,j} \right)_{i,j=k}^{n-k} \succeq 0 \quad \text{for } k = 0, 1, \dots, \left\lfloor \frac{n}{2} \right\rfloor. \tag{3.3}$$

To show this, Schrijver [31] constructs an orthogonal matrix U having the following property:

$$U^T M U = \begin{pmatrix} \widehat{M}_0 & 0 & \dots & 0 \\ 0 & \widehat{M}_1 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & \widehat{M}_{\lfloor n/2 \rfloor} \end{pmatrix}, \quad \text{where } \widehat{M}_k = \begin{pmatrix} M_k & 0 & \dots & 0 \\ 0 & M_k & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & M_k \end{pmatrix},$$

with block M_k being repeated $\binom{n}{k} - \binom{n}{k-1}$ times, for $k = 0, \dots, \lfloor n/2 \rfloor$.

The result extends to a block matrix whose blocks all lie in the Terwilliger algebra and which has a border of a special form. We state Lemma 3.2 for a 2×2 block matrix, but the analogous result holds obviously for any number of blocks.

LEMMA 3.2. *Let $A, B, C \in \mathcal{A}_n$, say, $A = \sum_{i,j,p} a^p_{i,j} M^{p,n}_{i,j}$, $B = \sum_{i,j,p} b^p_{i,j} M^{p,n}_{i,j}$, and $C = \sum_{i,j,p} c^p_{i,j} M^{p,n}_{i,j}$, and define accordingly*

$$A_k = \left(\sum_p \alpha^{p,n}_{i,j,k} a^p_{i,j} \right)^{n-k}_{i,j=k}, \; B_k = \left(\sum_p \alpha^{p,n}_{i,j,k} b^p_{i,j} \right)^{n-k}_{i,j=k}, \; C_k = \left(\sum_p \alpha^{p,n}_{i,j,k} c^p_{i,j} \right)^{n-k}_{i,j=k}.$$

Then

$$\begin{pmatrix} A & B \\ B^T & C \end{pmatrix} \succeq 0 \iff \begin{pmatrix} A_k & B_k \\ B_k^T & C_k \end{pmatrix} \succeq 0 \; \forall k = 0, 1, \dots, \left\lfloor \frac{n}{2} \right\rfloor.$$

Proof. The proof follows directly from the above by using the orthogonal matrix $\left(\begin{smallmatrix} U & 0 \\ 0 & U \end{smallmatrix}\right)$. ∎

LEMMA 3.3 (see Lemma 1 in [21]). *Let $M = \sum_{i,j,p=0}^n x^p_{i,j} M^{p,n}_{i,j} \in \mathcal{A}_n$, $c = \sum_{i=0}^n c_i \chi^i$, where $\chi^i \in \{0,1\}^{\mathcal{P}(N)}$ with $\chi^i_I = 1$ if $|I| = i$ (for $I \in \mathcal{P}(N)$), and $d \in \mathbb{R}$. Then*

$$\begin{pmatrix} d & c^T \\ c & M \end{pmatrix} \succeq 0 \iff \begin{cases} M_k \succeq 0 \;\; \text{for } k = 1, \dots, \left\lfloor \frac{n}{2} \right\rfloor, \\ \tilde{M}_0 := \begin{pmatrix} d & \tilde{c}^T \\ \tilde{c} & M_0 \end{pmatrix} \succeq 0 \end{cases}$$

after setting $\tilde{c}^T := (c_i \sqrt{\binom{n}{i}})_{i=0}^n$.

3.2. Compact formulation for $\psi(G)$ for Hamming graphs. As the graph $G = H(n, \mathcal{D})$ is vertex-transitive, we have $\psi(G) = \frac{2^n}{\ell(G)}$ by (2.16). It is shown in [21] how to compute the parameter $\ell(G)$ (when $\mathcal{D}$ is an interval $[1, d]$, but the reasoning is the same for any $\mathcal{D}$). The basic idea is that the matrix $M_2(h; x)$ appearing in (2.12) is a block matrix whose blocks lie in the Terwilliger algebra, and thus it can be block-diagonalized. We recall the details, directly for the parameter $\psi(G)$ from (2.13), as they will be useful for our treatment of the parameter $\Psi_\ell(G)$ in the next section.

Let x be feasible for the program (2.13). As G is vertex-transitive it suffices to require the condition $M_2(h; x) \succeq 0$ in (2.13) for *one* choice of $h \in V(G)$. Moreover, we may assume that the variable x is invariant under the action of the automorphism group of G. To fix ideas, let us choose the node $h := \emptyset$ of G (the empty subset of N). The matrix $M_2(\emptyset; x)$ has the block form

$$M_2(\emptyset; x) = \begin{pmatrix} t & e^T & b^T \\ e & A & B \\ b & B & B \end{pmatrix}, \tag{3.4}$$

where A, B, e, and b are indexed by $V(G) = \mathcal{P}(N)$, $\mathrm{diag}(A) = e$, and $\mathrm{diag}(B) = b$. By Lemma 2.1, we have

$$M_2(\emptyset; x) \succeq 0 \Longleftrightarrow \begin{pmatrix} t-1 & e^T - b^T \\ e-b & A-B \end{pmatrix} \succeq 0 \text{ and } B \succeq 0. \tag{3.5}$$

As x is invariant under the action of $\mathrm{Aut}(G)$, it follows that $A_{I,J} = x_{\{I,J\}} = x_{\{I',J'\}} = A_{I',J'}$ if $|I \triangle J| = |I' \triangle J'|$. In other words, the matrix A lies in the Bose–Mesner algebra, say,

$$A = \sum_{k=0}^{n} x_k M_k^n = \sum_{i,j,p=0}^{n} x_{i+j-2p} M_{i,j}^{p,n} \tag{3.6}$$

for some reals x_k. Moreover, $B_{I,J} = x_{\{\emptyset,I,J\}} = x_{\{\emptyset,I',J'\}} = B_{I',J'}$ if $|I'| = |I|$, $|J'| = |J|$, and $|I' \cap J'| = |I \cap J|$. In other words, the matrix B lies in the Terwilliger algebra, say,

$$B = \sum_{i,j,p=0}^{n} x_{i,j}^p M_{i,j}^{p,n} \tag{3.7}$$

for some reals $x_{i,j}^p$. The following relations link the parameters x_i and $x_{i,j}^p$.

LEMMA 3.4. *For $i, j, p = 0, \ldots, n$,*

$$\begin{aligned} &x_i = x_{0,i}^0, \\ &x_{i,j}^p = x_{j,i}^p = x_{i+j-2p,j}^{j-p} = x_{i+j-2p,i}^{i-p}, \end{aligned} \tag{3.8}$$

and the edge equations read

$$x_{i,j}^p = 0 \quad \textit{if} \quad \{i, j, i+j-2p\} \cap \mathcal{D} \neq \emptyset. \tag{3.9}$$

Proof. If $|I| = i$, then $x_i = A_{\emptyset,I} = x_{\{\emptyset,I\}} = B_{\emptyset,I} = x_{0,i}^0$. Let $|I| = i$, $|J| = j$, and $|I \cap J| = p$. Then $x_{i,j}^p = B_{I,J} = B_{J,I} = x_{j,i}^p$. Moreover, $x_{i,j}^p = B_{I,J} = x_{\{\emptyset,I,J\}} = x_{\{I,\emptyset,I\triangle J\}} = B_{I,I\triangle J} = x_{i+j-2p,i}^{i-p}$. This shows (3.8). The edge conditions read $B_{I,J} = x_{\{I,\emptyset,J\}} = 0$ if $\{|I|, |J|, |I \triangle J|\} \cap \mathcal{D} \neq \emptyset$, giving (3.9). □

We can now use the results from the previous subsection (Theorem 3.1 and Lemma 3.3) for block-diagonalizing the matrices occurring in (3.5). For $k = 0, \ldots, \lfloor n/2 \rfloor$, define the matrices

$$A_k := \left(\sum_p \alpha_{i,j,k}^{p,n} x_{0,i+j-2p}^0 \right)_{i,j=k}^{n-k}, \qquad B_k := \left(\sum_p \alpha_{i,j,k}^{p,n} x_{i,j}^p \right)_{i,j=k}^{n-k} \tag{3.10}$$

corresponding, respectively, to the matrices A, and B in (3.6) and (3.7). Define the vector

$$\tilde{c} := \left(\sqrt{\binom{n}{i}} \left(1 - x_{0,i}^0\right) \right)_{i=0}^{n} \in \mathbb{R}^{n+1}. \tag{3.11}$$

Then the parameter $\psi(H(n, \mathcal{D}))$ can be reformulated in the following way:

$$\begin{aligned} \psi(H(n,\mathcal{D})) = \min\ t \ \text{ s.t.} \quad & x_{0,0}^0 = 1 \text{ and } x_{i,j}^p \text{ satisfy (3.8) and (3.9), and} \\ & A_k - B_k \succeq 0 \ \text{ for } k = 1, \ldots, \lfloor n/2 \rfloor, \\ & B_k \succeq 0 \ \text{ for } k = 0, 1, \ldots, \lfloor n/2 \rfloor, \\ & \begin{pmatrix} t-1 & \tilde{c}^T \\ \tilde{c} & A_0 - B_0 \end{pmatrix} \succeq 0, \end{aligned} \tag{3.12}$$

where A_k, B_k, and $\tilde{c}$ are as in (3.10) and (3.11). To compute $\psi_{\geq 0}(H(n,\mathcal{D}))$, simply add the nonnegativity condition $x^p_{i,j} \geq 0$ to (3.12).

3.3. Compact formulation for $\Psi_\ell(G)$ for Hamming graphs. We now give a more compact formulation for the parameter $\Psi_\ell(G)$ when $G = H(n,\mathcal{D})$. As mentioned above, one has to evaluate $\ell(G_t)$ for various choices of $t \in \mathbb{N}$, with $\ell(G_t)$ being given by (2.27). As for the parameter $\psi(H(n,\mathcal{D}))$, we now observe that $A^1,\dots,B^4$, and thus all blocks in the matrices in (2.27) lie in the Terwilliger algebra. (As in the previous section we fix $h := \emptyset$, the empty subset of N.)

LEMMA 3.5. *The matrices A^s ($s = 1,2$) belong to the Bose–Mesner algebra $\mathcal{B}_n$, and the matrices B^s ($s = 1,2,3,4$) belong to the Terwilliger algebra $\mathcal{A}_n$, say, $A^s = \sum_{i=0}^n x(s)_i M^n_i$ ($s = 1,2$) and $B^s = \sum_{i,j,p=0}^n y(s)^p_{i,j} M^{p,n}_{i,j}$ ($s = 1,2,3,4$). Then*

$$
\begin{aligned}
&x(s)_i = y(s)^0_{0,i} \quad \text{for } s = 1,2,\ i = 1,\dots,n,\\
&y(s)^p_{i,j} = y(s)^p_{j,i} = y(s)^{j-p}_{i+j-2p,j} = y(s)^{i-p}_{i+j-2p,i} \ (\text{for } s = 1,4),\\
&y(2)^p_{i,j} = y(2)^{i-p}_{i,i+j-2p}, \qquad y(3)^p_{i,j} = y(3)^p_{j,i},\\
&y(3)^p_{i,j} = y(2)^{i-p}_{i+j-2p,i} \quad \text{for } i,j,p = 0,\dots,n.
\end{aligned} \tag{3.13}
$$

Moreover, the edge conditions can be reformulated as

$$
\begin{aligned}
&y(1)^p_{i,j} = 0 && \text{if } \{i,j,i+j-2p\} \cap \mathcal{D} \neq \emptyset,\\
&y(2)^i_{i,i} = y(4)^i_{i,i} = 0 && \text{for } i = 0,\dots,n,\\
&y(2)^p_{i,j} = 0 && \text{if } i \in \mathcal{D} \text{ or } j = 0,\\
&y(3)^p_{i,j} = 0 && \text{if } i+j-2p \in \mathcal{D}, \text{ or } i = 0, \text{ or } j = 0,\\
&y(4)^p_{i,j} = 0 && \text{if } i = 0 \text{ or } j = 0
\end{aligned} \tag{3.14}
$$

for distinct $i,j \in \{0,1,\dots,n\}$.

Proof. We use the fact that $A^1,\dots,B^4$ satisfy (2.24) and (2.25) where the variable y is assumed to be invariant under the action of $\mathrm{Sym}(t) \times \mathrm{Aut}(G) \subseteq \mathrm{Aut}(G_t)$. We have $A^1, A^2 \in \mathcal{B}_n$, since the entries $A^1_{I,J} = y_{\{1I,1J\}}$ and $A^2_{I,J} = y_{\{1I,2J\}}$ depend only on $|I \triangle J|$. (Indeed, if $|I' \triangle J'| = |I \triangle J|$, then there exists $\sigma \in \mathrm{Aut}(G)$ mapping $\{I,J\}$ to $\{I',J'\}$, and thus, by the invariance of y under action of σ, $y_{\{1I,1J\}} = y_{\{1I',1J'\}}$ and $y_{\{1I,2J\}} = y_{\{1I',2J'\}}$.) Similarly, for $s = 1,\dots,4$, $B^s \in \mathcal{A}_n$ since the entry $B^s_{I,J}$ depends only on $|I|, |J|$ and $|I \cap J|$. The proof for the identities $x(s)_i = y(s)^0_{0,i}$ ($s = 1,2$) and $y(1)^p_{i,j} = \cdots = y(1)^{i-p}_{i+j-2p,i}$ is identical to the proof of (3.8). Let $I,J \in \mathcal{P}(N)$, with $|I| = i$, $|J| = j$, and $|I \cap J| = p$. Then $y(4)^p_{i,j} = B^4_{I,J} = y_{\{1\emptyset,2I,3J\}} = y_{\{1\emptyset,3I,2J\}}$ (use the invariance of y under the permutation $(2,3) \in \mathrm{Sym}(t)$) and thus is equal to $B^4_{J,I} = y(4)^p_{j,i}$. Moreover, $y(4)^p_{i,j} = y_{\{1\emptyset,2I,3J\}} = y_{\{1I,2\emptyset,3I\triangle J\}} = y_{\{2I,1\emptyset,3I\triangle J\}}$ (first apply the switching mapping by I and then permute the indices $1,2$) and thus is equal to $B^4_{I,I\triangle J} = y(4)^{i-p}_{i,i+j-2p}$. Next we have $y(2)^p_{i,j} = B^2_{I,J} = y_{\{1I,1\emptyset,2J\}} = y_{\{1\emptyset,1I,2I\triangle J\}}$ (apply the switching mapping by I) and thus is equal to $B^2_{I,I\triangle J} = y(2)^{i-p}_{i,i+j-2p}$. Finally, $y(3)^p_{i,j} = B^3_{I,J} = y_{\{2I,1\emptyset,2J\}} = B^3_{J,I} = y(3)^p_{j,i}$, and $y(3)^p_{i,j} = y_{\{2I,1\emptyset,2J\}} = y_{\{2\emptyset,1I,2I\triangle J\}} = y_{\{1\emptyset,2I,1I\triangle J\}}$ (first switch by I and then permute $1,2$) and thus is equal to $B^2_{I\triangle J,I} = y(2)^{i-p}_{i+j-2p,i}$. The identities (3.14) follow directly from (2.25). ∎

As the blocks of the matrices in the program (2.27) lie in the Terwilliger algebra, the matrices in (2.27) can be block-diagonalized, as explained in section 3.1. For this, define the matrices

$$A_k^s := \left(\sum_p \alpha_{i,j,k}^{p,n} y(s)_{i+j-2p,0}^0\right)_{i,j=k}^{n-k}, \qquad B_k^s := \left(\sum_p \alpha_{i,j,k}^{p,n} y(s)_{i,j}^p\right)_{i,j=k}^{n-k} \tag{3.15}$$

corresponding, respectively, to the matrices A^s $(s = 1, 2)$ and B^s $(s = 1, 2, 3, 4)$, and define the vectors

$$\tilde{a} := \left(\sqrt{\binom{n}{i}}\,(y(1)_{0,0}^0 - y(1)_{i,i}^i)\right)_{i=0}^{n}, \ \tilde{b} := \left(\sqrt{\binom{n}{i}}\,(y(1)_{i,i}^i - y(3)_{i,i}^i)\right)_{i=0}^{n} \in \mathbb{R}^{n+1}. \tag{3.16}$$

By using Lemmas 3.2 and 3.3, we obtain the following reformulation for the parameter $\ell(G_t)$ from (2.27):

$$\begin{aligned}
\ell(G_t) = \ \max\ & 2^n t y(1)_{0,0}^0 \ \text{ s.t. } y(s)_{i,j}^p \ (s = 1, \dots, 4) \text{ satisfy (3.13) and (3.14), and} \\
& \begin{pmatrix} 1 - y(1)_{0,0}^0 & \tilde{a}^T & (t-1)\tilde{b}^T \\ & A_0^1 - B_0^1 & (t-1)(A_0^2 - B_0^2) \\ & & (t-1)(A_0^1 - B_0^3) + (t-1)(t-2)(A_0^2 - B_0^4) \end{pmatrix} \succeq 0, \\
& \begin{pmatrix} A_k^1 - B_k^1 & (t-1)(A_k^2 - B_k^2) \\ & (t-1)(A_k^1 - B_k^3) + (t-1)(t-2)(A_k^2 - B_k^4) \end{pmatrix} \succeq 0 \text{ for } k = 1, \dots, \lfloor n/2 \rfloor, \\
& \begin{pmatrix} B_k^1 & (t-1)B_k^2 \\ & (t-1)B_k^3 + (t-1)(t-2)B_k^4 \end{pmatrix} \succeq 0 \text{ for } k = 0, \dots, \lfloor n/2 \rfloor, \\
& A_k^1 - A_k^2 - B_k^3 + B_k^4 \succeq 0 \ \text{ for } k = 0, \dots, \lfloor n/2 \rfloor, \\
& B_k^3 - B_k^4 \succeq 0 \ \text{ for } k = 0, \dots, \lfloor n/2 \rfloor,
\end{aligned} \tag{3.17}$$

where A_k^s, B_k^s, $\tilde{a}$, and $\tilde{b}$ are as in (3.15) and (3.16). To compute $\ell_{\geq 0}(G_t)$ simply add the nonnegativity condition $y(s)_{i,j}^p \geq 0$ on all variables.

3.4. Numerical results for Hamming graphs. We have tested the various bounds on some instances of Hamming graphs. In what follows we use the following convention: For an integer $1 \leq d \leq n$, $H(n, d)$ (resp., $H^-(n, d)$, $H^+(n, d)$) denotes the graph $H(n, \mathcal{D})$, with $\mathcal{D} = \{d\}$ (resp., $\mathcal{D} = \{1, \dots, d\}$, $\{d, \dots, n\}$). The papers [9, 10, 11] give numerical results for the parameters $\overline{\vartheta}(G)$ and $\overline{\vartheta^+}(G)$ for such instances. Moreover, a bound related to copositive programming is computed in [11] (called the $\mathcal{K}_1$-bound in [11] or the $\kappa^{(1)}$ bound in [17]); it is shown in [17] that this bound is dominated by our parameter $\psi_{\geq 0}$.

In Table 1, the symbol "*" indicates the strict inequality $\Psi_\ell(G) > \lceil \psi(G) \rceil$, which happens for $H(10, 8)$ and $H^+(10, 8)$, and we indicate in bold the values satisfying LB $= \chi(G)$ for the obtained lower bound (LB). (Indeed, in these instances, LB $= 2^{n-1}$, while $\mathcal{P}(V)$ can be covered by the 2^{n-1} distinct pairs $\{I, V \setminus I\}$ $(I \subseteq V)$ which are stable sets as $n \notin \mathcal{D}$.)

The results in Table 1 indicate that the parameters $\psi(G)$ and $\psi_{\geq 0}(G)$ give in some instances a major improvement on Szegedy's bound $\overline{\vartheta^+}(G)$. On the other hand,

TABLE 1
Bounds for the chromatic number of Hamming graphs.

Graph	$\overline{\vartheta}(G)$	$\overline{\vartheta^+}(G)$	$\psi(G)$	$\Psi_\ell(G)$	$\psi_{\geq 0}(G)$	$\Psi_{\ell_{\geq 0}}(G)$
$H^-(7,4)$	36	42.6667	**64**	**64**	**64**	**64**
$H^-(8,5)$	72	85.3333	**128**	**128**	**128**	**128**
$H(10,6)$	6	8.7273	10.4366	11	10.8936	11
$H^-(10,6)$	207.36	320	**512**	**512**	**512**	**512**
$H(10,8)$	2.6667	3.2	3.9232	5*	3.9232	5*
$H^+(10,8)$	3.2	3.2	3.9232	5*	3.9232	5*
$H(11,4)$	16	21.5652	25.7351	26	25.7351	26
$H(11,6)$	12	12	12	12	15.2836	16
$H^-(11,7)$	414.72	640	**1024**	**1024**	**1024**	**1024**
$H^-(11,8)$	711.1111	819.2	**1024**	**1024**	**1024**	**1024**
$H(11,8)$	3.2	4.9383	5.7805	6	5.7805	6
$H(13,8)$	5.3333	9.4118	12.1429	13	13.6533	14
$H(15,6)$	27.7647	30.7368	46.4371	47	50.3036	51
$H(16,8)$	16	16	16	16	28.4444	29
$H(17,6)$	35	48.2222	86.3086	87	88.3204	89
$H(17,8)$	18	18	32	32	46.5122	47
$H(17,10)$	6.6666	12.6315	15.8750	16	25.8405	26
$H(18,10)$	10	16	18.3076	19	38.8844	-
$H(20,6)$	59.3735	59.3735	140.9586	141	140.9586	-
$H(20,8)$	41.7143	60.9524	107.1489	-	136.4115	-

in most cases, the parameter $\Psi_\ell(G)$ gives no improvement since $\Psi_\ell(G) = \lceil \psi(G) \rceil$. It could be that this feature is specific to Hamming graphs. As we will see in the next section, the bound $\Psi_\ell(G)$ does improve the bound $\lceil \psi(G) \rceil$ for Kneser graphs.

4. Bounds for Kneser graphs. We have seen that the parameter $\psi(G)$ is bounded by $\chi^*(G)$ and that, for vertex-transitive graphs, it coincides with the bound $|V(G)|/\ell(G)$. On the other hand, $\Psi_\ell(G)$ can sometimes be strictly greater than $\lceil \psi(G) \rceil$, e.g., for the Hamming graph $H(10,8)$ (recall Table 1). We present here some numerical results showing that $\Psi_\ell(G)$ can in fact be strictly greater than $\lceil \chi^*(G) \rceil$ for Kneser graphs.

Given integers $n \geq 2r$, the Kneser graph $K(n,r)$ is the graph whose vertices are the subsets of size r of a set N, with $|N| = n$, two vertices being adjacent if and only if they are disjoint. As shown in [24], $\alpha(K(n,r)) = \binom{n-1}{r-1}$, and thus $\chi^*(K(n,r)) = \frac{n}{r}$ in view of (2.2) as $K(n,r)$ is vertex-transitive. Lovász proved that $\chi(K(n,r)) = n-2r+2$ in his celebrated paper [23]. Thus the fractional chromatic number and the chromatic number of $K(n,r)$ can differ significantly, while the fractional chromatic number is close to the clique number $\omega(K(n,r)) = \lfloor \frac{n}{r} \rfloor$. Moreover, Lovász [24] proved that, for $G = K(n,r)$, $\alpha(G) = \vartheta(G)$. Hence, $\ell(G) = \alpha(G)$, implying that $\psi(G) = \frac{|V(G)|}{\ell(G)} = \chi^*(G) = n/r$. Therefore, $\Psi_\ell(G) \geq \lceil n/r \rceil$. We show in this section how to compute $\Psi_\ell(G)$.

The Kneser graph $K(n,r)$ coincides with the subgraph of the Hamming graph $H(n,\{2r\})$ induced by the subset $\mathcal{P}_{=r}(N) := \{I \in \mathcal{P}(N) \mid |I| = r\}$. It will be convenient to view the Kneser graph also in the following alternative way. Fix a set $T \subseteq N$, with $|T| = r$, and define

$$\mathcal{P}(N,T) := \{(I', I'') \in \mathcal{P}(T) \times \mathcal{P}(N \setminus T) \mid |I'| = |I''|\}.$$

The mapping

$$\begin{array}{ccc} \mathcal{P}_{=r}(N) & \longrightarrow & \mathcal{P}(N,T), \\ I & \mapsto & (T \setminus I, I \setminus T) \end{array} \tag{4.1}$$

is a bijection, and $|I \triangle J| = |(T \setminus I) \triangle (T \setminus J)| + |(I \setminus T) \triangle (J \setminus T)|$ holds for $I, J \in \mathcal{P}_{=r}(N)$. Hence $K(n,r)$ can also be viewed as the graph with node set $\mathcal{P}(N,T)$, with two nodes $(I', I''), (J', J'') \in \mathcal{P}(N,T)$ being adjacent if $|I' \triangle J'| + |I'' \triangle J''| = 2r$.

As we will see below, the matrices involved in the program (2.27) for the computation of $\Psi_\ell(K(n,r))$ lie in $\mathcal{B}_{r,r'}$ ($r' = n - r$), a subalgebra of a tensor product of two Terwilliger algebras, which has also been studied and block-diagonalized by Schrijver [31] (in connection with constant-weight codes). We follow the same steps as in section 3 for the computation of $\ell(G_t)$ for Hamming graphs, which we now carry out for Kneser graphs.

4.1. The subalgebra $\mathcal{B}_{r,r'}$. As above, $|N| = n$, and we fix a subset $T \subseteq N$, with $|T| = r$, and set $r' := n - r$. For $i, j, p = 0, 1, \ldots, r$ (resp., $i', j', q = 0, 1, \ldots, r'$), let $M^{p,r}_{i,j}$ (resp., $M^{q,r'}_{i',j'}$) be the matrices indexed by $\mathcal{P}(T)$ (resp., $\mathcal{P}(N \setminus T)$) defining the Terwilliger algebra $\mathcal{A}_r$ (resp., $\mathcal{A}_{r'}$) as in section 3.1. Let now $\mathcal{A}_{r,r'}$ be the algebra generated by the tensor products of matrices in $\mathcal{A}_r$ and $\mathcal{A}_{r'}$, that is,

$$\mathcal{A}_{r,r'} := \left\{ \sum_{i,j,p,i',j',q} x^{p,q}_{i,j,i',j'} M^{p,r}_{i,j} \otimes M^{q,r'}_{i',j'} \mid x^{p,q}_{i,j,i',j'} \in \mathbb{R} \right\}.$$

Matrices in $\mathcal{A}_{r,r'}$ are indexed by the set $\mathcal{P}(T) \times \mathcal{P}(N \setminus T)$. Consider the subalgebra

$$\mathcal{B}_{r,r'} := \left\{ \sum_{i,j,p,q} y^{p,q}_{i,j} M^{p,r}_{i,j} \otimes M^{q,r'}_{i,j} \mid y^{p,q}_{i,j} \in \mathbb{R} \right\}.$$

So $\mathcal{B}_{r,r'}$ consists of all matrices from $\mathcal{A}_{r,r'}$ satisfying $x^{p,q}_{i,j,i',j'} = 0$ if $i \neq i'$ or $j \neq j'$. Hence, for $M \in \mathcal{B}_{r,r'}$ and $(I, I'), (J, J') \in \mathcal{P}(T) \times \mathcal{P}(N \setminus T)$, $M_{(I,I'),(J,J')} = 0$ if $|I| \neq |I'|$ or if $|J| \neq |J'|$. Therefore any row/column of M indexed by $(I, I') \notin \mathcal{P}(N,T)$ is identically zero, and we may thus restrict matrices in $\mathcal{B}_{r,r'}$ to being indexed by the subset $\mathcal{P}(N,T)$ of $\mathcal{P}(T) \times \mathcal{P}(N \setminus T)$.

For $k \leq r$, let $M^{n,r}_k$ be the matrix indexed by $\mathcal{P}(N,T)$, whose $((I, I'), (J, J'))$th entry is equal to 1 if $|I \triangle J| + |I' \triangle J'| = 2k$ and to 0 otherwise. Thus $M^{n,r}_k$ corresponds to the principal submatrix of M^n_{2k} (in the Bose–Mesner algebra $\mathcal{B}_n$) indexed by the subset $\mathcal{P}_{=r}(N)$ and $M^{n,r}_k \in \mathcal{B}_{r,r'}$ as $M^{n,r}_k = \sum_{i,j,p,q \mid i+j-p-q=k} M^{p,r}_{i,j} \otimes M^{q,r'}_{i,j}$. Hence the set

$$\mathcal{B}^r_n := \left\{ \sum_{k=0}^{r} x_k M^{n,r}_k \mid x_k \in \mathbb{R} \right\}$$

is a subalgebra of $\mathcal{B}_{r,r'}$.

Schrijver [31] proved the following analogue of Theorem 3.1, giving the explicit block-diagonalization for matrices in $\mathcal{B}_{r,r'}$. For $k = 0, \ldots, \lfloor \frac{r}{2} \rfloor$, $l = 0, \ldots, \lfloor \frac{r'}{2} \rfloor$, set

$$W_{kl} := \{k, k+1, \ldots, r-k\} \cap \{l, l+1, \ldots, r'-l\}.$$

THEOREM 4.1 (see [31]). *For a matrix $M = \sum_{i,j,p,q} y^{p,q}_{i,j} M^{p,r}_{i,j} \otimes M^{q,r'}_{i,j}$ in $\mathcal{B}_{r,r'}$,*

$$M \succeq 0 \iff M_{k,l} := \left(\sum_{p,q} \alpha^{p,r}_{i,j,k} \alpha^{q,r'}_{i,j,l} y^{p,q}_{i,j} \right)_{i,j \in W_{kl}} \succeq 0 \tag{4.2}$$
$$\text{for each } k = 0, 1, \ldots, \lfloor \tfrac{r}{2} \rfloor \text{ and } l = 0, 1, \ldots, \left\lfloor \frac{r'}{2} \right\rfloor.$$

We have the following analogues of Lemmas 3.2 and 3.3.

LEMMA 4.2. *Let* $A = \sum_{i,j,p,q} a_{i,j}^{p,q} M_{i,j}^{p,r} \otimes M_{i,j}^{q,r'}$, $B = \sum_{i,j,p,q} b_{i,j}^{p,q} M_{i,j}^{p,r} \otimes M_{i,j}^{q,r'}$, *and* $C = \sum_{i,j,p,q} c_{i,j}^{p,q} M_{i,j}^{p,r} \otimes M_{i,j}^{q,r'}$ *be matrices in* $\mathcal{B}_{r,r'}$, *and define accordingly*

$$A_{kl} = \left(\sum_{p,q} \alpha_{i,j,k}^{p,r} \alpha_{i,j,l}^{q,r'} a_{i,j}^{p,q} \right)_{i,j \in W_{kl}}, \qquad B_{kl} = \left(\sum_{p,q} \alpha_{i,j,k}^{p,r} \alpha_{i,j,l}^{q,r'} b_{i,j}^{p,q} \right)_{i,j \in W_{kl}},$$
$$C_{kl} = \left(\sum_{p,q} \alpha_{i,j,k}^{p,r} \alpha_{i,j,l}^{q,r'} c_{i,j}^{p,q} \right)_{i,j \in W_{kl}}.$$

Then

$$\begin{pmatrix} A & B \\ B^T & C \end{pmatrix} \succeq 0 \Longleftrightarrow \begin{pmatrix} A_{kl} & B_{kl} \\ B_{kl}^T & C_{kl} \end{pmatrix} \succeq 0 \ \forall k = 0, 1, \ldots, \left\lfloor \frac{r}{2} \right\rfloor \text{ and } l = 0, 1, \ldots, \left\lfloor \frac{r'}{2} \right\rfloor.$$

LEMMA 4.3. *Let* $M = \sum_{i,j,p,q=0}^{n} x_{i,j}^{p,q} M_{i,j}^{p,r} \otimes M_{i,j}^{q,r'} \in \mathcal{B}_{r,r'}$ *and* $c = \sum_{i=0}^{n} c_i \chi^i$, *where* $\chi^i \in \{0,1\}^{\mathcal{P}(N,T)}$ *with* $\chi^i_{(I,I')} = 1$ *if* $|I| = i$ *(for* $(I, I') \in \mathcal{P}(N,T)$*) and* $d \in \mathbb{R}$. *Then*

$$\begin{pmatrix} d & c^T \\ c & M \end{pmatrix} \succeq 0 \Longleftrightarrow \begin{cases} M_{kl} \succeq 0 \ \text{ for } k = 0, \ldots, \left\lfloor \frac{r}{2} \right\rfloor, l = 0, \ \ldots, \left\lfloor \frac{r'}{2} \right\rfloor, \ k + l > 0; \\ \tilde{M}_{00} := \begin{pmatrix} d & \tilde{c}^T \\ \tilde{c} & M_{00} \end{pmatrix} \succeq 0 \end{cases}$$

after setting $\tilde{c}^T := (c_i \sqrt{\binom{r}{i}\binom{r'}{i}})_{i=0}^{r}$.

4.2. Compact formulation for $\Psi_\ell(G)$ for Kneser graphs. In order to compute $\Psi_\ell(G)$ for the Kneser graph $G = K(n,r)$, one has to evaluate $\ell(G_t)$ for various choices of t. As G is vertex-transitive, $\ell(G_t)$ can be computed by using the program (2.27). We now fix $h := T \in \mathcal{P}_{=r}(N)$ corresponding to $(\emptyset, \emptyset) \in \mathcal{P}(N,T)$ as a chosen node of G. We now show that the matrices $A^1, \ldots, B^4$ appearing in program (2.27) lie in the algebra $\mathcal{B}_{r,r'}$, and thus they can be block-diagonalized by using Theorem 4.1. The following lemma is the analogue of Lemma 3.5.

LEMMA 4.4. *The matrices* A^s $(s = 1,2)$ *belong to* $\mathcal{B}_n^r$, *and the matrices* B^s $(s = 1,2,3,4)$ *belong to* $\mathcal{B}_{r,r'}$, *say,* $A^s = \sum_{i=0}^{r} x(s)_i M_i^{n,r}$ $(s = 1,2)$ *and* $B^s = \sum_{i,j,p,q=0}^{r} y(s)_{i,j}^{p,q} M_{i,j}^{t,r} \otimes M_{i,j}^{q,r'}$ $(s = 1,2,3,4)$. *We have*

$$\begin{aligned} x(s)_i &= y(s)_{0,i}^{0,0} \ \text{ for } s = 1,2, \ i = 1, \ldots, r, \\ y(s)_{i,j}^{p,q} &= y(s)_{j,i}^{p,q} = y(s)_{i,i+j-p-q}^{i-q,i-p} = y_{j,i+j-p-q}^{j-q,j-p} \ \text{ for } s = 1,4, \\ y(2)_{i,j}^{p,q} &= y(2)_{i,i+j-p-q}^{i-q,i-p}, \qquad y(3)_{i,j}^{p,q} = y(3)_{j,i}^{p,q}, \\ y(3)_{i,j}^{p,q} &= y(2)_{i+j-p-q,i}^{i-q,i-p} \ \text{ for } i,j,p,q = 0, \ldots, r. \end{aligned} \tag{4.3}$$

Moreover, the edge conditions can be reformulated as

$$\begin{aligned} y(1)_{i,j}^{p,q} &= 0 \quad \text{if } \ i = r, \ \text{or } j = r, \ \text{or } i + j - p - q = r, \\ y(2)_{i,j}^{p,q} &= 0 \quad \text{if } i = r, \ \text{or } j = 0, \ \text{or } i + j - p - q = 0, \\ y(3)_{i,j}^{p,q} &= 0 \quad \text{if } i = 0, \ \text{or } j = 0, \ \text{or } i + j - p - q = r, \\ y(4)_{i,j}^{p,q} &= 0 \quad \text{if } i = 0, \ \text{or } j = 0, \ \text{or } i + j - p - q = 0. \end{aligned} \tag{4.4}$$

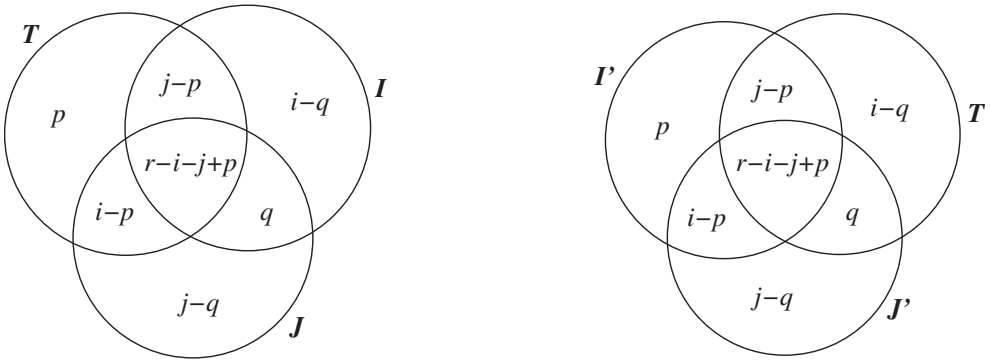


FIG. 4.1. *Venn diagrams.*

Proof. As in the proof of Lemma 3.5, the matrices $A^1, \dots, B^4$ satisfy (2.24) and (2.25), where the variable y is invariant under the action of $\mathrm{Sym}(t) \times \mathrm{Aut}(G)$. A main difference with the case of the Hamming graph is that, for the Kneser graph $G = K(n,r)$, $\mathrm{Aut}(G) \sim \mathrm{Sym}(n)$; i.e., the only automorphisms of G arise from the permutations of N. Recall that $\sigma \in \mathrm{Sym}(n)$ acts on $\mathcal{P}_{=r}(N)$ in the obvious way, by letting $\sigma(I) = \{\sigma(i) \mid i \in I\}$ for $I \in \mathcal{P}_{=r}(N)$.

Let us first show that $A^1 \in \mathcal{B}_n^r$; that is, $A^1_{I,J}$ depends only on $|I \triangle J|$ (for $I, J \in \mathcal{P}_{=r}(N)$). For this, let $I, J, I', J' \in \mathcal{P}_{=r}(N)$, with $|I \triangle J| = |I' \triangle J'|$. Then $|I \cap J| = |I' \cap J'|$, and thus there exists $\sigma \in \mathrm{Sym}(n)$ such that $\sigma(I) = I'$ and $\sigma(J) = J'$. Hence, $A^1_{I,J} = y_{\{1I,1J\}} = y_{\{1\sigma(I),1\sigma(J)\}} = A^1_{I',J'}$ since y is invariant under the action of σ. The proof for $A^2 \in \mathcal{B}_n^r$, $B^s \in \mathcal{B}_{r,r'}$, is along the same lines.

Let us now prove the identity $y(1)^{p,q}_{i,j} = y(1)^{i-q,i-p}_{i,i+j-p-q}$; the proofs for the remaining identities are along the same lines and thus are omitted. Say, $y(1)^{p,q}_{i,j} = B^1_{I,J}$, where $I, J \in \mathcal{P}_{=r}(N)$ with $|T \setminus I| = i$, $|T \setminus J| = j$, $|(T \setminus I) \cap (T \setminus J)| = p$, and $|(I \setminus T) \cap (J \setminus T)| = q$. See Figure 4.1 for the Venn diagram of the sets I, J, and T. Consider sets $I', J' \in \mathcal{P}_{=r}(N)$, which together with the set T have the Venn diagram shown in Figure 4.1. Then $B^1_{I',J'} = y(1)^{i-q,i-p}_{i,i+j-p-q}$, and there exists $\sigma \in \mathrm{Sym}(n)$ such that $\sigma(T) = I'$, $\sigma(I) = T$, and $\sigma(J) = J'$. Therefore, $y(1)^{p,q}_{i,j} = B^1_{I,J} = y_{\{1I,1J,1T\}} = y_{\{1\sigma(I),1\sigma(J),1\sigma(T)\}} = y_{\{1T,1J',1I'\}} = B^1_{I',J'} = y(1)^{i-q,i-p}_{i,i+j-p-q}$. ∎

For $k = 0, \dots, \lfloor r/2 \rfloor$, $l = 0, \dots, \lfloor r'/2 \rfloor$, define the matrices

(4.5)
$$A^s_{kl} = \left(\sum_{p,q} \alpha^{p,r}_{i,j,k} \alpha^{q,r'}_{i,j,l} y(s)^{0,0}_{0,i+j-p-q} \right)_{i,j \in W_{kl}}, \qquad B^s_{kl} = \left(\sum_{p,q} \alpha^{p,r}_{i,j,k} \alpha^{q,r'}_{i,j,l} y(s)^{p,q}_{i,j} \right)_{i,j \in W_{kl}}$$

corresponding, respectively, to the matrices A^s $(s = 1, 2)$ and B^s $(s = 1, 2, 3, 4)$, and define the vectors

(4.6)
$$\tilde{a} := \left(\sqrt{\tbinom{r}{i}\tbinom{r'}{i}} \left(y(1)^{0,0}_{0,0} - y(1)^{i,i}_{i,i} \right) \right)_{i=0}^{r}, \qquad \tilde{b} := \left(\sqrt{\tbinom{r}{i}\tbinom{r'}{i}} \left(y(1)^{i,i}_{i,i} - y(3)^{i,i}_{i,i} \right) \right)_{i=0}^{r}.$$

By using Lemmas 4.2 and 4.3, we obtain the following reformulation for the parameter

$\ell(G_t)$ from (2.27):

(4.7)

$$\ell(G_t) = \max \binom{n}{r} t y(1)_{0,0}^{0,0} \text{ s.t. } y(s)_{i,j}^{p,q},\ s = 1,\dots,4 \text{ satisfy (4.3) and (4.4), and}$$

$$\begin{pmatrix} 1 - y(1)_{0,0}^{0,0} & \tilde{a}^T & (t-1)\tilde{b}^T \\ & A_{00}^1 - B_{00}^1 & (t-1)(A_{00}^2 - B_{00}^2) \\ & & (t-1)(A_{00}^1 - B_{00}^3) + (t-1)(t-2)(A_{00}^2 - B_{00}^4) \end{pmatrix} \succeq 0;$$

$$\begin{pmatrix} A_{kl}^1 - B_{kl}^1 & (t-1)(A_{kl}^2 - B_{kl}^2) \\ & (t-1)(A_{kl}^1 - B_{kl}^3) + (t-1)(t-2)(A_{kl}^2 - B_{kl}^4) \end{pmatrix} \succeq 0$$
$$\text{for } k = 0,\dots,\lfloor r/2 \rfloor, l = 0,\dots,\lfloor r'/2 \rfloor,\ k + l > 0;$$

$$\begin{pmatrix} B_{kl}^1 & (t-1)B_{kl}^2 \\ & (t-1)B_{kl}^3 + (t-1)(t-2)B_{kl}^4 \end{pmatrix} \succeq 0 \text{ for } k = 0,\dots,\lfloor r/2 \rfloor, l = 0,\dots,\lfloor r'/2 \rfloor;$$

$$A_{kl}^1 - A_{kl}^2 - B_{kl}^3 + B_{kl}^4 \succeq 0 \text{ for } k = 0,\dots,\lfloor r/2 \rfloor, l = 0,\dots,\lfloor r'/2 \rfloor;$$

$$B_{kl}^3 - B_{kl}^4 \succeq 0 \text{ for } k = 0,\dots,\lfloor r/2 \rfloor, l = 0,\dots,\lfloor r'/2 \rfloor,$$

where A_{kl}^s, B_{kl}^s, $\tilde{a}$, and $\tilde{b}$ are as in (4.5) and (4.6). To compute $\ell_{\geq 0}(G_t)$ simply add the nonnegativity condition $y(s)_{i,j}^{p,q} \geq 0$ on all variables.

4.3. Numerical results for Kneser graphs. We show in Table 2 below our numerical results for the bounds $\Psi_\ell(G)$ and $\Psi_{\ell_{\geq 0}}(G)$ for several instances of Kneser graphs. We indicate in bold the values achieving the chromatic number.

5. Computing the new bound ψ_K for DIMACS benchmark graphs. So far we have been dealing with vertex-transitive graphs and with the bounds $\psi(\cdot)$ and $\Psi_\ell(\cdot)$. For the formulation of $\psi(G)$, it was observed in section 2 that, when G is vertex-transitive, it suffices to require in (2.13) positive semidefiniteness of $M_2(h,x)$ for *only one* $h \in V(G)$ instead of *for all* $h \in V(G)$. In the case of a nonsymmetric graph G one would need to require $M_2(h,x) \succeq 0$ for *all* $h \in V(G)$; therefore, with

TABLE 2
Bounds for the chromatic number of Kneser graphs.

Graph	$\lceil \chi^*(G) \rceil = \lceil n/r \rceil$	$\Psi_\ell(G)$	$\Psi_{\ell_{\geq 0}}(G)$	$\chi(G) = n - 2r + 2$
$K(6,2)$	3	**4**	**4**	**4**
$K(7,2)$	4	4	**5**	**5**
$K(8,3)$	3	**4**	**4**	**4**
$K(9,3)$	3	4	4	5
$K(10,4)$	3	3	**4**	**4**
$K(11,3)$	4	5	5	7
$K(11,4)$	3	4	4	5
$K(12,3)$	4	5	6	8
$K(12,4)$	3	4	4	6
$K(12,5)$	3	3	**4**	**4**
$K(13,5)$	3	4	4	5
$K(14,5)$	3	4	4	6
$K(15,3)$	5	6	6	11
$K(16,4)$	4	5	6	10
$K(24,6)$	4	4	6	14
$K(25,5)$	5	6	7	17
$K(34,7)$	5	6	7	22
$K(36,6)$	6	7	9	26

$n := |V(G)|$, in order to compute $\psi(G)$ (resp., $\ell(G_t)$, and thus $\Psi_\ell(G)$), one would have to solve a semidefinite program with $2n$ (resp., $4n$) matrices of order $\leq n+1$ (resp., $\leq 2n+1$). For graphs that are of interest, e.g., with $n \geq 100$, this cannot be done with the currently available software for semidefinite programming.

For nonsymmetric graphs we propose another variant of the bound $\psi^{(2)}(G)$. Given a clique K in G, let $M_2(K;x)$ denote the principal submatrix of $M_2(x)$ indexed by the multiset $\mathcal{P}_1(V) \cup (\cup_{h\in K}\{\{h,i\} \mid i \in V\})$. Now define the parameter

$$\psi_K(G) := \min\ t \ \text{ s.t. } \quad \begin{array}{l} x_{\mathbf{0}} = t,\ x_i = 1\ (i \in V),\ M_2(K;x) \succeq 0, \\ x_I = 0 \text{ for all } I \text{ containing an edge.} \end{array} \tag{5.1}$$

Then $\overline{\vartheta}(G) \leq \psi_K(G) \leq \chi^*(G)$. (The left inequality follows by using (2.4), and the right inequality follows from $\psi_K(G) \leq \psi^{(2)}(G) \leq \chi^*(G)$ by using (2.8) and (2.10).) Set $k := |K|$, and assume without loss of generality that $K = \{1, 2, \ldots, k\}$. With respect to the partition of its index set as $\{\mathbf{0}\} \cup \{\{i\} \mid i \in V\} \cup \cup_{h=1}^k \{\{h,i\} \mid i \in V\}$, the matrix $M_2(K;x)$ has the block form

$$M_2(K;x) = \begin{pmatrix} t & a_0^T & a_1^T & a_2^T & \ldots & a_k^T \\ a_0 & A_0 & A_1 & A_2 & \ldots & A_k \\ a_1 & A_1 & A_1 & 0 & \ldots & 0 \\ a_2 & A_2 & 0 & A_2 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ a_k & A_k & 0 & \ldots & 0 & A_k \end{pmatrix},$$

where $a_0, \ldots, a_k, A_0, \ldots, A_k$ are indexed by V, $a_i = \mathrm{diag}(A_i)$ $(0 \leq i \leq k)$, $a_0 = e$, $(A_0)_{ij} = x_{ij}$, and $(A_h)_{ij} = x_{\{h,i,j\}}$ for $h \in K$, $i, j \in V$. Note that for $h \in V$ the columns of A_0 and A_h indexed by $\{h\}$ are both equal to a_h. Hence, as in the proof of Lemma 2.1, we can do some row/column manipulations and verify that

$$M_2(K;x) \succeq 0 \Longleftrightarrow \begin{pmatrix} t-k & e^T - (\sum_{h=1}^k a_h)^T \\ e - \sum_{h=1}^k a_h & A_0 - \sum_{h=1}^k A_h \end{pmatrix} \succeq 0, \quad A_1, \ldots, A_k \succeq 0.$$

Hence $\psi_K(G)$ can be computed via a semidefinite program involving $k+1$ matrices of sizes $n+1$ (once) and n (k times).

We have conducted experiments for some DIMACS benchmark graphs (studied, e.g., in [4, 5, 8, 9, 12, 26, 27]). In Table 3 we present our lower bounds for the chromatic number of the graphs DSJC$a.b$. Recall that DSJC$a.b$ are random graphs with a vertices, two vertices being adjacent with probability $10^{-1}b$. The graph DSJR500.1 is a geometric graph with 500 nodes randomly distributed in the unit square, with an edge between two nodes if their distance is less than 0.1. The graph DSJR500.1c is the complement of DSJR500.1. The graphs can be downloaded from [34].

In Table 3, the column "LB" contains the previously best known lower bounds taken from [8, 26, 27], and the values in parentheses come from [3]; the bound 82 for DSJR500.1c is the size of a clique obtained by using the heuristic of [2]. The column "UB" contains the best known upper bounds taken from [4, 12, 13], i.e., the number of colors in the best colorings found so far. The column "K" contains the size of the clique used for computing the parameter $\psi_K(G)$ (the clique is found by using the heuristic from [2]). We also indicate the value of the theta number $\bar{\vartheta}(G)$ (also computed in [9, 10] for some instances), which already improves the best lower bound

TABLE 3
Bounds for the chromatic number of DIMACS instances.

Graph	LB	$\overline{\vartheta}(G)$	$\lceil\overline{\vartheta}(G)\rceil$	K	$\psi_K(G)$	$\lceil\psi_K(G)\rceil$	UB
DSJC125.1	5	4.1062	5	4	4.337	**5**	5
DSJC125.5	14 (17)	11.7844	12	10	13.942	**14**	17
DSJC125.9	42	37.768	38	34	42.53	**43**	43
DSJC250.1	6 (8)	4.906	5	4	5.208	**6**	8
DSJC250.5	14	16.234	17	12	19.208	**20**	28
DSJC250.9	48	55.152	56	43	66.15	**67**	72
DSJC500.1	6	6.217	**7**	5	6.542	7	12
DSJC500.5	13 (16)	20.542	21	13	27.791	**28**	48
DSJC500.9	59	84.04	85	56	100.43	**101**	126
DSJC1000.1	6	8.307	**9**	5	-	-	20
DSJC1000.5	15 (17)	31.89	**32**	14	-	-	83
DSJC1000.9	66	122.67	**123**	65	-	-	224
DSJR500.1c	82 (83)	83.74	84	77	84.12	**85**	85

in several instances. We indicate in bold our best new lower bounds for the chromatic number. In several instances they give a significant improvement on the best known lower bound. Moreover, in two instances, we are able to close the gap as our lower bound matches the upper bound; indeed we find the exact value of the chromatic number for the graphs DSJC125.9 ($\chi(G) = 43$) and DSJR500.1c ($\chi(G) = 85$), which were not known before to the best of our knowledge. These results demonstrate that the bounds $\psi_K(G)$ can be quite strong.

One may wonder why we did not add nonnegativity constraints in the formulation for ψ_K. The reason is that for random graphs adding nonnegativity constraints gives only a negligible improvement. This fact was already observed for the Lovász theta number in [9].

Remarks about the computational results. The computational results reported in Tables 1 and 2 were carried out by using the open source codes for semidefinite programming CSDP 5.0 and DSDP 5.8 available, respectively, at [35] and [36].

For finding the large cliques reported in column "K" of Table 3, we used the heuristic Max-AO (based on [2]), available at [37]. The values in the columns "$\overline{\vartheta}(G)$" and "$\psi_K(G)$" of Table 3 were computed by using the boundary point method of Povh, Rendl, and Wiegele [29], whose code is available at [38].

The semidefinite program for the parameter ψ_K can indeed be quite large. For instance, for the graph DSJR500.1*c*, it contains one 501×501 block and 77 blocks of size at most 500×500, and such a big problem cannot be solved by using solvers based on interior point methods.

Experiments were conducted on a single machine with an AMD Athlon 64 3500 processor and 1024 MB RAM memory. Here is a rough indication of the times needed to compute the bounds in Tables 1–3. Each bound in Tables 1–2 could be computed in less than a minute, as it involves a relatively small SDP; for instance, computing $\Psi_\ell(H(20,6))$ is via an SDP with 1502 variables and 47 blocks with sizes ranging from 1 to 43. It was harder to compute the bounds ψ_K in Table 3. In fact, we had to rerun the boundary point code several times for each instance in order to tailor the parameters of the code and speed up the convergence to an optimal solution. The computation times for the parameter $\psi_K(G)$ vary from a few minutes (e.g., less than 3 minutes for DCJC125.5 and about 25 minutes for DCJC125.1) up to four days for the most demanding instance DSJR500.1c.

Acknowledgments. We are very grateful to two referees for their careful reading and their useful suggestions which helped improve the presentation of the paper. We also thank Marco Chiarandini and Michael Trick, for telling us about coloring results for DIMACS benchmark graphs, and Janez Povh, Franz Rendl, and Angelika Wiegele, for adapting their boundary point algorithm code in such a way that it now exploits the block-diagonal structure in semidefinite programs.

REFERENCES

[1] M. Bellare and M. Sudan, *Improved non-approximability results*, in Proceedings of the 26th Annual ACM Symposium on Theory of Computing, 1994, pp. 184–193.
[2] S. Burer, R. Monteiro, and Y. Zhang, *Maximum stable set formulations and heuristics based on continuous optimization*, Math. Program. Ser. A, 94 (2002), pp. 137–166.
[3] M. Caramia and P. Dell'Olmo, *Bounding vertex coloring by truncated multistage branch and bound*, Networks, 44 (2004), pp. 231–242.
[4] M. Caramia and P. Dell'Olmo, *Coloring graphs by iterated local search traversing feasible and infeasible solutions*, Discrete Appl. Math., 156 (2008), pp. 201–217.
[5] M. Chiarandini, *Stochastic Local Search Methods for Highly Constrained Combinatorial Optimisation Problems*, Ph.D. thesis, Darmstadt University of Technology, 2005.
[6] V. Chvátal, *Edmonds polytopes and a hierarchy of combinatorial problems*, Discrete Math., 4 (1973, pp. 305–337.
[7] P. Delsarte, *An Algebraic Approach to the Association Schemes of Coding Theory*, Philips Research Reports Supplements (1973) 10, Philips Research Laboratories, Eindhoven, 1973.
[8] C. Desrosiers, P. Galinier, and A. Hertz, *Efficient algorithms for finding critical subgraphs*, Discrete Appl. Math., 156 (2008), pp. 244–266.
[9] I. Dukanovic and F. Rendl, *Semidefinite programming relaxations for graph coloring and maximal clique problems*, Math. Program., 109 (2007), pp. 345–365.
[10] I. Dukanovic and F. Rendl, *A semidefinite programming based heuristic for graph coloring*, Discrete Appl. Math., 156 (2008), pp. 180–189.
[11] I. Dukanovic and F. Rendl, *Copositive Programming Motivated Bounds on the Clique and the Chromatic Number*, http://www.optimization-online.org/DB_HTML/2006/05/1403.html (2006).
[12] P. Galinier and J.-K. Hao, *Hybrid evolutionary algorithms for graph coloring*, J. Comb. Optim., 3 (1999), pp. 379–397.
[13] P. Galinier, A. Hertz, and N. Zufferey, *An adaptive memory algorithm for the k-colouring problem*, Discrete Appl. Math., 156 (2008), pp. 267–279.
[14] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, Freeman, San Francisco, 1979.
[15] M. Grötschel, L. Lovász, and A. Scrijver *Geometric Alghorithms and Combinatorial Optimization*, Springer-Verlag, Berlin, 1988.
[16] N. Gvozdenović, *Approximating the Stability Number and the Chromatic Number of a Graph via Semidefinite Programming*, Ph.D. thesis, University of Amsterdam, 2008.
[17] N. Gvozdenović and M. Laurent, *The operator Ψ for the chromatic number of a graph*, SIAM J. Optim., 19 (2008), pp. 572–591
[18] D. Karger, R. Motwani, and M. Sudan, *Approximate graph coloring by semidefinite programming*, J. ACM, 45 (1998), pp. 246–265.
[19] J. B. Lasserre, *An explicit exact SDP relaxation for nonlinear $0-1$ programs*, in Integer Programming and Combinatorial Optimization, Lecture Notes in Comput. Sci. 2081, K. Aardal and A. M. H. Gerards, eds., Springer-Verlag, Berlin, 2001, pp. 293–303.
[20] M. Laurent, *A comparison of the Sherali-Adams, Lovász-Schrijver and Lasserre relaxations for $0-1$ programming*, Math. Oper. Res., 28 (2003), pp. 470–496.
[21] M. Laurent, *Strengthened semidefinite programming bounds for codes*, Math. Program., 109 (2007), pp. 239–261.
[22] M. Laurent and F. Rendl, *Semidefinite programming and integer programming*, in Handbook on Discrete Optimization, K. Aardal, G. Nemhauser, and R. Weismantel, eds., Elsevier B.V., New York, 2005, pp 393–514.
[23] L. Lovász, *Kneser's conjecture, chromatic numbers and homotopy*, J. Combin. Theory Ser. A, 25 (1978), pp. 319–324.
[24] L. Lovász, *On the Shannon capacity of a graph*. IEEE Trans. Inform. Theory, 25 (1979, pp. 1–7.

[25] J. Matousek and G. Ziegler, *Topological lower bounds for the chromatic number: A hierarchy*, Jahresber. Deutsch. Math.-Verein., 106 (2004), pp. 71–90.
[26] I. M. Méndez-Diaz and P. Zabala, *A branch-and-cut algorithm for graph coloring*, Discrete Appl. Math., 154 (2006), pp. 826–847.
[27] I. Méndez-Diaz and P. Zabala, *A cutting plane algorithm for graph coloring*, Discrete Appl. Math., 156 (2008), pp. 159–179.
[28] P. Meurdesoif, *Strengthening the Lovász $\theta\left(\overline{G}\right)$ bound for graph coloring*, Math. Program., 102 (2005), pp. 577–588.
[29] J. Povh, F. Rendl, and A. Wiegele, *A boundary point method to solve semidefinite programs*, Computing, 78 (2006), pp. 277–286
[30] A. Schrijver, *Combinatorial Optimization - Polyhedra and Efficiency*, Springer-Verlag, Berlin, 2003.
[31] A. Schrijver, *New code upper bounds from the Terwilliger algebra and semidefinite programming*, IEEE Trans. Inform. Theory, 51 (2005), pp. 2859–2866.
[32] M. Szegedy, *A note on the theta number of Lovász and the generalized Delsarte bound*, in Proceedings of the 35th Annual IEEE Symposium on Foundations of Computer Science, 1994, pp. 36–39.
[33] G. Ziegler, *Coloring Hamming graphs, Optimal Binary Codes, and the 0/1-Borsuk Problem in Low Dimensions*, in Computational Discrete Mathematics, Lecture Notes in Comput. Sci. 2122, Springer-Verlag, Berlin, 2001, pp. 159–171.
[34] http://mat.gsia.cmu.edu/COLOR03/.
[35] http://infohost.nmt.edu/~borchers/csdp.html.
[36] http://www-unix.mcs.anl.gov/~benson/dsdp/.
[37] http://dollar.biz.uiowa.edu/~burer/software/Max-AO/index.html.
[38] http://www.math.uni-klu.ac.at/or/Software/theta_bp.m.

SIAM J. OPTIM.
Vol. 19, No. 2, pp. 616–643

SUFFICIENT SECOND-ORDER OPTIMALITY CONDITIONS FOR SEMILINEAR CONTROL PROBLEMS WITH POINTWISE STATE CONSTRAINTS*

EDUARDO CASAS†, JUAN CARLOS DE LOS REYES‡, AND FREDI TRÖLTZSCH§

Abstract. Second-order sufficient optimality conditions are established for the optimal control of semilinear elliptic and parabolic equations with pointwise constraints on the control and the state. In contrast to former publications on this subject, the cone of critical directions is the smallest possible in the sense that the second-order sufficient conditions are the closest to the associated necessary ones. The theory is developed for elliptic distributed controls in domains up to dimension three. Moreover, problems of elliptic boundary control and parabolic distributed control are discussed in spatial domains of dimension two and one, respectively.

Key words. optimal control, elliptic equations, parabolic equations, pointwise state constraints, second-order necessary optimality conditions, second-order sufficient optimality conditions

AMS subject classifications. 49K20, 90C48

DOI. 10.1137/07068240X

1. Introduction.

In this paper, we essentially improve the theory of second-order sufficient optimality conditions for state-constrained optimal control problems of elliptic and parabolic type. We derive second-order sufficient conditions that are as close as possible to the associated necessary ones. In this way, we are able to complete the theory of second-order sufficient conditions for this class of problems, if the dimension of the spatial domain is sufficiently small.

For the theory of nonconvex differentiable mathematical programming in finite-dimensional spaces, second-order sufficient optimality conditions are indispensible both in the numerical analysis and for reliable numerical methods. If second-order information is not available, then local minima will not in general be stable and numerical methods will most likely not converge. For instance, the convergence analysis of SQP methods relies heavily on second-order conditions.

In the numerical analysis of nonlinear optimal control problems, second-order sufficient optimality conditions are even more important. If they are not satisfied, then the (strong) convergence of optimal controls or states and/or error estimates for numerical discretizations of the problems can hardly be shown. Also, other types of perturbations are difficult to handle without second-order conditions.

As is well known from the calculus of variations and the control theory for nonlinear ordinary differential equations, the theory of second-order conditions is more delicate and rich in function spaces. We mention, for instance, the work by Maurer [20] or Maurer and Zowe [21]. In particular, the well-known two-norm discrepancy occurs that essentially complicates the analysis; cf. the expositions in Ioffe [16] or

*Received by the editors February 8, 2007; accepted for publication (in revised form) January 4, 2008; published electronically July 2, 2008.
http://www.siam.org/journals/siopt/19-2/68240.html
†Dpto. de Matemática Aplicada y Ciencias de la Computación, E.T.S.I. Industriales y de Telecomunicación, Universidad de Cantabria, E-39005 Santander, Spain (eduardo.casas@unican.es). This author was partially supported by Ministerio de Educación y Ciencia and "Ingenio Mathematica (i-MATH)" (Consolider Ingenio 2010 Spain).
‡Dpto. de Matemática, EPN Quito, Quito, Ecuador (jcdelosreyes@math.epn.edu.ec).
§Institut für Mathematik, Technische Universität Berlin, D-10623 Berlin, Germany (troeltzsch@math.tu-berlin.de).

Malanowski [18]. For the important but more difficult case of pointwise state constraints in the control of ordinary differential equations, we refer to Malanowski [19] and to the references therein.

At present, the control of distributed parameter systems with pointwise state constraints is a very active field of research. Although the majority of papers are still devoted to convex problems with linear equations, the important case of nonlinear state equations is attracting more interest. Here, second-order conditions are needed. However, when pointwise state constraints are imposed, the situation is more complicated, since the Lagrange multipliers associated with them are measures. In contrast to the theory for ordinary differential equations, this causes severe restrictions on the dimension of the spatial domains of the equations and reduces the regularity of the adjoint state.

To our best knowledge, there exist only two contributions to the theory of second-order sufficient conditions for distributed problems with pointwise state constraints. The elliptic case was discussed in [12], while parabolic problems were investigated in [22]. The method of these papers was inspired by the splitting technique used in [11]. When applied to pointwise state constraints, the cones of critical directions established by this technique are too large so that the second-order sufficient conditions are based on slightly too strong assumptions. Moreover, the method was fairly complicated.

For other contributions to second-order optimality conditions for distributed parameter systems, we mention, for instance, the work by Bonnans [3] and the exposition in the monography by Bonnans and Shapiro [4] on elliptic problems with control constraints. We also refer to [9], where second-order necessary optimality conditions were first treated for elliptic problems, [10] for an abstract framework with applications to elliptic and parabolic problems, and [7], where elliptic problems with control constraints and state constraints of integral type were considered Moreover, we refer to the references therein.

In this paper, the sufficiency of second-order conditions is proven by a method that is close to the theory of nonlinear optimization in finite-dimensional spaces. We establish a cone of critical directions that is sharp; i.e., it is the one closest to the cone for establishing second-order necessary conditions.

We present a detailed proof for the case of distributed elliptic problems in domains of spatial dimension $n \leq 3$. Moreover, we briefly sketch the extension of this result to elliptic boundary control problems for $n \leq 2$ and to the parabolic distributed case for $n = 1$.

2. Problem statement. Let Ω be an open and bounded domain in $\mathbb{R}^n$, $n \leq 3$, with a Lipschitz boundary Γ. In this domain we consider the following state equation:

$$\begin{cases} Ay + f(x,y) = u & \text{in } \Omega, \\ y = 0 & \text{on } \Gamma, \end{cases} \tag{2.1}$$

where $f : \Omega \times \mathbb{R} \longrightarrow \mathbb{R}$ is a Carathéodory function and A denotes a second-order elliptic operator of the form

$$Ay(x) = -\sum_{i,j=1}^{n} \partial_{x_j}(a_{ij}(x)\partial_{x_i} y(x));$$

the coefficients $a_{ij} \in L^\infty(\Omega)$ satisfy

$$\lambda_A \|\xi\|^2 \leq \sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \quad \forall \xi \in \mathbb{R}^n \text{ for a.e. } x \in \Omega$$

for some $\lambda_A > 0$. In (2.1), the function u denotes the control, and y_u is the solution associated to the control u. We will state later the conditions leading to the existence and uniqueness of a solution of (2.1) in $C(\bar\Omega) \cap H^1(\Omega)$.

In this paper, we study the following optimal control problem:

$$(\mathrm{P}) \begin{cases} \min J(u) = \displaystyle\int_\Omega L(x, y_u(x), u(x))\, dx \\ \text{subject to } \ (y_u, u) \in (C(\bar\Omega) \cap H^1(\Omega)) \times L^\infty(\Omega), \\ \alpha(x) \le u(x) \le \beta(x) \quad \text{for a.e. } \ x \in \Omega, \\ g(x, y_u(x)) \le 0 \ \ \forall x \in K, \end{cases}$$

where $\alpha(x) < \beta(x)$ for almost all $x \in \Omega$, $\alpha, \beta \in L^\infty(\Omega)$, and $K \subset \bar\Omega$ is a compact set. Let us state the assumptions on the functions L, f, and g.

(A1) f is of class C^2 with respect to the second variable:

$$f(\cdot, 0) \in L^2(\Omega), \qquad \frac{\partial f}{\partial y}(x, y) \ge 0 \quad \text{for a.e. } x \in \Omega,$$

and for all $M > 0$ there exists a constant $C_{f,M} > 0$ such that

$$\left|\frac{\partial f}{\partial y}(x, y)\right| + \left|\frac{\partial^2 f}{\partial y^2}(x, y)\right| \le C_{f,M} \text{ for a.e. } x \in \Omega \text{ and } |y| \le M,$$

$$\left|\frac{\partial^2 f}{\partial y^2}(x, y_2) - \frac{\partial^2 f}{\partial y^2}(x, y_1)\right| \le C_{f,M}|y_2 - y_1| \ \text{ for } |y_1|, |y_2| \le M \text{ and for a.e. } x \in \Omega.$$

(A2) $L : \Omega \times (\mathbb{R} \times \mathbb{R}) \longrightarrow \mathbb{R}$ is a Carathéodory function of class C^2 with respect to the second and third variables, $L(\cdot, 0, 0) \in L^1(\Omega)$, and for all $M > 0$ there is a constant $C_{L,M} > 0$ and a function $\psi_M \in L^2(\Omega)$ such that

$$\left|\frac{\partial L}{\partial u}(x, y, u)\right| + \left|\frac{\partial L}{\partial y}(x, y, u)\right| \le \psi_M(x), \quad \|D^2_{(y,u)} L(x, y, u)\| \le C_{L,M},$$

$$\|D^2_{(y,u)} L(x, y_2, u_2) - D^2_{(y,u)} L(x, y_1, u_1)\| \le C_{L,M}(|y_2 - y_1| + |u_2 - u_1|)$$

for a.e. $x \in \Omega$ and $|y|, |y_i|, |u|, |u_i| \le M$, $i = 1, 2$, where $D^2_{(y,u)} L$ denotes the second derivative of L with respect to (y, u).

(A3) The function $g : K \times \mathbb{R} \longrightarrow \mathbb{R}$ is continuous, of class C^2 with respect to the second variable and $\partial_y g$, and $\partial^2_y g$ are also continuous functions in $K \times \mathbb{R}$. Moreover we will assume that $g(x, 0) < 0$ is satisfied for every $x \in K \cap \Gamma$.

The following result on the existence of a solution holds true for (2.1) as well as for the problem (P).

THEOREM 2.1. *Suppose that* (A1) *holds. Then, for every $u \in L^2(\Omega)$, the state equation* (2.1) *has a unique solution $y_u \in C(\bar\Omega) \cap H^1_0(\Omega)$. Furthermore, if $u_k \rightharpoonup u$ weakly in $L^2(\Omega)$, then $y_{u_k} \to y_u$ strongly in $C(\bar\Omega) \cap H^1_0(\Omega)$.*

The existence of a unique solution of (2.1) in $H^1(\Omega) \cap L^\infty(\Omega)$ is classical. It is a consequence of the monotonicity of f with respect to the second component. The continuity of y_u is also a well-known result; see, for instance, [15]. The continuity property is a consequence of the compactness of the inclusion $L^2(\Omega) \subset W^{-1,p}(\Omega)$ for any $p < 6$ and the fact that data $u \in W^{-1,p}(\Omega)$, with $6/5 < p < 6$, provide solutions in $C(\bar\Omega) \cap H^1_0(\Omega)$, the mapping $u \to y_u$ being continuous between these spaces.

THEOREM 2.2. *Let the function L be convex with respect to the third component and the set of feasible controls be nonempty. Then, under assumptions* (A1)–(A3)*, the control problem* (P) *has at least one solution.*

The proof of this theorem can be obtained by standard arguments.

Remark 2.3. We should remark that the differentiability of the functions f, L, and g is not necessary to prove the previous theorems. In fact, the only properties we need are the continuity of g and f with respect to the second variable, the continuity of L with respect to the second and third variables, the monotonicity of f with respect to y, the convexity of L with respect to u, and, for every $M > 0$, the existence of two functions $\phi_{f,M} \in L^2(\Omega)$ and $\phi_{L,M} \in L^1(\Omega)$ such that

$$|f(x,y)| \le \phi_{f,M}(x) \quad \text{and} \quad |L(x,y,u)| \le \phi_{L,M}(x) \text{ for a.e. } x \in \Omega \text{ and } |y|, |u| \le M.$$

These properties are an immediate consequence of the assumptions (A1)–(A3).

We finish this section by recalling some results about the differentiability of the nonlinear mappings involved in the control problem. For the detailed proofs, the reader is referred to Casas and Mateos [7].

THEOREM 2.4. *If* (A1) *holds, then the mapping $G : L^2(\Omega) \longrightarrow C(\bar{\Omega}) \cap H_0^1(\Omega)$, defined by $G(u) = y_u$, is of class C^2. Moreover, for all $v, u \in L^2(\Omega)$, $z_v = G'(u)v$ is defined as the solution of*

$$\begin{cases} Az_v + \dfrac{\partial f}{\partial y}(x, y_u) z_v = v \ \text{ in } \Omega, \\ z_v = 0 \ \text{ on } \Gamma. \end{cases} \tag{2.2}$$

Finally, for every $v_1, v_2 \in L^2(\Omega)$, $z_{v_1 v_2} = G''(u)v_1 v_2$ is the solution of

$$\begin{cases} Az_{v_1v_2} + \dfrac{\partial f}{\partial y}(x, y_u) z_{v_1v_2} + \dfrac{\partial^2 f}{\partial y^2}(x, y_u) z_{v_1} z_{v_2} = 0 \text{ in } \Omega, \\ z_{v_1v_2} = 0 \text{ on } \Gamma, \end{cases} \tag{2.3}$$

where $z_{v_i} = G'(u)v_i$, $i = 1, 2$.

Remark 2.5. This theorem shows why we assume $n \le 3$: To prove Theorem 4.1 on second-order sufficient conditions, we need the operator G to be differentiable from $L^2(\Omega)$ to $C(\bar{\Omega})$. This result holds true only for $n \le 3$.

The proof can be obtained by the implicit function theorem; see, for instance, [7, Thm. 2.5] for the proof in the case of a Neumann problem, which can be translated straightforwardly to the Dirichlet case.

THEOREM 2.6. *Suppose that* (A1) *and* (A2) *hold. Then $J : L^\infty(\Omega) \to \mathbb{R}$ is a functional of class C^2. Moreover, for every $u, v, v_1, v_2 \in L^\infty(\Omega)$,*

$$J'(u)v = \int_\Omega \left(\frac{\partial L}{\partial u}(x, y_u, u) + \varphi_{0u} \right) v \, dx \tag{2.4}$$

and

$$\begin{aligned} J''(u)v_1v_2 = \int_\Omega \Bigg[& \frac{\partial^2 L}{\partial y^2}(x, y_u, u) z_{v_1} z_{v_2} + \frac{\partial^2 L}{\partial y \partial u}(x, y_u, u)(z_{v_1} v_2 + z_{v_2} v_1) \\ & + \frac{\partial^2 L}{\partial u^2}(x, y_u, u) v_1 v_2 - \varphi_{0u} \frac{\partial^2 f}{\partial y^2}(x, y_u) z_{v_1} z_{v_2} \Bigg] \, dx, \end{aligned} \tag{2.5}$$

where $y_u = G(u)$ *and* $\varphi_{0u} \in W^{2,p}(\Omega)$ *is the unique solution of the problem*

$$\begin{cases} A^*\varphi + \dfrac{\partial f}{\partial y}(x, y_u)\varphi = \dfrac{\partial L}{\partial y}(x, y_u, u) \quad \text{in } \Omega, \\ \varphi = 0 \quad \text{on } \Gamma, \end{cases} \tag{2.6}$$

A^* *being the adjoint operator of* A *and* $z_{v_i} = G'(u)v_i$, $i = 1, 2$.

Let us remark that the linear and quadratic functionals $J'(u)$ and $J''(u)$ can be extended from $L^\infty(\Omega)$ to $L^2(\Omega)$ by the formulas (2.4) and (2.5). To check this point it is enough to use the assumptions (A1) and (A2). This extension will be used in the rest of the paper.

The previous theorem and the next one follow easily from Theorem 2.4 and the chain rule.

THEOREM 2.7. *Suppose that* (A1) *and* (A3) *hold. Then the mapping* $F : L^2(\Omega) \to C(K)$, *defined by* $F(u) = g(\cdot, y_u(\cdot))$, *is of class* C^2. *Moreover, for every* $u, v, v_1, v_2 \in L^2(\Omega)$,

$$F'(u)v = \frac{\partial g}{\partial y}(\cdot, y_u(\cdot))z_v(\cdot) \tag{2.7}$$

and

$$F''(u)v_1v_2 = \frac{\partial^2 g}{\partial y^2}(\cdot, y_u(\cdot))z_{v_1}(\cdot)z_{v_2}(\cdot) + \frac{\partial g}{\partial y}(\cdot, y_u(\cdot))z_{v_1v_2}(\cdot), \tag{2.8}$$

where $z_{v_i} = G'(u)v_i$, $i = 1, 2$, *and* $z_{v_1v_2} = G''(u)v_1v_2$.

Remark 2.8. A functional L that is very frequently appearing in the applications is given by

$$L(x, y, u) = L_0(x, y) + \frac{N}{2}u^2.$$

In this case, the functional J is twice differentiable not only in $L^\infty(\Omega)$ but also in $L^2(\Omega)$. Indeed, $J : L^2(\Omega) \longrightarrow \mathbb{R}$ is of class C^2, and the derivatives are given by the expressions

$$J'(u)v = \int_\Omega (Nu(x) + \varphi_{0u})\, v\, dx \tag{2.9}$$

and

$$J''(u)v_1v_2 = \int_\Omega \left[\frac{\partial^2 L_0}{\partial y^2}(x, y_u)z_{v_1}z_{v_2} + Nv_1v_2 - \varphi_{0u}\frac{\partial^2 f}{\partial y^2}(x, y_u)z_{v_1}z_{v_2} \right] dx. \tag{2.10}$$

Remark 2.9. The adjoint state φ_{0u} allows us to get a simple expression of $J'(u)$, but it is not the complete adjoint state of the control problem because the adjoint state equation (2.6) does not include the Lagrange multiplier associated to the state constraint; see (3.2) below for the full definition.

3. First-order optimality conditions. We define the Hamiltonian associated with problem (P), $H^\lambda : \Omega \times \mathbb{R}^3 \longrightarrow \mathbb{R}$, by

$$H^\lambda(x, y, u, \varphi) = \lambda \cdot L(x, y, u) + \varphi\,[u - f(x, y)].$$

We denote by $M(K)$ the Banach space of all real and regular Borel measures in K, which is identified with the dual space of $C(K)$.

In the rest of the paper, a local minimum of (P) is assumed to be a local solution in the sense of the topology of $L^\infty(\Omega)$. More precisely, we will say that $\bar u$ is a local minimum or a local solution of (P) in the sense of $L^q(\Omega)$, $1 \le q \le \infty$, if it is an admissible control of (P) and there exists $\varepsilon_{\bar u} > 0$ such that the minimum of J in the admissible set of (P) intersected with the ball $\bar B_{\varepsilon_{\bar u}}(\bar u) \subset L^q(\Omega)$ is achieved at $\bar u$.

The following result concerning Pontryagin's principle for problem (P) is well known; look into [8] and [17] as well as in the references therein for the proof.

THEOREM 3.1. *Let $\bar u$ be a local solution of* (P)*, and suppose that the assumptions* (A1)–(A3) *hold. Then there exist a real number $\bar\lambda \ge 0$, a measure $\bar\mu \in M(K)$, and a function $\bar\varphi \in W_0^{1,s}(\Omega)$, for all $1 \le s < n/(n-1)$, such that*

$$\bar\lambda + \|\bar\mu\| > 0 \tag{3.1}$$

$$\begin{cases} A^*\bar\varphi + \dfrac{\partial f}{\partial y}(x,\bar y(x))\bar\varphi = \bar\lambda\, \dfrac{\partial L}{\partial y}(x,\bar y,\bar u) + \dfrac{\partial g}{\partial y}(x,\bar y(x))\bar\mu \ \text{ in } \Omega, \\ \bar\varphi = 0 \ \text{ on } \Gamma, \end{cases} \tag{3.2}$$

$$\int_K (z(x) - g(x,\bar y(x)))\,d\bar\mu(x) \le 0 \quad \forall z \in C(K) \text{ such that } z(x) \le 0\ \forall x \in K, \tag{3.3}$$

$$H^{\bar\lambda}(x,\bar y(x),\bar u(x),\bar\varphi(x)) = \min_{t\in[\alpha_{\varepsilon_{\bar u}}(x),\beta_{\varepsilon_{\bar u}}(x)]} H^{\bar\lambda}(x,\bar y(x),t,\bar\varphi(x)) \ \text{ for a.e. } x \in \Omega, \tag{3.4}$$

where

$$\alpha_{\varepsilon_{\bar u}}(x) = \max\{\alpha(x), \bar u(x) - \varepsilon_{\bar u}\} \ \text{ and } \ \beta_{\varepsilon_{\bar u}}(x) = \min\{\beta(x), \bar u(x) + \varepsilon_{\bar u}\},$$

assuming that $\bar u$ is a minimum of (P) *in the ball $\bar B_{\varepsilon_{\bar u}}(\bar u) \subset L^\infty(\Omega)$. Moreover, if the following linearized Slater condition holds:*

$$\begin{aligned} &\exists u_0 \in L^\infty(\Omega), \text{with } \alpha(x) \le u_0(x) \le \beta(x) \ \text{ for a.e. } x \in \Omega, \text{such that} \\ &g(x,\bar y(x)) + \frac{\partial g}{\partial y}(x,\bar y(x)) z_{u_0-\bar u}(x) < 0 \quad \forall x \in K, \end{aligned} \tag{3.5}$$

where $\bar y$ is the state associated to $\bar u$ and $z_{u_0-\bar u} = G'(\bar u)(u_0 - \bar u)$, then the choice $\bar\lambda = 1$ can be made.

From now on, we take $\bar\lambda = 1$ and denote the Hamilton function by $H := H^1$.

Remark 3.2. Together with the inequality $g(x,\bar y(x)) \le 0$, relation (3.3) is equivalent to the well-known complementarity conditions

$$g(x,\bar y(x)) \le 0\ \forall x \in K, \quad \bar\mu \ge 0 \text{ in } M(K), \ \text{ and } \int_K g(x,\bar y(x))\,d\bar\mu(x) = 0.$$

It is also well known that (3.3) implies that $\bar\mu$ is a positive measure concentrated on the set of points

$$K_0 = \{x \in K : g(x,\bar y(x)) = 0\};$$

see, for instance, the references given before the statement of the previous theorem. From this property and assumption (A3), we deduce that $\bar\mu(K \cap \Gamma) = 0$.

Remark 3.3. By using elementary calculus, we obtain from (3.4) that

$$\frac{\partial H}{\partial u}(x,\bar y(x),\bar u(x),\bar\varphi(x))(k-\bar u(x))\ge 0\quad\forall k\in[\alpha(x),\beta(x)] \tag{3.6}$$

and

$$\frac{\partial^2 H}{\partial u^2}(x,\bar y(x),\bar u(x),\bar\varphi(x))\ge 0\ \text{ if }\ \frac{\partial H}{\partial u}(x,\bar y(x),\bar u(x),\bar\varphi(x))=0 \tag{3.7}$$

for a.e. $x\in\Omega$. On the other hand, notice that

$$\frac{\partial^2 L}{\partial u^2}(x,y,u)=\frac{\partial^2 H}{\partial u^2}(x,y,u,\varphi). \tag{3.8}$$

The inequality (3.6) implies that

$$\begin{cases}\dfrac{\partial H}{\partial u}(x,\bar y(x),\bar u(x),\bar\varphi(x))\ge 0 & \text{if }\ \bar u(x)=\alpha(x),\\[2mm] \dfrac{\partial H}{\partial u}(x,\bar y(x),\bar u(x),\bar\varphi(x))\le 0 & \text{if }\ \bar u(x)=\beta(x),\\[2mm] \dfrac{\partial H}{\partial u}(x,\bar y(x),\bar u(x),\bar\varphi(x))=0 & \text{if }\ \alpha(x)<\bar u(x)<\beta(x).\end{cases} \tag{3.9}$$

Reciprocally we also deduce from (3.6) that

$$\begin{cases}\bar u(x)=\alpha(x) & \text{if }\ \dfrac{\partial H}{\partial u}(x,\bar y(x),\bar u(x),\bar\varphi(x))>0,\\[2mm] \bar u(x)=\beta(x) & \text{if }\ \dfrac{\partial H}{\partial u}(x,\bar y(x),\bar u(x),\bar\varphi(x))<0.\end{cases} \tag{3.10}$$

The properties given by (3.8) and (3.9) are satisfied almost everywhere in Ω.

Remark 3.4. If we consider $\bar u$ in Theorem 3.1 to be a local minimum of (P) in the sense of $L^q(\Omega)$, $1\le q<+\infty$, then (3.4) can be written in the form (see [8])

$$H^{\bar\lambda}(x,\bar y(x),\bar u(x),\bar\varphi(x))=\min_{t\in[\alpha(x),\beta(x)]}H^{\bar\lambda}(x,\bar y(x),t,\bar\varphi(x))\ \text{ for a.e. } x\in\Omega.$$

Let us formulate the Lagrangian version of the optimality conditions (3.2)–(3.4). The Lagrange function $\mathcal{L}:L^\infty(\Omega)\times M(K)\longrightarrow\mathbb{R}$ associated to problem (P) is defined by

$$\mathcal{L}(u,\mu)=J(u)+\int_K g(x,y_u(x))\,d\mu(x)=\int_\Omega L(x,y_u(x),u(x))\,dx+\int_K g(x,y_u(x))\,d\mu(x).$$

By using (2.4) we find that

$$\frac{\partial\mathcal{L}}{\partial u}(u,\mu)v=\int_\Omega\left(\frac{\partial L}{\partial u}(x,y_u(x),u(x))+\varphi_u(x)\right)v(x)\,dx=\int_\Omega H_u(x)v(x)\ dx, \tag{3.11}$$

where

$$H_u(x)=\frac{\partial H}{\partial u}(x,y_u(x),u(x),\varphi(x)) \tag{3.12}$$

and $\varphi_u \in W_0^{1,s}(\Omega)$, for all $1 \le s < n/(n-1)$, is the solution of the Dirichlet problem

$$\begin{cases} A^*\varphi + \dfrac{\partial f}{\partial y}(x, y_u)\varphi = \dfrac{\partial L}{\partial y}(x, y_u, u) + \dfrac{\partial g}{\partial y}(x, y_u(x))\mu \ \text{ in } \Omega, \\ \varphi = 0 \ \text{ on } \Gamma. \end{cases} \tag{3.13}$$

Notice that the subscript u in y_u and H_u has a different meaning. While y_u is used to indicate that y is the state associated with u, H_u denotes the partial derivative of H with respect to u. This short notation for partial derivatives is frequently used in the following and will not cause confusion. Later, we also write H_{uu}, H_{yu}, or H_{yu} for $\partial^2 H/\partial u^2$, $\partial^2 H/\partial y \partial u$, etc.

If we insert $(\bar y(x), \bar u(x), \bar\varphi(x))$ into expression (3.12), then we denote $H_u(x)$ by $\bar H_u(x)$.

Now the inequality (3.6) along with (3.11) leads to

$$\frac{\partial \mathcal{L}}{\partial u}(\bar u, \bar\mu)(u - \bar u) \ge 0 \ \text{ if } \alpha(x) \le u(x) \le \beta(x) \text{ for a.e. } x \in \Omega \tag{3.14}$$

for any local solution $\bar u$, where $\bar y$ is the associated state and $\bar\varphi$ is the adjoint state given by (3.2), provided that (3.5) holds.

Before finishing this section we provide the expression of the second derivative of the Lagrangian with respect to the control, which will be used in the next section. From (2.8) we get

$$\frac{\partial^2 \mathcal{L}}{\partial u^2}(u, \mu)v_1 v_2 = J''(u)v_1 v_2 + \int_K \left[\frac{\partial^2 g}{\partial y^2}(x, y_u(x))z_{v_1}(x)z_{v_2}(x) + \frac{\partial g}{\partial y}(x, y_u(x))z_{v_1 v_2}(x)\right] d\mu(x).$$

By (2.5), this is equivalent to

$$\begin{aligned} \frac{\partial^2 \mathcal{L}}{\partial u^2}(u, \mu)v_1 v_2 = \int_\Omega \Bigg[& \frac{\partial^2 L}{\partial y^2}(x, y_u, u)z_{v_1} z_{v_2} + \frac{\partial^2 L}{\partial y \partial u}(x, y_u, u)(z_{v_1} v_2 + z_{v_2} v_1) \\ & + \frac{\partial^2 L}{\partial u^2}(x, y_u, u)v_1 v_2 - \varphi_u \frac{\partial^2 f}{\partial y^2}(x, y_u)z_{v_1} z_{v_2}\Bigg] dx \\ & + \int_K \frac{\partial^2 g}{\partial y^2}(x, y_u(x))z_{v_1}(x)z_{v_2}(x)\, d\mu(x), \end{aligned} \tag{3.15}$$

where φ_u is the solution of (3.13).

4. Second-order optimality conditions. Let $\bar u$ be a feasible control of problem (P) and $\bar y$ be the associated state. We assume that there exist $\bar\mu \in M(K)$ and $\bar\varphi \in W_0^{1,s}(\Omega)$, $1 \le s < n/(n-1)$, such that (3.2)–(3.4) are satisfied. As in the previous section, we use the notation

$$\bar H_u(x) := \frac{\partial H}{\partial u}(x, \bar y(x), \bar u(x), \bar\varphi(x)).$$

The partial derivative of H with respect to y at $(x, \bar y(x), \bar u(x), \bar\varphi(x))$ is denoted analogously by $\bar H_y(x)$.

Associated with $\bar{u}$, we define the cone of critical directions by

$$C_{\bar{u}} = \{h \in L^2(\Omega) : h \text{ satisfies (4.1), (4.2), and (4.3)}\},$$

$$h(x) = \begin{cases} \geq 0 \text{ if } \bar{u}(x) = \alpha(x), \\ \leq 0 \text{ if } \bar{u}(x) = \beta(x), \\ = 0 \text{ if } \bar{H}_u(x) \neq 0, \end{cases} \tag{4.1}$$

$$\frac{\partial g}{\partial y}(x, \bar{y}(x)) z_h(x) \leq 0 \quad \text{if} \quad g(x, \bar{y}(x)) = 0, \tag{4.2}$$

$$\int_K \frac{\partial g}{\partial y}(x, \bar{y}(x)) z_h(x)\, d\bar{\mu}(x) = 0. \tag{4.3}$$

If we think in terms of the finite-dimensional case, inequality (4.2) says that the derivative of the state constraint in the direction h is nonpositive if the constraint is active, and (4.3) states that this derivative is zero whenever the corresponding Lagrange multiplier is strictly positive. The relations (4.2)–(4.3) provide a convenient extension of the usual conditions in the finite-dimensional case.

We should mention that (4.3) is new in the context of infinite-dimensional optimization problems. In earlier papers on this subject, other extensions to the infinite-dimensional case were suggested. For instance, Maurer and Zowe [21] used first-order sufficient conditions to consider strict positivity of Lagrange multipliers. Inspired by their approach, in [12] an application to state-constrained elliptic boundary control was suggested. In terms of our problem, (4.3) was relaxed by

$$\int_K \frac{\partial g}{\partial y}(x, \bar{y}(x)) z_h(x)\, d\bar{\mu}(x) \geq -\varepsilon \int_{\Omega \setminus \Omega^\tau} |h(x)|\, dx$$

for some $\varepsilon > 0$; cf. [12, ineq. (5.15)]. Here $\Omega^\tau \subset \Omega$ is the set of points where $|\bar{H}_u(x)| \geq \tau$ holds true. We will prove that this relaxation is not necessary, which leads to a smaller cone of critical directions that seems to be optimal.

The sufficient second-order optimality conditions are given by the expressions (4.4) and (4.5) in the next theorem.

THEOREM 4.1. *Let $\bar{u}$ a feasible control of problem* (P)*, $\bar{y}$ be the associated state, and $(\bar{\varphi}, \bar{\mu}) \in W_0^{1,s}(\Omega) \times M(K)$, for all $1 \leq s < n/(n-1)$, satisfying* (3.2)–(3.4)*. Assume further that there exist two constants $\omega > 0$ and $\tau > 0$ such that*

$$\frac{\partial^2 L}{\partial u^2}(x, \bar{y}(x), \bar{u}(x)) \geq \omega \quad \textit{if} \quad |\bar{H}_u(x)| \leq \tau \quad \textit{for a.e. } x \in \Omega, \tag{4.4}$$

$$\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\mu}) h^2 > 0 \quad \forall h \in C_{\bar{u}} \setminus \{0\}. \tag{4.5}$$

Then there exist $\varepsilon > 0$ and $\delta > 0$ such that for every admissible control u of problem (P) *the following inequality holds:*

$$J(\bar{u}) + \frac{\delta}{2} \|u - \bar{u}\|^2_{L^2(\Omega)} \leq J(u) \quad \textit{if} \quad \|u - \bar{u}\|_{L^\infty(\Omega)} < \varepsilon. \tag{4.6}$$

Remark 4.2. Thanks to (3.8), we can compare the second-order necessary condition (3.7) with the sufficient one given by (4.4). We do not require only the strict positivity on the second derivative of the Hamiltonian with respect to the control at the points where the first derivative vanishes, as in the finite-dimensional case. We

also impose the second derivative to be strictly positive whenever the first derivative is "small." This is the usual case when we pass from finite to infinite dimension. For an instructive example the reader is referred to [14].

Inequality (4.4) is satisfied if the second derivative of L with respect to u is strictly positive for any $(y,u,\varphi)\in\mathbb{R}^3$ and almost all $x\in\Omega$. This assumption implies that L is strictly convex with respect to u. We recall that the convexity of L with respect to u was necessary to prove the existence of an optimal control. Under this strict convexity assumption, the sufficient second-order optimality conditions are reduced to (4.5). This is the case when $L(x,y,u)=L_0(x,y)+Nu^2/2$ if $N>0$.

The condition (4.5) seems to be natural. In fact, under some regularity assumption, we can expect the inequality

$$\frac{\partial^2\mathcal{L}}{\partial u^2}(\bar u,\bar\mu)h^2\ge 0\quad\forall h\in C_{\bar u}$$

to be a necessary condition for local optimality. At least this is the case when the state constraints are of integral type (see [7]) or when K is a finite set of points (see [6]).

Proof of Theorem 4.1. We argue by contradiction. Suppose that $\bar u$ does not satisfy the quadratic growth condition (4.6). Then there exists a sequence $\{u_k\}_{k=1}^\infty\subset L^2(\Omega)$ of feasible controls of (P) such that $u_k\to\bar u$ in $L^\infty(\Omega)$ and

$$J(\bar u)+\frac{1}{k}\|u_k-\bar u\|^2_{L^2(\Omega)}>J(u_k)\quad\forall k. \tag{4.7}$$

Let us take

$$\rho_k=\|u_k-\bar u\|_{L^2(\Omega)}\quad\text{and}\quad h_k=\frac{1}{\rho_k}(u_k-\bar u).$$

Since $\|h_k\|_{L^2(\Omega)}=1$, we can extract a subsequence, denoted in the same way, such that $h_k\rightharpoonup h$ weakly in $L^2(\Omega)$. Now we split the proof into several steps.

Step 1: $\frac{\partial\mathcal{L}}{\partial u}(\bar u,\bar\mu)h=0$. In the following, we write $y_k=y_{u_k}$. Since u_k is feasible, it holds that $g(x,y_k(x))\le 0$ for every $x\in K$. By using (3.3) and (4.7) we obtain

$$J(\bar u)+\frac{1}{k}\|u_k-\bar u\|^2_{L^2(\Omega)}=\mathcal{L}(\bar u,\bar\mu)+\frac{1}{k}\|u_k-\bar u\|^2_{L^2(\Omega)}>J(u_k)\ge\mathcal{L}(u_k,\bar\mu). \tag{4.8}$$

From the mean value theorem we know that

$$\mathcal{L}(u_k,\bar\mu)=\mathcal{L}(\bar u,\bar\mu)+\rho_k\frac{\partial\mathcal{L}}{\partial u}(v_k,\bar\mu)h_k,$$

with v_k a point between $\bar u$ and u_k. This identity and (4.8) imply that

$$\frac{\partial\mathcal{L}}{\partial u}(v_k,\bar\mu)h_k<\frac{1}{k\rho_k}\|u_k-\bar u\|^2_{L^2(\Omega)}=\frac{1}{k}\|u_k-\bar u\|_{L^2(\Omega)}.$$

Since $h_k\rightharpoonup h$ weakly in $L^2(\Omega)$, $v_k\to\bar u$ in $L^\infty(\Omega)$, $y_{v_k}\to\bar y$ in $C(\bar\Omega)\cap H^1_0(\Omega)$, and $\varphi_{v_k}\to\bar\varphi$ in $W^{1,s}_0(\Omega)\subset L^2(\Omega)$ for s close to $n/(n-1)$, we deduce from the above inequality and the expression of the derivative of the Lagrangian given by (3.11) that

$$\frac{\partial\mathcal{L}}{\partial u}(\bar u,\bar\mu)h=\lim_{k\to\infty}\frac{\partial\mathcal{L}}{\partial u}(v_k,\bar\mu)h_k\le 0. \tag{4.9}$$

On the other hand, since $\alpha(x) \le u_k(x) \le \beta(x)$ holds for almost all $x \in \Omega$, we deduce from the variational inequality (3.14) that

$$\frac{\partial \mathcal{L}}{\partial u}(\bar u, \bar\mu)h_k = \rho_k \frac{\partial \mathcal{L}}{\partial u}(\bar u, \bar\mu)(u_k - \bar u) \ge 0,$$

which implies that

$$\frac{\partial \mathcal{L}}{\partial u}(\bar u, \bar\mu)h = \lim_{k\to\infty} \frac{\partial \mathcal{L}}{\partial u}(v_k, \bar\mu)h_k \ge 0.$$

This inequality, along with (4.9), leads to

$$\frac{\partial \mathcal{L}}{\partial u}(\bar u, \bar\mu)h = 0. \tag{4.10}$$

Step 2: $h \in C_{\bar u}$. We have to confirm (4.1)–(4.3). The set of functions of $L^2(\Omega)$ that are nonnegative if $\bar u(x) = \alpha(x)$ and nonpositive if $\bar u(x) = \beta(x)$, almost everywhere, is convex and closed. Therefore, it is weakly closed. Moreover $u_k - \bar u$ obviously belongs to this set, and thus every h_k also does. Consequently, h belongs to the same set. Then (3.10), together with (3.12), implies that

$$\int_\Omega |\bar H_u(x)h(x)|\,dx = \int_\Omega \bar H_u(x)h(x)\,dx = \frac{\partial \mathcal{L}}{\partial u}(\bar u, \bar\mu)h = 0,$$

and hence $h(x) = 0$ if $\bar H_u(x) \ne 0$, which concludes the proof of (4.1).

Let us prove (4.2). From Theorem 2.4 we have

$$z_h = G'(\bar u)h = \lim_{k\to\infty} \frac{(y_{\bar u+\rho_k h_k} - \bar y)}{\rho_k} \quad \text{in } C(\bar\Omega) \cap H_0^1(\Omega),$$

which implies for every $x \in K$ such that $g(x, \bar y(x)) = 0$ that

$$\frac{\partial g}{\partial y}(x, \bar y(x))z_h(x) = \lim_{k\to\infty} \frac{[g(x, y_{\bar u+\rho_k h_k}(x)) - g(x, \bar y(x))]}{\rho_k} \le 0. \tag{4.11}$$

The last inequality follows from the fact that u_k is feasible, $\bar u + \rho_k h_k = u_k$, and consequently $g(x, y_{\bar u+\rho_k h_k}(x)) = g(x, y_{u_k}(x)) \le 0$ for every $x \in K$.

Finally, we prove (4.3). By taking $z = g(\cdot, y_{u_k}(\cdot))$ in (3.3), we get

$$\begin{aligned}\int_K \frac{\partial g}{\partial y}(x, \bar y(x))z_h(x)\,d\bar\mu(x) &= \lim_{k\to\infty} \frac{1}{\rho_k} \int_K [g(x, y_{\bar u+\rho_k h_k}(x)) - g(x, \bar y(x))]\,d\bar\mu(x) \\ &= \lim_{k\to\infty} \frac{1}{\rho_k} \int_K [g(x, y_{u_k}(x)) - g(x, \bar y(x))]\,d\bar\mu(x) \le 0. \end{aligned} \tag{4.12}$$

On the other hand, from (4.7) we find

$$J'(\bar u)h = \lim_{k\to\infty} \frac{J(\bar u + \rho_k h_k) - J(\bar u)}{\rho_k} = \lim_{k\to\infty} \frac{J(u_k) - J(\bar u)}{\rho_k} \le \lim_{k\to\infty} \frac{\rho_k}{k} = 0. \tag{4.13}$$

Then (4.10), (4.12), (4.13), and the fact that

$$\frac{\partial \mathcal{L}}{\partial u}(\bar u, \bar\mu)h = J'(\bar u)h + \int_K \frac{\partial g}{\partial y}(x, \bar y(x))z_h(x)\,d\bar\mu(x)$$

imply that

$$J'(\bar{u})h = \int_K \frac{\partial g}{\partial y}(x, \bar{y}(x)) z_h(x)\, d\bar{\mu}(x) = 0.$$

Thus (4.3) holds, and we know that $h \in C_{\bar{u}}$.

Step 3: $h = 0$. By taking into account (4.5), it is enough to prove that

$$\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\mu})h \le 0. \tag{4.14}$$

For this purpose, we evaluate the Lagrangian. By a second-order Taylor expansion, we derive

$$\mathcal{L}(u_k, \bar{\mu}) = \mathcal{L}(\bar{u}, \bar{\mu}) + \rho_k \frac{\partial \mathcal{L}}{\partial u}(\bar{u}, \bar{\mu})h_k + \frac{\rho_k^2}{2}\frac{\partial^2 \mathcal{L}}{\partial u^2}(w_k, \bar{\mu})h_k^2, \tag{4.15}$$

w_k being an intermediate point between $\bar{u}$ and u_k. From here we get

$$\begin{aligned} &\rho_k \frac{\partial \mathcal{L}}{\partial u}(\bar{u}, \bar{\mu})h_k + \frac{\rho_k^2}{2}\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\mu})h_k^2 \\ &= \mathcal{L}(u_k, \bar{\mu}) - \mathcal{L}(\bar{u}, \bar{\mu}) + \frac{\rho_k^2}{2}\left[\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\mu}) - \frac{\partial^2 \mathcal{L}}{\partial u^2}(w_k, \bar{\mu})\right] h_k^2. \end{aligned} \tag{4.16}$$

Now (4.8) can be written

$$\mathcal{L}(u_k, \bar{\mu}) - \mathcal{L}(\bar{u}, \bar{\mu}) \le \frac{\rho_k^2}{k}. \tag{4.17}$$

On the other hand, by taking into account the expression (3.15) of the second derivative of the Lagrangian, assumptions (A1)–(A3) and Theorems 2.1 and 2.4, and the fact that $u_k \to \bar{u}$ in $L^\infty(\Omega)$ and $\|h_k\|_{L^2(\Omega)} = 1$, we obtain

$$\begin{aligned} \left|\left[\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\mu}) - \frac{\partial^2 \mathcal{L}}{\partial u^2}(w_k, \bar{\mu})\right] h_k^2\right| &\le \left\|\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\mu}) - \frac{\partial^2 \mathcal{L}}{\partial u^2}(w_k, \bar{\mu})\right\|_{B(L^2(\Omega))} \|h_k\|^2_{L^2(\Omega)} \\ &= \left\|\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u}, \bar{\mu}) - \frac{\partial^2 \mathcal{L}}{\partial u^2}(w_k, \bar{\mu})\right\|_{B(L^2(\Omega))} \to 0 \quad \text{when} \quad k \to \infty, \end{aligned} \tag{4.18}$$

where $B(L^2(\Omega))$ is the space of quadratic forms in $L^2(\Omega)$.

Let us define

$$\Omega^\tau = \{x \in \Omega : |\bar{H}_u(x)| > \tau\}.$$

From (3.10) and the definition of h_k we know that $\bar{H}_u(x)h_k(x) \ge 0$ in Ω; therefore

$$\frac{\partial \mathcal{L}}{\partial u}(\bar{u}, \bar{\mu})h_k = \int_\Omega \bar{H}_u(x)h_k(x)\, dx \ge \int_{\Omega^\tau} |\bar{H}_u(x)||h_k(x)|\, dx \ge \tau \int_{\Omega^\tau} |h_k(x)|\, dx. \tag{4.19}$$

For any $\varepsilon > 0$ we can take k_ε such that

$$\|\rho_k h_k\|_{L^\infty(\Omega)} = \|\bar{u} - u_k\|_{L^\infty(\Omega)} < \varepsilon \quad \forall k \ge k_\varepsilon \quad \text{for a.e. } x \in \Omega,$$

and therefore

$$\frac{\rho_k^2 h_k^2(x)}{\varepsilon} \le \rho_k |h_k(x)| \quad \forall k \ge k_\varepsilon \quad \text{for a.e. } x \in \Omega.$$

From this inequality and (4.19) it follows that

$$\rho_k \frac{\partial \mathcal{L}}{\partial u}(\bar u, \bar\mu) h_k \ge \rho_k \tau \int_{\Omega^\tau} |h_k(x)|\, dx \ge \frac{\rho_k^2 \tau}{\varepsilon} \int_{\Omega^\tau} h_k^2(x)\, dx. \tag{4.20}$$

By collecting (4.16)–(4.18) and (4.20) and dividing by $\rho_k^2/2$, we obtain for any $k \ge k_\varepsilon$

$$\frac{2\tau}{\varepsilon} \int_{\Omega^\tau} h_k^2(x)\, dx + \frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar u, \bar\mu) h_k^2 \le \frac{2}{k} + \left\| \frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar u, \bar\mu) - \frac{\partial^2 \mathcal{L}}{\partial u^2}(w_k, \bar\mu) \right\|_{B(L^2(\Omega))}. \tag{4.21}$$

Next, we study the left-hand side of this inequality. First of all let us notice that from (3.15) we obtain for any $v \in L^2(\Omega)$

$$\begin{aligned} \frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar u, \bar\mu) v^2 = &\int_\Omega \left[\bar H_{uu}(x) v^2(x) + 2 \bar H_{uy}(x) z_v(x) v(x) + \bar H_{yy}(x) z_v^2(x) \right] dx \\ &+ \int_K \frac{\partial^2 g}{\partial y^2}(x, \bar y(x)) z_v^2(x)\, d\bar\mu(x), \end{aligned}$$

where

$$\bar H_{uu}(x) = \frac{\partial^2 H}{\partial u^2}(x, \bar y(x), \bar u(x), \bar\varphi(x))$$

and $\bar H_{uy}$ and $\bar H_{yy}$ are defined analogously. We also recall that

$$\bar H_{uu}(x) = \frac{\partial^2 L}{\partial u^2}(x, \bar y(x), \bar u(x)). \tag{4.22}$$

Then we have

$$\begin{aligned} &\frac{2\tau}{\varepsilon} \int_{\Omega^\tau} h_k^2(x)\, dx + \frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar u, \bar\mu) h_k^2 \\ &\quad = \int_{\Omega^\tau} \left(\frac{2\tau}{\varepsilon} + \bar H_{uu}(x) \right) h_k^2(x)\, dx + \int_{\Omega \setminus \Omega^\tau} \bar H_{uu}(x) h_k^2(x)\, dx \\ &\qquad + \int_\Omega \left[2 \bar H_{uy}(x) z_{h_k}(x) h_k(x) + \bar H_{yy}(x) z_{h_k}^2(x) \right] dx \\ &\qquad + \int_K \frac{\partial^2 g}{\partial y^2}(x, \bar y(x)) z_{h_k}^2(x)\, d\bar\mu(x). \end{aligned} \tag{4.23}$$

From assumptions (A1)–(A3) we deduce the existence of $C > 0$ such that $|\bar H_{uu}(x)| \le C$ for a.e. $x \in \Omega$. Therefore we can take $\varepsilon > 0$ small enough so that the following inequality holds:

$$\frac{2\tau}{\varepsilon} + \bar H_{uu}(x) \ge \frac{2\tau}{\varepsilon} - C > 0 \quad \text{for a.e. } x \in \Omega^\tau.$$

Thus

$$\liminf_{k \to \infty} \int_{\Omega^\tau} \left(\frac{2\tau}{\varepsilon} + \bar H_{uu}(x) \right) h_k^2(x)\, dx \ge \int_{\Omega^\tau} \left(\frac{2\tau}{\varepsilon} + \bar H_{uu}(x) \right) h^2(x)\, dx. \tag{4.24}$$

Moreover from (4.4) we have $\bar{H}_{uu}(x) \geq \omega > 0$ in $\Omega \setminus \Omega^\tau$, and therefore we also get

$$\liminf_{k\to\infty} \int_{\Omega\setminus\Omega^\tau} \bar{H}_{uu}(x)h_k^2(x)\,dx \geq \int_{\Omega\setminus\Omega^\tau} \bar{H}_{uu}(x)h^2(x)\,dx. \tag{4.25}$$

Finally, by taking into account that $z_{h_k} \to z_h$ strongly in $C(\bar{\Omega}) \cap H_0^1(\Omega)$, we deduce from (4.21)–(4.24) and (4.18) that

$$\begin{aligned}&\int_{\Omega^\tau} \left(\frac{2\tau}{\varepsilon} + \bar{H}_{uu}(x)\right) h^2(x)\,dx + \int_{\Omega\setminus\Omega^\tau} \bar{H}_{uu}(x)h^2(x)\,dx\\ &+ \int_\Omega [2\bar{H}_{uy}(x)h(x)z_h(x) + \bar{H}_{yy}(x)z_h^2(x)]dx + \int_K \frac{\partial^2 g}{\partial y^2}(x,\bar{y}(x))z_h^2(x)\,d\bar{\mu}(x) \leq 0.\end{aligned} \tag{4.26}$$

This expression can be written as follows:

$$\frac{2\tau}{\varepsilon}\int_{\Omega^\tau} h^2(x)\,dx + \frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u},\bar{\mu})h^2 \leq 0,$$

which along with (4.5) and the fact that $h \in C_{\bar{u}}$ implies that $h = 0$.

Step 4: $h_k \to 0$ strongly in $L^2(\Omega)$. We have already proved that $h_k \rightharpoonup 0$ weakly in $L^2(\Omega)$; therefore $z_{h_k} \to 0$ strongly in $C(\bar{\Omega}) \cap H_0^1(\Omega)$. By using (4.21) and (4.23) and the fact that $\|h_k\|_{L^2(\Omega)} = 1$, we conclude that

$$\begin{aligned}0 &< \min\left\{\omega, \frac{2\tau}{\varepsilon} - C\right\} = \min\left\{\omega, \frac{2\tau}{\varepsilon} - C\right\} \limsup_{k\to\infty} \int_\Omega h_k^2(x)\,dx\\ &\leq \limsup_{k\to\infty} \left\{\int_{\Omega^\tau} \left(\frac{2\tau}{\varepsilon} + \bar{H}_{uu}(x)\right) h_k^2(x)\,dx + \int_{\Omega\setminus\Omega^\tau} \bar{H}_{uu}(x)h_k^2(x)\,dx\right\}\\ &\leq \limsup_{k\to\infty} \left\{\frac{2}{k} + \left\|\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar{u},\bar{\mu}) - \frac{\partial^2 \mathcal{L}}{\partial u^2}(w_k,\bar{\mu})\right\|_{B(L^2(\Omega))}\right.\\ &\left. - \int_K \frac{\partial^2 g}{\partial y^2}(x,\bar{y}(x))z_{h_k}^2(x)\,d\bar{\mu}(x) - \int_\Omega \left[2\bar{H}_{uy}(x)z_{h_k}(x)h_k(x) + \bar{H}_{yy}(x)z_{h_k}^2(x)\right]dx\right\} = 0.\end{aligned}$$

Thus we have the contradiction. □

There is a very interesting particular case of (P) where Theorem 4.1 has a stronger formulation.

THEOREM 4.3. *Assume that $L(x,y,u) = L_0(x,y) + Nu^2/2$, with $N > 0$. If $\bar{u}$ is a feasible control of problem* (P)*, $\bar{y}$ is the associated state, $(\bar{\varphi},\bar{\mu}) \in W_0^{1,s}(\Omega) \times M(K)$, for all $1 \leq s < n/(n-1)$, and $(\bar{y},\bar{u},\bar{\varphi},\bar{\mu})$ satisfies* (3.2)–(3.4) *and* (4.5)*, then there exist $\varepsilon > 0$ and $\delta > 0$ such that for every admissible control u of problem* (P) *the following inequality holds:*

$$J(\bar{u}) + \frac{\delta}{2}\|u-\bar{u}\|_{L^2(\Omega)}^2 \leq J(u) \quad \text{if} \quad \|u-\bar{u}\|_{L^2(\Omega)} < \varepsilon. \tag{4.27}$$

We have already mentioned in Remark 4.2 that the first-order optimality conditions along with (4.5) are sufficient for optimality when $L(x,y,u) = L_0(x,y) + Nu^2/2$, with $N > 0$. But the above theorem includes more very important information. Relation (4.27) says that $\bar{u}$ is a strict local minimum of (P) in $L^2(\Omega)$. The fact that the control appears linearly in the state equation and quadratically in the cost functional allows us to get sufficient optimality conditions for a local minimum not only in $L^\infty(\Omega)$ but also in $L^2(\Omega)$. This fact is very important in the analysis of stability and convergence of numerical algorithms to solve (P). The proof of Theorem 4.3 follows the same arguments and steps as those given in the proof of Theorem 4.1. The essential fact is that the functional J is of class C^2 in $L^2(\Omega)$; see Remark 2.8.

5. Bilateral state constraints. In this section we will consider the extension of the control problem to the case of bilateral state constraints. More precisely we formulate the control problem as follows:

$$(\mathrm{P})\quad \begin{cases} \min J(u) = \displaystyle\int_\Omega L(x, y_u(x), u(x))\,dx \\ \text{subject to } \ (y_u, u) \in (C(\bar\Omega) \cap H^1(\Omega)) \times L^\infty(\Omega), \\ \alpha(x) \le u(x) \le \beta(x) \quad \text{for a.e. } \ x \in \Omega, \\ g_a(x) \le g(x, y_u(x)) \le g_b(x) \ \ \forall x \in K, \end{cases}$$

where $g_a, g_b : K \mapsto \mathbb{R}$ are continuous functions and $g_a(x) < g_b(x)$ for every $x \in K$. We assume the same hypotheses as in the previous sections. All of the previous theorems remain valid with some obvious modifications that we are going to mention. The Slater assumption required in Theorem 3.5 is now formulated as follows:

$$\begin{aligned} &\exists u_0 \in L^\infty(\Omega), \text{with } \alpha(x) \le u_0(x) \le \beta(x) \text{ for a.e. } x \in \Omega, \text{ such that} \\ &g_a(x) < g(x, \bar y(x)) + \frac{\partial g}{\partial y}(x, \bar y(x)) z_{u_0 - \bar u}(x) < g_b(x) \quad \forall x \in K. \end{aligned} \tag{5.1}$$

Under this assumption, Theorem 3.1 remains valid except for (3.3), which is written now in the following way:

$$\int_K (z(x)) - g(x, \bar y(x))\, d\bar\mu(x) \le 0 \ \forall z \in C(K), \text{ with } g_a(x) \le z(x) \le g_b(x) \ \forall x \in K. \tag{5.2}$$

From (5.2) we deduce that $\bar\mu$ is concentrated at the set of points K_0 where the state constraint is active:

$$K_0 = K_- \cup K_+ = \{x \in K : g(x, \bar y(x)) = g_a(x)\} \cup \{x \in K : g(x, \bar y(x)) = g_b(x)\}.$$

Now the Lagrange multiplier $\bar\mu$ is not necessarily a positive measure. However, its Jordan decomposition into nonnegative measures $\bar\mu^+$, $\bar\mu^-$ is as follows:

$$\bar\mu = \bar\mu^+ - \bar\mu^-, \ \text{ with } \operatorname{supp}(\bar\mu^+) \subset K_+ \text{ and } \operatorname{supp}(\bar\mu^-) \subset K_-.$$

The cone of critical directions $C_{\bar u}$ is formed by the functions $h \in L^2(\Omega)$ satisfying (4.1) and

$$\frac{\partial g}{\partial y}(x, \bar y(x)) z_h(x) \le 0 \ \text{ if } \ g(x, \bar y(x)) = g_b(x), \tag{5.3}$$

$$\frac{\partial g}{\partial y}(x, \bar y(x)) z_h(x) \ge 0 \ \text{ if } \ g(x, \bar y(x)) = g_a(x), \tag{5.4}$$

$$\int_K \left| \frac{\partial g}{\partial y}(x, \bar y(x)) z_h(x) \right| d|\bar\mu|(x) = 0, \tag{5.5}$$

where $|\bar\mu| = \bar\mu^+ + \bar\mu^-$. Then Theorem 4.1 is still true, and the only changes of the proof appear in Steps 1 and 2. In particular, (4.8) can be rewritten with the help

of (3.3) in the following way:

$$\mathcal{L}(\bar u,\bar\mu)+\frac{1}{k}\|u_k-\bar u\|^2_{L^2(\Omega)}-\int_K g(x,\bar y(x))\,d\bar\mu(x)=J(\bar u)+\frac{1}{k}\|u_k-\bar u\|^2_{L^2(\Omega)}$$
$$>J(u_k)\ge\mathcal{L}(u_k,\bar\mu)-\int_K g(x,\bar y(x))\,d\bar\mu(x)\ge\mathcal{L}(u_k,\bar\mu),$$

and the proof can continue as in Theorem 4.1.

On the other hand, relation (4.11) in Step 2 must be replaced by

$$\frac{\partial g}{\partial y}(x,\bar y(x))z_h(x)=\begin{cases}\le 0 & \forall x\in K_+,\\ \ge 0 & \forall x\in K_-.\end{cases}\tag{5.6}$$

Relations (4.12) and (4.13) remain valid. Finally, by using (4.10) and (5.6) we deduce the identity (5.5) as follows:

$$\int_K\left|\frac{\partial g}{\partial y}(x,\bar y(x))z_h(x)\right|\,d|\bar\mu|(x)=-\int_K\frac{\partial g}{\partial y}(x,\bar y(x))z_h(x)\,d\bar\mu(x)=J'(\bar u)h=0.$$

6. Elliptic boundary control.

6.1. Problem statement. The method of the preceding sections can be extended to other types of equations in a straightforward way. Here we discuss the case of boundary control, while the next section is devoted to a one-dimensional distributed parabolic control problem. Instead of (2.1), we consider now

$$\begin{cases} Ay+f(x,y)=0 & \text{in }\Omega,\\ \partial_\nu y+\gamma\,y=u & \text{on }\Gamma,\end{cases}\tag{6.1}$$

where ∂_ν denotes the conormal-derivative associated with A and $\gamma\in L^\infty(\Gamma)$ is nonnegative with $\gamma\not\equiv 0$. In contrast to section 1, we assume here that $n=2$. We need this stronger assumption, since now the control-to-state mapping G must be twice continuously differentiable from $L^2(\Gamma)$ to $C(\bar\Omega)$; cf. Remark 2.5. The differential operator A is defined as in section 1.

We consider the optimal boundary control problem

$$(\mathrm{PB})\begin{cases}\min J(u)=\displaystyle\int_\Omega L(x,y_u(x))\,dx+\int_\Gamma \ell(x,y_u(x),u(x))\,ds(x)\\ \text{subject to}\quad (y_u,u)\in(C(\bar\Omega)\cap H^1(\Omega))\times L^\infty(\Gamma),\\ \alpha(x)\le u(x)\le\beta(x)\quad\text{for a.e. } x\in\Gamma,\\ g(x,y_u(x))\le 0\ \ \forall x\in K.\end{cases}\tag{6.2}$$

Here α and β are now functions from $L^\infty(\Gamma)$, with $\alpha(x)\le\beta(x)$ for a.a. $x\in\Gamma$, ds denotes the surface measure on Γ, y_u is the solution of (6.1) associated with $u\in L^2(\Gamma)$, and $K\subset\bar\Omega$ is again a compact set.

The following assumptions are imposed on the data: We assume (A1)–(A3) on f, L, and g (where, of course, the dependence of L on u in (A2) is redundant). Moreover, we require the following.

(A4) The function $\ell:\Gamma\times(\mathbb{R}\times\mathbb{R})\longrightarrow\mathbb{R}$ satisfies assumption (A2) with ℓ substituted for L and Γ substituted for Ω.

Remark 6.1. We confine ourselves to a linear boundary condition. An extension to a nonlinear condition of the type $\partial_\nu y + b(x,y) = u$ is possible under associated assumptions on b. On the other hand, the assumption $\gamma \not\equiv 0$, that allows us to deduce the existence of a unique solution of (6.1), can be replaced by

$$\frac{\partial f}{\partial y}(x,t) > 0 \quad \text{for all } x \in E \text{ and } t \in \mathbb{R},$$

where E is a measurable subset of Ω with a strictly positive measure.

The proof of the next theorems is completely analogous to that of Theorems 2.2 and 2.4; see Alibert and Raymond [1].

THEOREM 6.2. *Suppose that* (A1) *holds. Then, for every* $u \in L^2(\Gamma)$*, the state equation* (6.1) *has a unique solution* $y_u \in C(\bar\Omega) \cap H^1(\Omega)$*. Furthermore, if* $u_k \rightharpoonup u$ *weakly in* $L^2(\Gamma)$*, then* $y_{u_k} \to y_u$ *strongly in* $C(\bar\Omega) \cap H^1(\Omega)$*.*

Notice that controls of $L^2(\Gamma)$ are transformed continuously to states in the Hölder space $C^{0,\kappa}(\Omega)$, with some $0 < \kappa < 1$; cf. Stampacchia [23, Thm. 14.2]. The second part of the statement is an immediate conclusion.

THEOREM 6.3. *Assume that* (A1)–(A4) *are fulfilled, the function* ℓ *is convex with respect to the third component, and the set of feasible controls is nonempty. Then the control problem* (PB) *has at least one solution.*

The proof can be performed by standard methods.

6.2. Necessary optimality conditions. We first state results on the first- and second-order derivatives of the control-to-state mapping $G(u) = y_u$ and of the reduced objective functional J. The results are analogous to Theorems 2.6–2.7 so that we only collect them without proof, since the associated modifications are obvious. Under assumptions (A1)–(A4), all mappings listed below are of class C^2 from $L^\infty(\Gamma)$ to their respective image spaces. The associated derivatives can be obtained as follows.

We define, for $v \in L^2(\Gamma)$, the function z_v as the unique solution to

$$\begin{cases} Az_v + \dfrac{\partial f}{\partial y}(x,y_u)z_v = 0 \ \text{ in } \Omega, \\ \partial_\nu z_v + \gamma z_v = v \ \text{ on } \Gamma. \end{cases} \tag{6.3}$$

Then G' is given by $G'(u)v = z_v$. Moreover, for v_1, $v_2 \in L^2(\Gamma)$, we introduce $z_{v_i} = G'(u)v_i$, $i = 1, 2$, and obtain $G''(u)v_1v_2 = z_{v_1v_2}$, where $z_{v_1v_2}$ is the solution to

$$\begin{cases} Az_{v_1v_2} + \dfrac{\partial f}{\partial y}(x,y_u)z_{v_1v_2} + \dfrac{\partial^2 f}{\partial y^2}(x,y_u)z_{v_1}z_{v_2} = 0 \text{ in } \Omega, \\ \partial_\nu z_{v_1v_2} + \gamma\, z_{v_1v_2} = 0 \text{ on } \Gamma. \end{cases} \tag{6.4}$$

The adjoint state $\varphi_{0u} \in H_0^1(\Omega)$ associated with u and J is introduced as the unique solution to

$$\begin{cases} A^*\varphi + \dfrac{\partial f}{\partial y}(x,y_u)\varphi = \dfrac{\partial L}{\partial y}(x,y_u) \ \text{ in } \Omega, \\ \partial_\nu \varphi + \gamma\, \varphi = \dfrac{\partial \ell}{\partial y}(x,y_u,u) \ \text{ on } \Gamma. \end{cases} \tag{6.5}$$

It holds that

$$J'(u)v = \int_\Gamma \left(\frac{\partial \ell}{\partial u}(x, y_u, u) + \varphi_{0u} \right) v \, ds, \tag{6.6}$$

$$\begin{aligned} J''(u)v_1v_2 = &\int_\Omega \left[\frac{\partial^2 L}{\partial y^2}(x, y_u, u) z_{v_1} z_{v_2} - \varphi_{0u} \frac{\partial^2 f}{\partial y^2}(x, y_u) z_{v_1} z_{v_2} \right] dx \\ &+ \int_\Gamma \left[\frac{\partial^2 \ell}{\partial y^2}(x, y_u, u) z_{v_1} z_{v_2} + \frac{\partial^2 \ell}{\partial y \partial u}(x, y_u, u)(z_{v_1} v_2 + z_{v_2} v_1) \right. \\ &\left. + \frac{\partial^2 \ell}{\partial u^2}(x, y_u, u) v_1 v_2 \right] ds. \end{aligned} \tag{6.7}$$

Under (A1) and (A3), the mapping $F : L^2(\Gamma) \to C(K)$, defined by $F(u) = g(\cdot, y_u(\cdot))$, is of class C^2. For every $u, v, v_1, v_2 \in L^2(\Gamma)$, its first- and second-order derivatives are given again by (2.7) and (2.8), respectively.

Now we introduce the Hamiltonian H by

$$H(x, y, u, \varphi) = \ell(x, y, u) + \varphi \, [u - \gamma \, y].$$

The first-order necessary conditions admit the following form.

THEOREM 6.4. *Let $\bar u$ be a local solution of* (PB). *Suppose that assumptions* (A1)–(A4) *hold, and assume the linearized Slater condition* (3.5) *with some $u_0 \in L^\infty(\Gamma)$, $\alpha(x) \le u_0(x) \le \beta(x)$ for a.e. $x \in \Gamma$. Then there exist a measure $\bar\mu \in M(K)$ and a function $\bar\varphi \in W^{1,s}(\Omega)$ for all $1 \le s < n/(n-1)$ such that*

$$\begin{cases} A^* \bar\varphi + \dfrac{\partial f}{\partial y}(x, \bar y(x)) \bar\varphi = \dfrac{\partial L}{\partial y}(x, \bar y, \bar u) + \dfrac{\partial g}{\partial y}(x, \bar y(x)) \bar\mu_{|\Omega} \ \text{ in } \Omega, \\ \partial_\nu \bar\varphi + \gamma \, \bar\varphi = \dfrac{\partial g}{\partial y}(x, \bar y(x)) \bar\mu_{|\Gamma} \ \text{ on } \Gamma, \end{cases} \tag{6.8}$$

$$\int_K (z(x) - g(x, \bar y(x))) d\bar\mu(x) \le 0 \quad \forall z \in C(K) \text{ such that } z(x) \le 0 \ \forall x \in K, \tag{6.9}$$

$$H(x, \bar y(x), \bar u(x), \bar\varphi(x)) = \min_{t \in [\alpha_{\varepsilon \bar u}(x), \beta_{\varepsilon \bar u}(x)]} H(x, \bar y(x), t, \bar\varphi(x)) \quad \text{for a.e. } x \in \Gamma, \tag{6.10}$$

where $\alpha_{\varepsilon_{\bar u}}$ and $\beta_{\varepsilon_{\bar u}}$ are defined similarly as in Theorem 3.1 *and $\bar\mu_{|\Omega}$ and $\bar\mu_{|\Gamma}$ denote the restrictions of μ to Ω and Γ, respectively.*

At the optimal point, the derivatives of H fulfill the relations (3.6)–(3.10) with an obvious modification: We have to substitute $x \in \Gamma$ for $x \in \Omega$. Moreover, we have to replace (3.8) by

$$\frac{\partial^2 \ell}{\partial u^2}(x, y, u) = \frac{\partial^2 H}{\partial u^2}(x, y, u, \varphi). \tag{6.11}$$

The Lagrangian function $\mathcal{L} : L^\infty(\Gamma) \times M(K) \longrightarrow \mathbb{R}$ associated to problem (PB) is defined by

$$\mathcal{L}(u, \mu) = \int_\Omega L(x, y_u(x)) \, dx + \int_\Gamma \ell(x, y_u(x), u(x)) \, ds + \int_K g(x, y_u(x)) \, d\mu(x).$$

By using (6.6) we deduce that

$$\frac{\partial \mathcal{L}}{\partial u}(u,\mu)v = \int_\Gamma H_u(x)v(x)\,ds, \tag{6.12}$$

where

$$H_u(x) = \frac{\partial H}{\partial u}(x, y(x), u(x), \varphi_u(x)) \tag{6.13}$$

and φ_u is obtained from the adjoint equation (6.8), where y_u is substituted for $\bar y$, u for $\bar u$, and μ for $\bar\mu$, respectively. We finally indicate the expression for the second-order derivative of $\mathcal{L}$:

$$\begin{aligned}\frac{\partial^2 \mathcal{L}}{\partial u^2}(u,\mu)v_1v_2 = &\int_\Omega \left[\frac{\partial^2 L}{\partial y^2}(x,y_u)z_{v_1}z_{v_2} - \varphi_u \frac{\partial^2 f}{\partial y^2}(x,y_u)z_{v_1}z_{v_2}\right] dx\\ &+ \int_\Gamma \left[\frac{\partial^2 \ell}{\partial y^2}(x,y_u,u)z_{v_1}z_{v_2} + \frac{\partial^2 \ell}{\partial y \partial u}(x,y_u,u)(z_{v_1}v_2 + z_{v_2}v_1)\right.\\ &\left. + \frac{\partial^2 \ell}{\partial u^2}(x,y_u,u)v_1v_2\right] ds\\ &+ \int_K \frac{\partial^2 g}{\partial y^2}(x,y_u(x))z_{v_1}(x)z_{v_2}(x)\,d\mu(x),\end{aligned} \tag{6.14}$$

where φ_u is defined as after (6.13).

6.3. Second-order sufficient optimality conditions. Let $\bar u$ be a feasible control of problem (PB) and $\bar y$ be the associated state. We assume that there exist $\bar\mu \in M(K)$ and $\bar\varphi \in W_0^{1,s}(\Omega)$, $1 \le s < n/(n-1)$, such that the first-order necessary conditions (6.8)–(6.10) are satisfied. Associated with $\bar u$, we introduce the function

$$\bar H_u(x) = \frac{\partial H}{\partial u}(x, \bar y(x), \bar u(x), \bar\varphi(x))$$

and define the cone of critical directions by

$$C_{\bar u} = \{h \in L^2(\Gamma) : h \text{ satisfies (4.1)–(4.3) with } x \in \Gamma\}. \tag{6.15}$$

Notice that this cone is only formally the same as in (4.1)–(4.3), since x varies here through Γ. The second-order sufficient condition admits now the following form.

THEOREM 6.5. *Assume that $n = 2$, and let $\bar u$ be a feasible control of problem* (PB)*, $\bar y$ the associated state, and $(\bar\varphi, \bar\mu) \in W^{1,s}(\Omega) \times M(K)$, for all $1 \le s < n/(n-1)$, satisfying* (6.8)–(6.10)*. Let there exist two constants $\omega > 0$ and $\tau > 0$ such that*

$$\frac{\partial^2 \ell}{\partial u^2}(x, \bar y(x), \bar u(x)) \ge \omega \quad \text{if} \quad |\bar H_u(x)| \le \tau \quad \text{for a.e. } x \in \Gamma, \tag{6.16}$$

$$\frac{\partial^2 \mathcal{L}}{\partial u^2}(\bar u, \bar\mu)h^2 > 0 \quad \forall h \in C_{\bar u} \setminus \{0\}, \tag{6.17}$$

where $C_{\bar u}$ is defined in (6.15) *and $\partial^2\mathcal{L}/\partial u^2$ is taken from* (6.14)*, with $u := \bar u$ and $\mu := \bar\mu$.*

Then there exist $\varepsilon > 0$ and $\delta > 0$ such that, for every admissible control u of problem (PB)*, the following inequality holds:*

$$J(\bar u) + \frac{\delta}{2}\|u - \bar u\|^2_{L^2(\Gamma)} \le J(u) \quad \text{if} \quad \|u - \bar u\|_{L^\infty(\Gamma)} < \varepsilon. \tag{6.18}$$

Proof. The proof is almost identical with the one of Theorem 4.1. Therefore, we mention only where essential changes occur.

Throughout the proof, we have to perform the obvious modification that $L^2(\Gamma)$, $L^\infty(\Gamma)$, and $H^1(\Omega)$ must be substituted for $L^2(\Omega)$, $L^\infty(\Omega)$, and $H_0^1(\Omega)$, respectively. Moreover, in some integrals, Ω must obviously be replaced by Γ. Then Steps 1 and 2 can be adopted without further changes.

Step 3: The arguments up to (4.18) do not need changes. Next, we modify Ω^τ by

$$\Gamma^\tau = \{x \in \Gamma \,:\, |\bar H_u(x)| > \tau\}.$$

Hereafter, Ω and Ω^τ are replaced by Γ and Γ^τ, respectively. In (4.22), ℓ must be substituted for L, and in (4.23) we add the integral over $\partial^2 L/\partial u^2$ to arrive at

$$\begin{aligned}
&\frac{2\tau}{\varepsilon}\int_{\Gamma^\tau} h_k^2(x)\,ds + \frac{\partial^2\mathcal{L}}{\partial u^2}(\bar u,\bar\mu)h_k^2 \\
&\quad= \int_{\Gamma^\tau}\left(\frac{2\tau}{\varepsilon} + \bar H_{uu}(x)\right) h_k^2(x)\,ds + \int_{\Gamma\setminus\Gamma^\tau} \bar H_{uu}(x)h_k^2(x)\,ds \\
&\qquad+ \int_\Gamma \left[2\bar H_{uy}(x)z_{h_k}(x)h_k(x) + \bar H_{yy}(x)z_{h_k}^2(x)\right] ds \\
&\qquad+ \int_K \frac{\partial^2 g}{\partial y^2}(x,\bar y(x))z_{h_k}^2(x)\,d\bar\mu(x) + \int_\Omega \frac{\partial^2 L}{\partial y^2}(x,\bar y(x))z_{h_k}^2(x)\,dx. \qquad (6.19)
\end{aligned}$$

Analogously, this term must be added to the left-hand side of (4.26).

Step 4: First, we conclude from $h_k \rightharpoonup 0$ in $L^2(\Gamma)$ that $z_{h_k} \to 0$ strongly in $C(\bar\Omega)$. Proceeding as in the former Step 4, we finally conclude with

$$\begin{aligned}
0 < \limsup_{k\to\infty}\Bigg\{ &\frac{2}{k} + \left\|\frac{\partial^2\mathcal{L}}{\partial u^2}(\bar u,\bar\mu) - \frac{\partial^2\mathcal{L}}{\partial u^2}(w_k,\bar\mu)\right\|_{B(L^2(\Gamma))} \\
&- \int_K \frac{\partial^2 g}{\partial y^2}(x,\bar y(x))z_{h_k}^2(x)\,d\bar\mu(x) - \int_\Omega \frac{\partial^2 L}{\partial y^2}(x,\bar y(x))z_{h_k}^2(x)\,dx \\
&- \int_\Omega \left[2\bar H_{uy}(x)z_{h_k}(x)h_k(x) + \bar H_{yy}(x)z_{h_k}^2(x)\right] dx\Bigg\} = 0. \qquad \square
\end{aligned}$$

7. The parabolic case.

7.1. Problem statement. Finally we prove that our method can also be extended to one-dimensional parabolic problems with distributed control. This extension is addressed here. To define the parabolic problem, we consider the one-dimensional domain $\Omega = (a,b)$ and the time interval $[0,T]$ for given $T > 0$. We fix an initial value $y_0 \in C[a,b]$ and introduce the set $Q = (a,b)\times(0,T)$. Moreover, we introduce the space $W(0,T) = \{y \in L^2(0,T;H^1(\Omega)) \,:\, \frac{dy}{dt} \in L^2(0,T;H^1(\Omega)')\}$.

Remark 7.1. Again, the restriction on the dimension of Ω comes from the requirement that the control-to-state mapping is of class C^2 from $L^2(Q)$ to $C(\bar Q)$. This holds true only for $n = 1$. We should mention here that boundary controls cannot be handled by our approach. Neumann boundary data from $L^2(0,T)$ are not, in general, transformed into continuous states.

The parabolic equation is defined by

$$\text{(7.1)}\qquad \begin{cases} \dfrac{dy}{dt} + Ay + f(x,t,y) = u & \text{in } (a,b)\times(0,T),\\ -\partial_x y(a,t) = 0 & \text{in } (0,T),\\ \partial_x y(b,t) = 0 & \text{in } (0,T),\\ y(\cdot,0) = y_0 & \text{in } (a,b), \end{cases}$$

where ∂_x denotes the partial derivative with respect to x. The associated optimal control problem is

$$\text{(7.2)}\quad \text{(PP)}\quad \begin{cases} \min J(u) = \displaystyle\int_0^T \int_a^b L(x,t,y_u(x,t),u(x,t))\,dxdt + \int_a^b r(x,y(x,T))\,dx \\ \qquad\qquad + \displaystyle\int_0^T \ell_a(t,y_u(a,t))\,dt + \int_0^T \ell_b(t,y_u(b,t))\,dt \\ \text{subject to } \ (y_u,u)\in (C(\bar{Q})\cap W(0,T))\times L^\infty(Q), \\ \alpha(x,t)\le u(x,t)\le \beta(x,t) \quad \text{for a.e. } \ (x,t)\in Q, \\ g(x,t,y_u(x,t))\le 0 \ \ \forall (x,t)\in K. \end{cases}$$

Here α and β are functions from $L^\infty(Q)$, with $\alpha(x,t)\le\beta(x,t)$ for a.a. $(x,t)\in Q$, y_u is the solution of (7.1) associated with $u\in L^2(Q)$, and $K\subset\bar{Q}$ is a compact set.

The following assumptions are required.

(A5) The function $f: Q\times\mathbb{R}\longrightarrow\mathbb{R}$ satisfies the modification of assumption (A1) that is obtained by substituting Q for Ω and (x,t) for x, respectively.

(A6) The function $L: Q\times(\mathbb{R}\times\mathbb{R})\longrightarrow\mathbb{R}$ satisfies the modified assumption (A2) obtained by substituting Q for Ω and (x,t) for x, respectively.

(A7) The function $g: K\times\mathbb{R}\longrightarrow\mathbb{R}$ is continuous and is of class C^2 with respect to the second variable, and $\partial_y g$ and $\partial_y^2 g$ are also continuous functions in $K\times\mathbb{R}$. Moreover, the strict inequality

$$\text{(7.3)}\qquad g(x,0,y_0(x))<0$$

holds for every $x\in K\cap\bar{\Omega}$.

(A8) The functions $\ell_k:[0,T]\times\mathbb{R}\longrightarrow\mathbb{R}$, $k\in\{a,b\}$, are Carathéodory functions of class C^2 with respect to the second variable with $\ell_k(\cdot,0)\in L^1(0,T)$. For all $M>0$, there exist a constant $C_M>0$ and a function $\psi_M\in L^2(0,T)$ such that

$$\left|\frac{\partial\ell_k}{\partial u}(t,y)\right|\le\psi_M(x), \quad \left|\frac{\partial^2\ell_k}{\partial y^2}(t,y)\right|\le C_M,$$
$$\left|\frac{\partial^2\ell_k}{\partial y^2}(t,y_2)-\frac{\partial^2\ell_k}{\partial y^2}(t,y_1)\right|\le C_M\,|y_2-y_1|$$

holds for $k\in\{a,b\}$ for a.e. $t\in[0,T]$ and $|y|,|y_i|\le M$, $i=1,2$.

Analogously, $r:[a,b]\times\mathbb{R}\longrightarrow\mathbb{R}$ is a Carathéodory function of class C^2 with respect to the second variable with $r(\cdot,0)\in L^1(a,b)$. It satisfies the assumptions on ℓ_k above with ℓ_k replaced by r, (a,b) substituted for $(0,T)$, and x substituted for t.

For the parabolic equation, the following result on existence and regularity holds true.

THEOREM 7.2. *Suppose that* (A5) *is satisfied. Then, for every* $u \in L^2(Q)$*, the state equation* (7.1) *has a unique solution* $y_u \in C(\bar{Q}) \cap W(0,T)$*. If* $u_k \rightharpoonup u$ *weakly in* $L^2(Q)$*, then* $y_{u_k} \to y_u$ *strongly in* $C(\bar{Q})$*.*

The proof of the theorem is postponed to section 7.4.

THEOREM 7.3. *Assume that* (A5)–(A8) *are fulfilled, the function* L *is convex with respect to the third component, and the set of feasible controls is nonempty. Then the control problem* (PP) *has at least one solution.*

This theorem is a standard consequence of Theorem 7.2.

7.2. Necessary optimality conditions. Also here, the control-to-state mapping $G(u) = y_u$, $G : L^2(Q) \to C(\bar{Q}) \cap W(0,T)$, and the reduced objective functional J are of class C^2 from $L^\infty(Q)$ to their image spaces, provided that assumptions (A5)–(A8) are satisfied. Since this is known (see [5]) we state the associated derivatives for convenience below.

We define, for $v \in L^2(Q)$, the function z_v as the unique solution to

$$
\left\{
\begin{aligned}
\frac{dz_v}{dt} + Az_v + \frac{\partial f}{\partial y}(x,t,y_u)z_v &= v \quad \text{in } Q,\\
-\partial_x z_v(a,t) &= 0 \quad \text{in } (0,T),\\
\partial_x z_v(b,t) &= 0 \quad \text{in } (0,T),\\
y(x,0) &= 0 \quad \text{in } (a,b).
\end{aligned}
\right. \tag{7.4}
$$

Then $G'(u)$, $G : L^2(Q) \to C(\bar{Q}) \cap W(0,T)$, is given by $G'(u)v = z_v$. Moreover, for v_1, $v_2 \in L^2(Q)$, we introduce $z_{v_i} = G'(u)v_i$, $i = 1, 2$, and obtain $G''(u)v_1v_2 = z_{v_1v_2}$, where $z_{v_1v_2}$ is the solution to

$$
\left\{
\begin{aligned}
\frac{dz_{v_1v_2}}{dt} + Az_{v_1v_2} + \frac{\partial f}{\partial y}(x,t,y_u)z_{v_1v_2} + \frac{\partial^2 f}{\partial y^2}(x,t,y_u)z_{v_1}z_{v_2} &= 0 \quad \text{in } Q,\\
-\partial_x z_{v_1v_2}(a,t) &= 0 \quad \text{in } (0,T),\\
\partial_x z_{v_1v_2}(b,t) &= 0 \quad \text{in } (0,T),\\
z_{v_1v_2}(x,0) &= 0 \quad \text{in } (a,b).
\end{aligned}
\right. \tag{7.5}
$$

The adjoint state $\varphi_{0u} \in W(0,T)$ associated with u and J is introduced as the unique solution to

$$
\left\{
\begin{aligned}
-\frac{d\varphi}{dt} + A^*\varphi + \frac{\partial f}{\partial y}(x,t,y_u)\varphi &= \frac{\partial L}{\partial y}(x,t,y_u,u) \quad \text{in } Q,\\
-\partial_x\varphi(a,t) &= \frac{\partial \ell_a}{\partial y}(t,y_u(a,t)) \quad \text{in } (0,T),\\
\partial_x\varphi(b,t) &= \frac{\partial \ell_b}{\partial y}(t,y_u(b,t)) \quad \text{in } (0,T),\\
\varphi(x,T) &= \frac{\partial r}{\partial y}(x,y_u(x,T)) \quad \text{in } (a,b).
\end{aligned}
\right. \tag{7.6}
$$

We have

$$J'(u)v = \int_Q \left(\frac{\partial L}{\partial u}(x,t,y_u,u) + \varphi_{0u} \right) v \, dxdt, \tag{7.7}$$

$$\begin{aligned} J''(u)v_1v_2 = \int_Q \Bigg[& \frac{\partial^2 L}{\partial y^2}(x,t,y_u,u) z_{v_1} z_{v_2} + \frac{\partial^2 L}{\partial y \partial u}(x,t,y_u,u)(z_{v_1}v_2 + z_{v_2}v_1) \\ & + \frac{\partial^2 L}{\partial u^2}(x,t,y_u,u)v_1v_2 - \varphi_{0u} \frac{\partial^2 f}{\partial y^2}(x,t,y_u) z_{v_1} z_{v_2} \Bigg] dxdt \\ & + \int_0^T \frac{\partial^2 \ell_a}{\partial y^2}(t, y_u(a,t)) z_{v_1}(a,t) z_{v_2}(a,t) \, dt \\ & + \int_0^T \frac{\partial^2 \ell_b}{\partial y^2}(t, y_u(b,t)) z_{v_1}(b,t) z_{v_2}(b,t) \, dt \\ & + \int_\Omega \frac{\partial^2 r}{\partial y^2}(x, y_u(x,T)) z_{v_1}(x,T) z_{v_2}(x,T) \, dx. \end{aligned} \tag{7.8}$$

We require the following linearized Slater condition: There exists $u_0 \in L^\infty(Q)$ with $\alpha(x,t) \le u_0(x,t) \le \beta(x,t)$ for a.e. $(x,t) \in Q$ such that

$$g(x,t,\bar y(x,t)) + \frac{\partial g}{\partial y}(x,t,\bar y(x,t)) z_{u_0-\bar u}(x,t) < 0 \quad \forall (x,t) \in K. \tag{7.9}$$

Notice that we have assumed (7.3), since this is needed to satisfy (7.9). The Hamiltonian H is defined by

$$H(x,t,y,u,\varphi) = L(x,t,y,u) + \varphi\,[u - f(x,t,y)],$$

and the first-order necessary conditions admit the following form (see Casas [5]).

THEOREM 7.4. *Let $\bar u$ be a local solution of* (PP). *Suppose that assumptions* (A5)–(A8) *hold, and assume the Slater condition* (7.9) *with some $u_0 \in L^\infty(Q)$, $\alpha(x,t) \le u_0(x,t) \le \beta(x,t)$ for a.e. $(x,t) \in Q$. Then there exist a measure $\bar\mu \in M(K)$ and a function $\bar\varphi \in L^\tau(0,T;W^{1,\sigma}(\Omega))$, for all τ, $\sigma \in [1,2)$, with $\frac{1}{\tau} + \frac{1}{\sigma} > \frac{3}{2}$, such that*

$$\begin{cases} -\dfrac{d\bar\varphi}{dt} + A^*\varphi + \dfrac{\partial f}{\partial y}(x,t,\bar y)\,\bar\varphi = \dfrac{\partial L}{\partial y}(x,t,\bar y,\bar u) + \dfrac{\partial g}{\partial y}(x,t,\bar y)\bar\mu_{|_Q}, \\ -\partial_x\bar\varphi(a,t) = \dfrac{\partial \ell_a}{\partial y}(t,y_u(a,t)) + \dfrac{\partial g}{\partial y}(a,t,\bar y(a,t))\bar\mu_{|_{\{a\}\times(0,T)}}, \\ \partial_x\bar\varphi(b,t) = \dfrac{\partial \ell_b}{\partial y}(t,\bar y(b,t)) + \dfrac{\partial g}{\partial y}(b,t,\bar y(b,t))\bar\mu_{|_{\{b\}\times(0,T)}}, \\ \bar\varphi(x,T) = \dfrac{\partial r}{\partial y}(x,\bar y(x,T)) + \dfrac{\partial g}{\partial y}(x,T,\bar y(x,T))\bar\mu_{|_{\Omega\times\{T\}}} \end{cases} \tag{7.10}$$

for a.a. $x \in (a,b)$, $t \in (0,T)$, where $\bar\mu_{|_Q}$, $\bar\mu_{|_{\{a\}\times(0,T)}}$, $\bar\mu_{|_{\{b\}\times(0,T)}}$, and $\bar\mu_{|_{\Omega\times\{T\}}}$ denote the restrictions of μ to Q, $\{a\}\times(0,T)$, $\{b\}\times(0,T)$, and $\Omega\times\{T\}$, respectively,

$$\int_K (z(x,t) - g(x,t,\bar y(x,t)))d\bar\mu(x,t) \le 0 \ \forall z \in C(K), \ \text{with } z(x,t) \le 0 \ \forall (x,t) \in K, \tag{7.11}$$

and, for almost all $(x,t) \in Q$,

$$H(x,t,\bar y(x,t),\bar u(x,t),\bar\varphi(x,t)) = \min_{s\in[\alpha_{\varepsilon\bar u}(x,t),\beta_{\varepsilon\bar u}(x,t)]} H(x,t,\bar y(x,t),s,\bar\varphi(x,t)), \tag{7.12}$$

where $\alpha_{\varepsilon\bar u}$ and $\beta_{\varepsilon\bar u}$ are defined along the lines of Theorem 3.1.

The Lagrange function is defined in a standard way by

$$\mathcal{L}(u,\mu) = \int_Q L(x,t,y_u(x,t),u(x,t))\,dxdt + \int_0^T \ell_a(t,y_u(a,t))\,dt + \int_0^T \ell_b(t,y_u(b,t))\,dt + \int_K g(x,t,y_u(x,t))\,d\mu(x,t).$$

For convenience, we establish only the second-order derivative of $\mathcal{L}$:

(7.13)
$$\begin{aligned}&\frac{\partial^2\mathcal{L}}{\partial u^2}(u,\mu)v_1v_2\\ &= \int_Q \left[\frac{\partial^2 L}{\partial y^2}(x,t,y_u,u)z_{v_1}z_{v_2} + \frac{\partial^2 L}{\partial y\partial u}(x,t,y_u,u)(z_{v_1}v_2+z_{v_2}v_1) + \frac{\partial^2 L}{\partial u^2}(x,t,y_u,u)v_1v_2\right.\\ &\quad \left. - \varphi_u\frac{\partial^2 f}{\partial y^2}(x,t,y_u)z_{v_1}z_{v_2}\right]dxdt + \int_\Omega \frac{\partial^2 r}{\partial y^2}(x,y_u(x,T))z_{v_1}(x,T)z_{v_2}(x,T)\,dx\\ &\quad + \int_0^T \left[\frac{\partial^2\ell_a}{\partial y^2}(t,y_u(a,t))z_{v_1}(a,t)z_{v_2}(a,t) + \frac{\partial^2\ell_b}{\partial y^2}(t,y_u(b,t))z_{v_1}(b,t)z_{v_2}(b,t)\right]dt\\ &\quad + \int_K \frac{\partial^2 g}{\partial y^2}(x,t,y_u(x,t))z_{v_1}(x)z_{v_2}(x)\,d\bar\mu(x,t),\end{aligned}$$

where φ_u is the solution of (7.10), where u is taken for $\bar u$, y_u instead of $\bar y$, and μ for $\bar\mu$.

7.3. Second-order sufficient optimality conditions. With the prerequisites of the preceding section at hand, the extension of the second-order sufficient optimality conditions to the parabolic case is straightforward. We define the cone of critical directions associated with $\bar u$ by

$$C_{\bar u} = \{h \in L^2(Q) : h \text{ satisfies (7.14), (7.15), and (7.16) below}\},$$

$$h(x,t) = \begin{cases} \geq 0 \text{ if } \bar u(x,t) = \alpha(x,t),\\ \leq 0 \text{ if } \bar u(x,t) = \beta(x,t),\\ = 0 \text{ if } \bar H_u(x,t) \neq 0,\end{cases} \tag{7.14}$$

$$\frac{\partial g}{\partial y}(x,t,\bar y(x,t))z_h(x,t) \leq 0 \quad \text{if} \quad g(x,t,\bar y(x,t)) = 0, \tag{7.15}$$

$$\int_K \frac{\partial g}{\partial y}(x,t,\bar y(x,t))z_h(x,t)\,d\bar\mu(x,t) = 0. \tag{7.16}$$

The sufficient second-order optimality conditions for $\bar u$ are stated in the following result.

THEOREM 7.5. *Let $\bar u$ be a feasible control of problem* (PP) *that satisfies, together with the associated state $\bar y$ and $(\bar\varphi, \bar\mu) \in L^\tau(0,T;W^{1,\sigma}(\Omega))\times M(K)$ for all τ, $\sigma \in [1,2)$, with $\frac{1}{\tau}+\frac{1}{\sigma} > \frac{3}{2}$, the first-order conditions* (7.10)–(7.12). *Assume in addition that there exist two constants $\omega > 0$ and $\tau > 0$ such that*

$$\frac{\partial^2 L}{\partial u^2}(x,t,\bar y(x,t),\bar u(x,t)) \geq \omega \quad \text{if} \quad |\bar H_u(x,t)| \leq \tau \quad \text{for a.e. } (x,t) \in Q, \tag{7.17}$$

$$\frac{\partial^2\mathcal{L}}{\partial u^2}(\bar u,\bar\mu)h^2 > 0 \quad \forall h \in C_{\bar u}\setminus\{0\}. \tag{7.18}$$

Then there exist $\varepsilon > 0$ and $\delta > 0$ such that, for every admissible control u of problem (PP)*, the following inequality holds:*

$$J(\bar u) + \frac{\delta}{2}\|u - \bar u\|^2_{L^2(Q)} \le J(u) \quad \text{if} \quad \|u - \bar u\|_{L^\infty(Q)} < \varepsilon. \tag{7.19}$$

The proof is analogous to the one of Theorem 4.1. We have to perform obvious modifications that are along the line of the ones explained in the proof of Theorem 6.5. Therefore, we skip these details.

7.4. Proof of Theorem 7.2. To prepare the proof of Theorem 7.2, we first state some results on maximal parabolic regularity of the elliptic differential operator A. In the one-dimensional case we study here, A admits the form

$$A = \frac{\partial}{\partial x}\left[a_{11}(\cdot)\frac{\partial}{\partial x}\right].$$

Let us consider A on its natural domain

$$D := D(A) = \left\{ w \in H^2(\Omega) \,:\, \frac{\partial w}{\partial x}(a) = \frac{\partial w}{\partial x}(b) = 0 \right\} \tag{7.20}$$

that is dense in $L^2(\Omega)$. It is known that, for all $\tau \in (0,1)$,

$$D(A^\tau) = \begin{cases} H^{2\tau}(\Omega) \cap \left\{ w \,:\, \dfrac{\partial w}{\partial x}(a) = \dfrac{\partial w}{\partial x}(b) = 0 \right\} & \text{if } \tau > \frac{3}{4}, \\ H^{2\tau}(\Omega), & \text{if } \tau < \frac{3}{4}; \end{cases}$$

cf. [24]. In particular, we have $D(A^{\frac12}) = H^1(\Omega)$. To shorten the notation, we write below $S := (0,T)$ with closure $\bar S$. Moreover, for a Banach space $X \subset L^1(\Omega)$ and $1 < p < \infty$, we introduce the space

$$W^{1,p}(S,X) = \left\{ y \in L^p(S,X) \,:\, \frac{\partial y}{\partial t} \in L^p(S,X) \right\}.$$

It is known that, for all $1 < p < \infty$, A exhibits maximal parabolic $L^p(S,L^p(\Omega))$-regularity. This means that, for all $f \in L^p(S,L^p(\Omega))$, there is a unique solution $y \in W^{1,p}(S,L^p(\Omega)) \cap L^p(S,D(A))$ of

$$\frac{\partial y}{\partial t} + A\,y = f \text{ in } S, \quad y(0) = 0, \tag{7.21}$$

where the differential equation is to be understood in the distributional sense; cf. [13]. Here the definition of $D(A)$ must be adapted by replacing $W^{2,p}(\Omega)$ for $H^2(\Omega)$ in (7.20). In all that follows, we apply this result with $p = 2$ for $X = H := L^2(\Omega)$. Therefore, for all $f \in L^2(S,H) \cong L^2(Q)$, there is a unique solution $y \in W^{1,2}(S,H) \cap L^2(S,D)$ of (7.21). The mapping $f \mapsto y$ is surjective and hence continuous.

Our proof relies on the following result.

LEMMA 7.6. *For all $0 < \tau < \eta < 1$ and $\kappa = \frac{\eta-\tau}{2\eta}$, there holds the continuous injection $W^{1,2}(S,H) \cap L^2(S,D) \hookrightarrow C^\kappa(S,H^\tau(\Omega))$.*

Proof. We show first that $W^{1,2}(S,H) \hookrightarrow C^{\frac12}(S,H)$. To this aim, let $y \in W^{1,2}(S,H)$ and $t,s \in \bar S$ be given. Then

$$\begin{aligned} \|y(t) - y(s)\|_H &= \left\| \int_s^t y'(\rho)\,d\rho \right\|_H \le \int_s^t \|y'(\rho)\|_H\,d\rho \\ &\le \left(\int_s^t \|y'(\rho)\|_H^2\,d\rho \right)^{\frac12} \left(\int_s^t d\rho \right)^{\frac12} \le \|y\|_{W^{1,2}(S,H)}\,|t-s|^{\frac12} \end{aligned}$$

verifies the injection claimed above. Next, we prove the statement of the lemma. We denote by $[\cdot,\cdot]_\theta$ the complex interpolation functor; see Triebel [24]. It follows from [2, Chap. III, Thm. 4.10.2] and [24, Chap. 1.8] that the continuous injection

$$W^{1,2}(S,H)\cap L^2(S,D)\hookrightarrow C(\bar S,[H,D]_{1/2})=C(\bar S,H^1(\Omega)) \tag{7.22}$$

takes place. The interpolation identity $[H,D]_{1/2}=H^1(\Omega)$ is well known and can be found, for instance, in [24].

We fix now τ and η by $0<\tau<\eta<1/2$ and put $\lambda=\tau/\eta$. Then we obtain with a generic constant c that

$$\begin{aligned}\frac{\|y(t)-y(s)\|_{[H,D]_\tau}}{|t-s|^{\frac12(1-\lambda)}} &\le c\,\frac{\|y(t)-y(s)\|_{[H,[H,D]_\eta]_\lambda}}{|t-s|^{\frac12(1-\lambda)}} \qquad (7.23)\\ &\le c\,\frac{\|y(t)-y(s)\|_H^{1-\lambda}}{|t-s|^{\frac12(1-\lambda)}}\,\|y(t)-y(s)\|^\lambda_{[H,D]_\eta}\\ &\le c\left(\frac{\|y(t)-y(s)\|_H}{|t-s|^{\frac12}}\right)^{1-\lambda}\|y(t)-y(s)\|^\lambda_{[H,D]_\eta}, \qquad (7.24)\end{aligned}$$

where we have applied the complex reiteration theorem, [24, Chap. 1.9.3]. In the last estimate, the first factor is bounded, since $W^{1,2}(S,H)\hookrightarrow C^{\frac12}(S,H)$. In view of the injection (7.22) and $[H,D]_\eta=H^{2\eta}(\Omega)$, with $0<2\eta<1$, the second factor can be estimated by

$$\begin{aligned}\|y(t)-y(s)\|^\lambda_{[H,D]_\eta}\le c\,\|y(t)-y(s)\|^\lambda_{H^{2\eta}(\Omega)} &\le \Big(2c\,\|y\|_{C(\bar S,H^{2\eta}(\Omega))}\Big)^\lambda\\ &\le \Big(c\,\|y\|_{C(\bar S,H^1(\Omega))}\Big)^\lambda\le c\,\|y\|^\lambda_{W^{1,2}(S,H)\cap L^2(S,D)}.\end{aligned}$$

In the last estimate, we have used the embedding (7.22). Moreover, we took advantage of the equivalence of the norms of $[H,D]_\eta$ and $H^{2\eta}(\Omega)$. Therefore, the second factor in (7.24) is bounded, too. The statement of the lemma follows now from (7.24) after inserting $\tau:=2\tau$, $\eta:=2\eta$, $[H,D]_\tau=H^{2\tau}(\Omega)$, and $\kappa=\frac12(1-\lambda)$ into (7.23). □

Proof of Theorem 7.2. The existence result of Theorem 7.2 is well known; we refer to Casas [5]. Therefore, we show only that weakly converging sequences of controls are transformed to strongly converging sequences of states.

Let a sequence (u_k) be given that converges weakly in $L^2(Q)$ to u. Consider the equation for y_k and u_k:

$$\begin{aligned}\frac{\partial y_k}{\partial t}+Ay_k+f(x,y_k)&=u_k \quad \text{in } Q,\\ \frac{\partial y_k}{\partial x}(a,t)&=0 \quad \text{in } S,\\ \frac{\partial y_k}{\partial x}(b,t)&=0 \quad \text{in } S,\\ y_k(0)&=y_0 \quad \text{in } \Omega.\end{aligned}$$

Standard arguments show that $y_k\rightharpoonup y$ in $W(0,T)\cap C(\bar Q)$, where $y=y_u$. The functions y_k are uniformly bounded in $C(\bar Q)$, hence the sequence $\big(d(\cdot,y_k)\big)$ is bounded in $L^2(Q)$, and we can select a weakly converging subsequence indexed by k_l. We write $f_k=u_k-d(\cdot,y_k)$ and split $y_k=v+w_k$, where w_k solves

$$\frac{\partial w_k}{\partial t}+Aw_k=f_k$$

with homogeneous inital and boundary conditions, while v solves

$$\frac{\partial v}{\partial t} + Av = 0$$

with inhomogeneous initial condition $v(0) = y_0$ and homogeneous boundary conditions. Thanks to Lemma 7.6, the sequence (w_{k_l}) converges weakly in $C^\kappa(S, H^\tau(\Omega))$, where $\kappa > 0$ and $\tau > 1/2$ can be chosen. Therefore, the functions w_{k_l} belong to a space $C^\sigma(\bar{Q})$ with some positive σ so that, by compact embedding into $C(\bar{Q})$, the sequence converges strongly in $C(\bar{Q})$. Consequently, $y_{k_l} = v + w_{k_l}$ converges strongly in $C(\bar{Q})$ towards y. Moreover, it follows by standard arguments that $y = y_u$. Since this holds for all subsequences with the same limit y, the whole sequence (y_k) converges uniformly to y_u. □

Acknowledgment. The authors are grateful to J. Rehberg (Weierstrass Institute Berlin (WIAS)) for pointing out the proof of Lemma 7.6.

REFERENCES

[1] J. J. Alibert and J. P. Raymond, *Boundary control of semilinear elliptic equations with discontinuous leading coefficients and unbounded controls*, Numer. Funct. Anal. Optim., 18 (1997), pp. 235–250.
[2] H. Amann, *Linear and Quasilinear Parabolic Problems*, Birkhäuser, Basel, 1995.
[3] F. Bonnans, *Second-order analysis for control constrained optimal control problems of semilinear elliptic systems*, Appl. Math. Optim., 38 (1998), pp. 303–325.
[4] F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer-Verlag, New York, 2000.
[5] E. Casas, *Pontryagin's principle for state-constrained boundary control problems of semilinear parabolic equations*, SIAM J. Control Optim., 35 (1997), pp. 1297–1327.
[6] E. Casas, *Necessary and sufficient optimality conditions for elliptic control problems with finitely many pointwise state constraints*, ESAIM Control Optim. Calc. Var., to appear.
[7] E. Casas and M. Mateos, *Second order optimality conditions for semilinear elliptic control problems with finitely many state constraints*, SIAM J. Control Optim., 40 (2002), pp. 1431–1454.
[8] E. Casas, J. Raymond, and H. Zidani, *Pontryagin's principle for local solutions of control problems with mixed control-state constraints*, SIAM J. Control Optim., 39 (2000), pp. 1182–1203.
[9] E. Casas and F. Tröltzsch, *Second-order necessary optimality conditions for some state-constrained control problems of semilinear elliptic equations*, Appl. Math. Optim., 39 (1999), pp. 211–227.
[10] E. Casas and F. Tröltzsch, *Second-order necessary and sufficient optimality conditions for optimization problems and applications to control theory*, SIAM J. Optim., 13 (2002), pp. 406–431.
[11] E. Casas, F. Tröltzsch, and A. Unger, *Second order sufficient optimality conditions for a nonlinear elliptic control problem*, J. Anal. Appl., 15 (1996), pp. 687–707.
[12] E. Casas, F. Tröltzsch, and A. Unger, *Second order sufficient optimality conditions for some state-constrained control problems of semilinear elliptic equations*, SIAM J. Control Optim., 38 (2000), pp. 1369–1391.
[13] G. Dore, *l^p-regularity for abstract differential equations*, in Functional Analysis and Related Topics, in Proceedings of the International Conference in memory of Prof. K. Yosida held at RIMS, Kyoto University, 1991, Lecture Notes Math. 1540, H. Komatsu, ed., Springer-Verlag, Berlin, 1993, pp. 25–38.
[14] J. Dunn, *l^2 sufficient optimality conditions for end-constrained optimal control problems with inputs in a polyhedron*, SIAM J. Control Optim., 36 (1998), pp. 1833–1851.
[15] D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1977.
[16] A. D. Ioffe, *Necessary and sufficient conditions for a local minimum. 3: Second order conditions and augmented duality*, SIAM J. Control Optim., 17 (1979), pp. 266–288.
[17] X. Li and J. Yong, *Necessary conditions for optimal control of distributed parameter systems*, SIAM J. Control Optim., 29 (1991), pp. 895–908.

[18] K. Malanowski, *Two-norm approach in stability and sensitivity analysis of optimization and optimal control problems*, Adv. Math. Sci. Appl., 2 (1993), pp. 397–443.
[19] K. Malanowski, *Sufficient optimality conditions for optimal control subject to state constraints*, SIAM J. Control Optim., 35 (1997), pp. 205–227.
[20] H. Maurer, *First and second order sufficient optimality conditions in mathematical programming and optimal control*, Math. Program. Study, 14 (1981), pp. 163–177.
[21] H. Maurer and J. Zowe, *First- and second-order conditions in infinite-dimensional programming problems*, Math. Program., 16 (1979), pp. 98–110.
[22] J.-P. Raymond and F. Tröltzsch, *Second order sufficient optimality conditions for nonlinear parabolic control problems with state constraints*, Discrete Contin. Dyn. Syst., 6 (2000), pp. 431–450.
[23] G. Stampacchia, *Problemi al contorno ellittici, con dati discontinui, dotati di soluzioni hölderiane*, Ann. Mat., 51 (1960), pp. 1–37.
[24] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, North–Holland, Amsterdam, 1978.

SIAM J. OPTIM.
Vol. 19, No. 2, pp. 644–654

ON STABLE FEASIBLE SETS IN GENERALIZED SEMI-INFINITE PROGRAMMING*

HARALD GÜNZEL†, HUBERTUS TH. JONGEN†, AND JAN-J. RÜCKMANN‡

Abstract. We consider the feasible set of a generalized semi-infinite programming problem with a one-dimensional index set of inequality constraints depending on the state variable. The latter dependence on the state variable gives rise to a complicated structure of the feasible set. Under appropriate transversality conditions we partially present the local description of feasible sets in new coordinates by means of the finitely many basic functions.

Key words. generalized semi-infinite programming, feasible set, transversal zero-point, transversality, normal form

AMS subject classifications. 90C34, 90C31, 90C30, 65C05

DOI. 10.1137/070694259

1. Introduction.

We consider the feasible set M of a generalized semi-infinite programming (GSIP) problem, which is defined as

$$M := \{x \in \mathbb{R}^n \mid v_0(x,y) \geq 0,\ y \in Y(x)\}$$

with the index set

$$Y(x) := \{y \in \mathbb{R} \mid v_i(x,y) \geq 0,\ i \in I = \{1,\dots,p\}\}.$$

For $x \in \mathbb{R}^n$ define the corresponding set of active constraints as

$$Y_0(x) := \{y \in Y(x) \mid v_0(x,y) = 0\}.$$

In GSIP the index set $Y(x)$ may depend on the state variable x. This is in contrast to standard semi-infinite programming (SIP) where the corresponding index set is fixed. GSIP became in recent years a substantial research area in mathematical programming. There exists a wide range of applications, we refer, e.g., to design-centering problems, reverse Chebyshev approximation, robust optimization, time-minimal control problems, and others. For a more detailed study see, e.g., the recent book [11] and the compilations [2], [10].

Throughout the paper we make the following two assumptions.

(A1) The set-valued mapping $x \in \mathbb{R}^n \mapsto Y(x) \subset \mathbb{R}$ is upper-semicontinuous in the sense of Berge [1], and $Y(x)$ is compact.

(A2) $v_i \in C^\infty(\mathbb{R}^n \times \mathbb{R}, \mathbb{R}),\ i = 0,\dots,p.$

Assumption (A1) is an usual assumption in GSIP. If $\bar{x} \in M$ and $Y_0(\bar{x}) = \emptyset$, then (A1) implies that $\bar{x} \in \operatorname{int} M$. We assume (A2) in order to avoid unnecessary technicalities. Note that (A2) is a mild differentiability assumption since the C^∞-functions are C^l_s-dense in the space of C^l-functions (here, C^l_s denotes the strong (Whitney) C^l-topology [5]).

*Received by the editors June 11, 2007; accepted for publication (in revised form) January 30, 2008; published electronically July 2, 2008.
http://www.siam.org/journals/siopt/19-2/69425.html

†Department of Mathematics, RWTH Aachen University, Templergraben 55, 52056 Aachen, Germany (guenzel@mathc.rwth-aachen.de, jongen@rwth-aachen.de).

‡School of Mathematics, University of Birmingham, Edgbaston, Birmingham B152TT, United Kingdom (ruckmanj@maths.bham.ac.uk).

The feasible set M in GSIP may exhibit some topological properties which do not appear in standard SIP, e.g.,

- M need not be closed;
- M may have a disjunctive structure (for details see [3], [6], [7], and [11]).

If $\bar{x} \in M$, then each $\bar{y} \in Y_0(\bar{x})$ is a global minimizer of the function $v_0(\bar{x}, .)$ subject to $y \in Y(\bar{x})$. By using the corresponding optimal value function, M can be described as

$$M = \{x \in \mathbb{R}^n \mid \min_{y \in Y(x)} v_0(x, y) \geq 0\}.$$

In the case where $Y(\bar{x}) = \emptyset$, the minimum in the latter description of M is defined to be ∞; in particular, in that case $\bar{x}$ belongs to M. If for $\bar{x} \in M$ each $y \in Y_0(\bar{x})$ is a nondegenerate minimizer of $v_0(\bar{x}, .)$ restricted to $Y(\bar{x})$, then the so-called reduction approach is available and in a neighborhood of $\bar{x}$ the set M can be described by finitely many differentiable functions (e.g., [4], [9]). Unfortunately, one cannot avoid degenerate minimizers $y \in Y_0(x)$ for some $x \in M$. Moreover, some of these degeneracies also remain stable under small perturbations.

The goal of this paper is to describe the local structure of M around $\bar{x} \in \partial M$ (the boundary of M) in new coordinates by means of the finitely many basic functions. If $\bar{x} \in \operatorname{int} M$, then the local structure of M is trivial since an open neighborhood of $\bar{x}$ belongs completely to M. The knowledge of the local structure of M gives insight into the complicated structure of this type of optimization problem, and might be useful for the design of corresponding solution methods.

DEFINITION 1. *Let $v \in C^\infty(\mathbb{R}^n \times \mathbb{R}, \mathbb{R})$. For $k \geq 1$, a point $\bar{y} \in \mathbb{R}$ is called a* zero-point *for $v(\bar{x}, .)$* of an order *k, if the following holds at $(\bar{x}, \bar{y})$:*

$$\left(\frac{\partial}{\partial y}\right)^i v = 0, \ i = 0, \ldots, k-1, \quad \left(\frac{\partial}{\partial y}\right)^k v \neq 0. \tag{1}$$

Moreover, the zero-point $\bar{y}$ for $v(\bar{x}, .)$ of an order k is called transversal*, if in addition to* (1)*, the Jacobian matrix with respect to (x, y) of the system of the equations $(\frac{\partial}{\partial y})^i v = 0$, $i = 0, \ldots, k-1$, has rank k at $(\bar{x}, \bar{y})$.*

The local description of M around $\bar{x} \in \partial M$ may become extremely complicated if the functions $v_i(\bar{x}, .)$, $i = 0, \ldots, p$, have the zero-points of arbitrary order. Therefore, we restrict our forthcoming analysis to the four most simple cases, where $\bar{x} \in \partial M$ and $\bar{y} \in Y(\bar{x})$ (note: If $Y(\bar{x}) = \emptyset$, then $\bar{x} \in M$ with $Y_0(\bar{x}) = \emptyset$, and, hence $\bar{x} \in \operatorname{int} M$). In the following we present a first characterization of these four cases; their complete definitions will be given in section 3.

Case A. $\bar{y}$ is a transversal zero-point for $v_0(\bar{x}, .)$ of an order k, $k \geq 2$, and $v_i(\bar{x}, \bar{y}) > 0$, $i \in I$.

Case B. $\bar{y}$ is a transversal zero-point for $v_0(\bar{x}, .)$ of an order k, $k \geq 2$, there exists an index $i_0 \in I$ such that $\bar{y}$ is a transversal zero-point for $v_{i_0}(\bar{x}, .)$ of an order one, and $v_i(\bar{x}, \bar{y}) > 0$, $i \in I \setminus \{i_0\}$.

Case C. $\bar{y}$ is a transversal zero-point for $v_i(\bar{x}, .)$ of an order one for each $i \in \{0\} \cup I$ with $v_i(\bar{x}, \bar{y}) = 0$.

Case D. The set $\{i \in I \mid v_i(\bar{x}, \bar{y}) = 0\}$ is a singleton, say $\{1\}$, and

- either $v_0(\bar{x}, \bar{y}) \neq 0$ and $\bar{y}$ is a transversal zero-point for $v_1(\bar{x}, .)$ of an order two,
- or $v_0(\bar{x}, \bar{y}) = 0$ and with $\{i_1, i_2\} = \{0, 1\}$, $\bar{y}$ is a transversal zero-point for $v_{i_1}(\bar{x}, .)$ of an order two and a transversal zero-point for $v_{i_2}(\bar{x}, .)$ of an order one or two.

Remark 2. The local analysis of these four cases will describe the existence of the above-mentioned phenomena (nonclosedness and the disjunctive structure of M) while a further analysis which includes the zero-points of higher order would only illustrate that the local fine structure of M becomes more complicated. This can already be seen by comparing the analysis of Cases A, B, and C with that of Case D. In this sense, the choice of these four cases can be considered as complete.

The results of this paper can be extended to the more general case, where M is described by finitely many equality constraints and finitely many inequality constraints of the type $v_0(x,y) \geq 0$, $y \in Y(x)$.

This paper is organized as follows. Section 2 contains some basic results, and section 3 presents the discussion of the four cases and a decomposition theorem in which the feasible set M around $\bar{x} \in \partial M$ is locally described in new coordinates by means of the finitely many basic functions. Since this decomposition theorem mainly uses a transversality condition, the description of M remains stable under sufficiently small smooth perturbations of the defining functions.

2. Basic results. We start with a definition and a lemma that is a consequence of [8], Lemma 1.

DEFINITION 3. *A local C^∞-coordinate transformation on $\mathbb{R}^n \times \mathbb{R}$, mapping $(x,y) \mapsto (u,w)$, is called* admissible*, if it can be written in the form*

$$\begin{aligned} u &= \xi^1(x), \\ w &= \xi^2(x,y). \end{aligned}$$

LEMMA 4. *Let $v \in C^\infty(\mathbb{R}^n \times \mathbb{R}, \mathbb{R})$, $\bar{x} \in \mathbb{R}^n$, and let $\bar{y} \in \mathbb{R}$ be a transversal zero-point for $v(\bar{x},.)$ of an order k, $k \geq 2$. Let $\varepsilon := \operatorname{sign}(\frac{\partial}{\partial y})^k v(\bar{x},\bar{y})$. Then, there exists an admissible C^∞-coordinate transformation sending $(\bar{x},\bar{y})$ to $(0,0)$ such that v has in new coordinates the form*

$$\varepsilon y^k + \sum_{i=1}^{k-1} x_i y^{k-1-i}, \tag{2}$$

where the new coordinates are again denoted by x and y. Let the associated set of vectors $S(\bar{y})$ be defined as

$$S(\bar{y}) := \left\{ \left(\frac{\partial}{\partial y} \right)^i v_x(\bar{x},\bar{y}),\ i = 0, \ldots, k-2 \right\}$$

(v_x denotes the partial derivative of v with respect to x). Then, the vectors in $S(\bar{y})$ are linearly independent.

Remark 5. Let $\varepsilon = 1$, and k be even. Then, the local optimal value function $\min_y v(x,y)$, takes in new coordinates the form

$$x \mapsto \eta^{k-1}(x_1, \ldots, x_{k-1}) = \min_{y \in [-1,1]} y^k + \sum_{i=1}^{k-1} x_i y^{k-1-i}. \tag{3}$$

In particular, if $k = 2$, we have $\eta^1(x_1) = x_1$.

Remark 6. The associated set $S(\bar{y})$ is related to a fundamental part of the Jacobian matrix $D\xi^1(\bar{x})$. Moreover, $\frac{\partial}{\partial y}\xi^2(\bar{x},\bar{y}) > 0$. See [8] for details.

As done in the lemma, we will always denote the resulting coordinates after an admissible coordinate transformation for (x,y) by (x,y).

Now, we consider the following special situation, which will be related later to case D. Let $\bar{x} \in \mathbb{R}^n$, and let $\bar{y} \in \mathbb{R}$ be a zero-point for $v_0(\bar{x}, .)$ of an order two and a zero-point for $v_1(\bar{x}, .)$ of an order two such that $v_i(\bar{x}, \bar{y}) > 0$ for $i = 2, \ldots, p$. Furthermore, assume that the Jacobian matrix of the system

$$v_0 = 0, \quad v_1 = 0, \quad v_{0,y} = 0, \quad v_{1,y} = 0$$

has rank four at $(\bar{x}, \bar{y})$ (by $v_{0,x}$, $v_{0,y}$, $v_{0,xx}$, $v_{0,xy} \ldots$ we denote the corresponding partial derivatives of v_0). Omitting all arguments $(\bar{x}, \bar{y})$ the latter Jacobian is the first entry in (4), and since the following matrices in (4)–(5) are the results of simple row manipulations, the matrix in (5) also has rank four:

$$\begin{pmatrix} v_{0,x} & 0 \\ v_{1,x} & 0 \\ v_{0,yx} & v_{0,yy} \\ v_{1,yx} & v_{1,yy} \end{pmatrix} \qquad \begin{pmatrix} v_{0,x} & 0 \\ v_{1,x} & 0 \\ v_{0,yx}(v_{0,yy})^{-1} & 1 \\ v_{1,yx}(v_{1,yy})^{-1} & 1 \end{pmatrix} \tag{4}$$

$$\left(\begin{array}{l|l} v_{0,x} & 0 \\ v_{1,x} & 0 \\ v_{0,yx}(v_{0,yy})^{-1} - v_{1,yx}(v_{1,yy})^{-1} & 0 \\ \hline v_{1,yx}(v_{1,yy})^{-1} & 1 \end{array}\right). \tag{5}$$

The rows of the upper-left submatrix, marked in (5), must therefore be linearly independent. They form the associated set of vectors

$$S(\bar{y}) := \left\{ v_{0,x},\ v_{1,x},\ v_{0,yx}(v_{0,yy})^{-1} - v_{1,yx}(v_{1,yy})^{-1} \right\}.$$

Let $y_i(x)$ denote the locally unique solution of the system $v_{i,y}(x, .) = 0$, $i = 0, 1$ for x close to $\bar{x}$. Applying the chain rule, one obtains

$$Dy_i(x) = -v_{i,yx}(v_{i,yy})^{-1}\big|_{y=y_i(x)}, \quad i = 0, 1.$$

Hence, $S(\bar{y})$ is linearly independent if and only if the Jacobian at $x = \bar{x}$ of the mapping $x \mapsto (v_0(x, y_0(x)),\ v_1(x, y_1(x)),\ y_1(x) - y_0(x))$ has rank three. The latter mapping is related to the coordinate transformation in the following lemma.

LEMMA 7. *Let the assumptions of the above special situation be satisfied. Then, there exists an admissible local C^∞-coordinate transformation sending $(\bar{x}, \bar{y})$ to $(0, 0)$ such that in the new coordinates the functions v_0, v_1 take the form:*

$$\begin{aligned} v_0(x, y) &= \pm y^2 \mp x_1, \\ v_1(x, y) &= \alpha(x, y)(y - x_2)^2 + x_3, \end{aligned}$$

where $\alpha \in C^\infty(\mathbb{R}^n \times \mathbb{R}, \mathbb{R})$ with $\alpha(0, 0) \neq 0$.

Proof. First we show that there exists a local representation of v_i with smooth functions γ_i, $\gamma_i(\bar{x}, \bar{y}) \neq 0$, $i = 0, 1$ such that

$$v_i(x, y) = \gamma_i(x, y)(y - y_i(x))^2 + v_i(x, y_i(x)). \tag{6}$$

The formula (6) can be derived as follows (omitting the index i):

$$\begin{aligned} v(x, y) &= v(x, y) - v(x, y(x)) + v(x, y(x)) \\ v(x, y) - v(x, y(x)) &= \int_0^1 \frac{d}{dt} v(x, ty + (1 - t)y(x))dt \\ &= \int_0^1 v_y(x, ty + (1 - t)y(x))(y - y(x))dt \\ &= (y - y(x))h(x, y), \end{aligned} \tag{7}$$

where $h(x,y)$ stands for the part separated by $y - y(x)$. Noting $h(x,y(x)) = 0$, we can use the same integral representation for $h(x,y)$, obtaining (with $\gamma(x,y)$ denoting the analogous part separated by $y - y(x)$)

$$h(x,y) = (y - y(x))\gamma(x,y). \tag{8}$$

Substituting (8) in (7), we get (6).

Now we take $(y - y_0(x))\sqrt{|\gamma_0(x,y)|}$ as a new y-coordinate and $v_0(x, y_0(x))$ as new x-coordinate, say, x_1 if $\gamma_0(\bar{x},\bar{y}) < 0$ (or $-x_1$ if $\gamma_0(\bar{x},\bar{y}) > 0$). Let $\tilde{v}_1$ denote the function v_1 in the new coordinates, and let $\tilde{y}_1(x)$ be the corresponding solution of $\tilde{v}_{1,y}(x,.) = 0$ for x near $\bar{x}$. In an analogous way as in (6) we can write $\tilde{v}_1$ as follows:

$$\tilde{v}_1(x,y) = \tilde{\gamma}_1(x,y)(y - \tilde{y}_1(x))^2 + \tilde{v}_1(x, \tilde{y}_1(x)).$$

Finally, we take $\tilde{y}_1(x)$ and $\tilde{v}_1(x,\tilde{y}_1(x))$ as new coordinates, say, x_2 and x_3 (without affecting x_1 and y). Then, the linear independence of the vectors in $S(\bar{y})$ guarantees that the above procedure results in an admissible local C^∞-coordinate transformation, indeed. This completes the proof. □

3. Main results.

3.1. The four cases. In this subsection we will discuss the cases A, B, C, and D mentioned in section 1. Throughout this subsection we assume that $\bar{x} \in \partial M$, $\bar{y} \in Y(\bar{x})$ (hence $v_i(\bar{x},\bar{y}) \geq 0$, $i \in I$), and that the set

$$I^* := \{i \in \{0\} \cup I \mid v_i(\bar{x},\bar{y}) = 0\}$$

is nonempty. Note that $\bar{x}$ need not be feasible since M is not necessarily closed. If $\bar{y}$ is a transversal zero-point for $v_i(\bar{x},.)$ of an order k for some $i \in I^*$, then let $y_i(x)$ denote the locally unique solution of $(\frac{\partial}{\partial y})^{k-1} v_i(x,.) = 0$ (with $y_i(\bar{x}) = \bar{y}$). Furthermore, if the functions v_i, $i \in \{0\} \cup I$, or their derivatives are evaluated at $(\bar{x},\bar{y})$, then we omit the argument $(\bar{x},\bar{y})$ in the remainder of this paper. Moreover, all considered coordinate transformations in this subsection will be admissible.

In what follows we will define certain sets M^k, $k = 1,\ldots,11$. These sets partially describe the feasible set M. In fact, they define local normal forms, say, basic sets. Finally, the feasible set will be the intersection of such basic sets (see section 3.2).

Case A. $I^* = \{0\}$ and $\bar{y}$ is a transversal zero-point for $v_0(\bar{x},.)$ of an order k, $k \geq 2$.

Then, the associated set of vectors $S(\bar{y})$ is

$$S(\bar{y}) = \left\{ \left(\frac{\partial}{\partial y}\right)^i v_{0,x},\ i = 0,\ldots,k-2 \right\},$$

and its cardinality $|S(\bar{y})|$ is called the *rank of* $\bar{y}$, i.e., we have $\operatorname{rank}(\bar{y}) = k - 1$. Since $\bar{x} \in \partial M$, $\bar{y}$ is a transversal zero-point for $v_0(\bar{x},.)$ of an even order with $\varepsilon = 1$ (recall that, $\varepsilon := \operatorname{sign}(\frac{\partial}{\partial y})^k v_0$) and, by Lemma 4, there exists an admissible local coordinate transformation sending $(\bar{x},\bar{y})$ to $(0,0)$ such that the local optimal value function takes the form $\eta^{k-1}(x_1,\ldots,x_{k-1})$ (in (3)).

We say that $\bar{y}$ is of *Type* 1 and define the set

$$M^1(x_1,\ldots,x_{k-1}) := \left\{x \in U \mid \eta^{k-1}(x_1,\ldots,x_{k-1}) \geq 0\right\}$$

for an appropriate neighborhood U of $0 \in \mathbb{R}^n$. In particular, if $k = 4$, then the set M^1 is the upper part of the so-called swallow tail (see [8], where a picture is also included).

Case B. $\bar{y}$ is a transversal zero-point for $v_0(\bar{x}, .)$ of an order k, $k \geq 2$; there exists an index $i_0 \in I$ (say $i_0 = 1$) such that $I^* = \{0, 1\}$, and $\bar{y}$ is a transversal zero-point for $v_1(\bar{x}, .)$ of an order one. Furthermore, the Jacobian matrix of the system

$$\left(\frac{\partial}{\partial y}\right)^i v_0 = 0, \quad i = 0, \ldots, k-1, \quad v_1 = 0$$

has rank $k + 1$.

The associated set of vectors $S(\bar{y})$ is

$$S(\bar{y}) = \left\{ \left(\frac{\partial}{\partial y}\right)^{k-1} v_{0,x} \left[\left(\frac{\partial}{\partial y}\right)^k v_0\right]^{-1} - v_{1,x}(v_{1,y})^{-1}, \ \left(\frac{\partial}{\partial y}\right)^i v_{0,x}, \ i = 0, \ldots, k-2 \right\},$$

and we have $\operatorname{rank}(\bar{y}) = k$.

The linear independence of the vectors in $S(\bar{y})$ implies that the following C^∞-coordinate transformation is admissible. First, by Lemma 4, in new coordinates, v_0 takes the form

$$\varepsilon y^k + \sum_{i=1}^{k-1} x_i y^{k-i-1},$$

and we obtain $y_0(x) = 0$ for x near $\bar{x}$. Then, we take $y_1(x)$ $(= y_1(x) - y_0(x))$ as a new x-coordinate, say, x_k.

Note that $\operatorname{sign} v_{1,y}$ is invariant under the former admissible coordinate transformation, since $\frac{\partial}{\partial y}\xi^2(\bar{x}, \bar{y}) > 0$. If $v_{1,y} > 0$, then we say that $\bar{y}$ is of *Type* 2 and define the set

$$M^2(x_1, \ldots, x_k) := \left\{ x \in U \mid \min_{y \in [x_k, 1]} y^k + \sum_{i=1}^{k-1} x_i y^{k-i-1} \geq 0 \right\},$$

where, without loss of generality, we have taken (and will also take in the following cases) the same neighborhood U of $0 \in \mathbb{R}^n$ as for M^1 (see [8] for a picture of M^2). Note that in the latter case we have $\varepsilon = 1$.

If $v_{1,y} < 0$, then we say the $\bar{y}$ is of *Type* 3 and define the set

$$M^3(x_1, \ldots, x_k) := \left\{ x \in U \mid \min_{y \in [-1, x_k]} (-1)^k y^k + \sum_{i=1}^{k-1} x_i y^{k-i-1} \geq 0 \right\}.$$

In the latter case there exist the following two options: Either k is even and $\varepsilon = 1$, or k is odd and $\varepsilon = -1$.

Remark 8. In [8], the index y belongs to a fixed interval Y. In our context, the boundary points of $Y(x)$ may vary as a function of the state variable x. This is the reason why, in the definitions of the sets M^2, M^3, the variable boundary x_k appears.

Case C. $\bar{y}$ is a transversal zero-point for $v_i(\bar{x}, .)$ of an order one, $i \in I^*$, and the Jacobian matrix of the system

$$v_i = 0, \quad i \in I^*$$

has rank $|I^*|$.

We define the sets $I_0 := I^* \cap I$, $I_0^+ := \{i \in I_0 \mid v_{i,y} > 0\}$, and $I_0^- := I_0 \setminus I_0^+$ and consider the following two subcases.

Subcase C1. $0 \notin I^*$.

If $v_0 > 0$, then $v_0(x, y) > 0$ for all x near $\bar{x}$ and all $y \in Y(x)$ locally around $\bar{y}$.

Now, let $v_0 < 0$. If $I_0^+ = \emptyset$ or $I_0^- = \emptyset$, then $\bar{x} \in \text{int } \complement M$, where $\complement M := \mathbb{R}^n \setminus M$. Next, let I_0^+ and I_0^- both be nonempty. Furthermore, let $I_0 =: \{1, \ldots, k\}$, $k \in I_0^+$ and define the associated set of vectors $S(\bar{y})$ as

$$S(\bar{y}) := \left\{ v_{k,x}(v_{k,y})^{-1} - v_{i,x}(v_{i,y})^{-1} \mid i = 1, \ldots, k-1 \right\}.$$

The vectors in $S(\bar{y})$ are linearly independent, and we put $\text{rank}(\bar{y}) = k - 1$. Note that $\bar{x} \notin M$ (since $\bar{y} \in Y(\bar{x})$ and $v_0 < 0$)!

Obviously, if $\bar{x} \in \partial M$, then there exist indices $i \in I_0^-$ and $j \in I_0^+$ with $y_i(x) < y_j(x)$ for some x near $\bar{x}$. After choosing $y_i(x) - y_k(x)$ as new x-coordinates x_i, $i = 1, \ldots, k-1$, we say that $\bar{y}$ is of *Type* 4 and define the set:

$$M^4(x_1, \ldots, x_{k-1}) := \bigcup_{i \in I_0^-} \{x \in U \mid x_i < 0\} \cup \left\{ x \in U \mid \exists\, i \in I_0^-, j \in I_0^+ \setminus \{k\} :\ x_i < x_j \right\}.$$

Note that M^4 (as well as some further sets M^i, $i \geq 5$ to be defined) are related to both of the topological features of M mentioned in section 1: It has a disjunctive structure, and its constraints do not describe a closed set.

Subcase C2. $0 \in I^*$.

If $I_0 = \emptyset$, then we have $\bar{x} \in \text{int } \complement M$. Now, let $I_0 =: \{1, \ldots, k\}$ with $k \geq 1$, define

$$S(\bar{y}) := \left\{ v_{0,x}(v_{0,y})^{-1} - v_{i,x}(v_{i,y})^{-1} \mid i = 1, \ldots, k \right\},$$

and hence, $\text{rank}(\bar{y}) = k$.

First, let $v_{0,y} > 0$ and define the new x-coordinates $x_i := y_i(x) - y_0(x)$, $i = 1, \ldots, k$. Assume for the moment that $I_0^+ \neq \emptyset$ and $I_0^- \neq \emptyset$. Then, locally around $\bar{y}$, the set $Y(\bar{x})$ reduces to the singleton $\{\bar{y}\}$. If there exist indices $i \in I_0^-$ and $j \in I_0^+$ with $x_i < x_j$, then $Y(x)$ becomes empty locally around $\bar{y}$. If $Y(x)$ is not empty around $\bar{y}$, we must have $x_i \geq x_j$ for all $i \in I_0^-$ and $j \in I_0^+$. In the latter case, it is necessary for the feasibility of x to have $\max\{x_j \mid j \in I_0^+\} \geq 0$. Now, we define the set:

$$\begin{aligned} M^5(x_1, \ldots, x_k) := {} & \textstyle\bigcup_{i \in I_0^-,\ j \in I_0^+} \{x \in U \mid x_i < x_j\} \\ & \cup \left\{ x \in U \mid x_j \leq x_i,\ i \in I_0^-,\ j \in I_0^+,\ \max\{x_j \mid j \in I_0^+\} \geq 0 \right\}. \end{aligned}$$

If $I_0^+ = I_0$, the latter set reduces to

$$M^5(x_1, \ldots, x_k) = \left\{ x \in U \mid \max\{x_j \mid j \in I_0^+\} \geq 0 \right\}.$$

The case $I_0^- = I_0$ is not possible, since it would imply $\bar{x} \in \text{int}\, \complement M$. Altogether, if $v_{0,y} > 0$, we say that $\bar{y}$ is of *Type* 5, and we have $\text{rank}(\bar{y}) = k$.

Now, let $v_{0,y} < 0$ and define the new x-coordinates $x_i := y_i(x) - y_0(x)$, $i = 1, \ldots, k$. If $I_0^+ \neq \emptyset$ and $I_0^- \neq \emptyset$, then, by analogous arguments, we define the set:

$$\begin{aligned} M^6(x_1, \ldots, x_k) \quad := \quad & \textstyle\bigcup_{i \in I_0^-,\ j \in I_0^+} \{x \in U \mid x_i < x_j\} \\ & \cup \left\{ x \in U \mid x_j \leq x_i,\ i \in I_0^-,\ j \in I_0^+,\ \min\{x_i \mid i \in I_0^-\} \leq 0 \right\}. \end{aligned}$$

For $I_0^- = I_0$, the latter set reduces to

$$M^6(x_1, \ldots, x_k) = \left\{ x \in U \mid \min\{x_i \mid i \in I_0^-\} \leq 0 \right\}.$$

If $v_{o,y} < 0$, we say that $\bar{y}$ is of *Type* 6, and we have $\text{rank}(\bar{y}) = k$.

Case D. The set I_0 is a singleton, say, $I_0 = \{1\}$, and

- either $I^* = \{1\}$, and $\bar{y}$ is a transversal zero-point for $v_1(\bar{x}, .)$ of an order two;
- or $I^* = \{0, 1\}$, and there exist indices $i_1, i_2 \in I^*$ such that $\bar{y}$ is a transversal zero-point for $v_{i_1}(\bar{x}, .)$ of order two, and a transversal zero-point for $v_{i_2}(\bar{x}, .)$ of an order one or two.

Furthermore, a corresponding rank condition for the Jacobian matrix is satisfied (to be specified).

We consider four subcases.

Subcase D1. $I^* = \{1\}$.

If $v_0 > 0$, then $v_0(x, y) > 0$ for all x near $\bar{x}$ and all $y \in Y(x)$ locally around $\bar{y}$. Now, let $v_0 < 0$ and $v_{1,yy} < 0$ ($v_{1,yy} > 0$ implies $\bar{x} \in \operatorname{int} \complement M$). Note that $\bar{x} \notin M$. We have $S(\bar{y}) = \{v_{1,x}\}$ and, by Lemma 4, in new coordinates the function v_1 takes the form $-y^2 + x_1$. We define the set:

$$M^7(x_1) := \{x \in U \mid x_1 < 0\}.$$

We say that $\bar{y}$ is of *Type* 7, and we have $\operatorname{rank}(\bar{y}) = 1$.

Subcase D2. $I^* = \{0, 1\}$, and $\bar{y}$ is a transversal zero-point for $v_0(\bar{x}, .)$ of an order one, and a transversal zero-point for $v_1(\bar{x}, .)$ of an order two. Furthermore, the Jacobian matrix of the system

$$v_0 = 0, \quad v_1 = 0, \quad v_{1,y} = 0$$

has rank 3.

Then, define

$$S(\bar{y}) := \left\{v_{1,yx}(v_{1,yy})^{-1} - v_{0,x}(v_{0,y})^{-1},\ v_{1,x}\right\}.$$

We have $v_{1,yy} < 0$ ($v_{1,yy} > 0$ would imply $\bar{x} \in \operatorname{int} \complement M$) and, by Lemma 4, in new coordinates, v_1 takes the form $-y^2 + x_1$ (here, $v_1(x, y_1(x))$ is the new coordinate x_1). First, let $v_{0,y} > 0$. We take $y_0(x)$ ($= y_0(x) - y_1(x)$) as a new x-coordinate, say, x_2. Now, if $x_1 < 0$, then $Y(x)$ becomes empty locally around $\bar{y}$. If $x_1 \geq 0$, then for the feasibility of x one needs that $x_2 \leq -\sqrt{x_1}$. Therefore, we define the set:

$$M^8(x_1, x_2) := \{x \in U \mid x_1 < 0\} \cup \{x \in U \mid x_1 \geq 0,\ x_2 \leq -\sqrt{x_1}\}.$$

If $v_{0,y} < 0$, then take $-y_0(x)$ ($= -y_0(x) + y_1(x)$) as the new coordinate x_2, and we obtain the same description M^8. We say that $\bar{y}$ is of *Type* 8, and we have $\operatorname{rank}(\bar{y}) = 2$.

Subcase D3. $I^* = \{0, 1\}$, and $\bar{y}$ is a transversal zero-point for $v_0(\bar{x}, .)$ of an order two, and a transversal zero-point for $v_1(\bar{x}, .)$ of an order one. Furthermore, the Jacobian matrix of the system

$$v_0 = 0, \quad v_{0,y} = 0, \quad v_1 = 0$$

has rank 3. This case is already included in Case B (for $k = 2$).

Subcase D4. $I^* = \{0, 1\}$, and $\bar{y}$ is a transversal zero-point for $v_i(\bar{x}, .)$ of an order two, $i = 0, 1$. The Jacobian matrix of the system

$$v_0 = 0, \quad v_{0,y} = 0, \quad v_1 = 0, \quad v_{1,y} = 0$$

has rank 4.

Then, we define the set:

$$S(\bar{y}) := \left\{v_{0,x},\ v_{1,x},\ v_{0,yx}(v_{0,yy})^{-1} - v_{1,yx}(v_{1,yy})^{-1}\right\}.$$

By Lemma 7, in new coordinates, the functions v_0, v_1 take the form:

$$\begin{aligned} v_0(x,y) &= \pm y^2 \mp x_1, \\ v_1(x,y) &= \alpha(x,y)(y-x_2)^2 + x_3, \end{aligned}$$

where $\alpha \in C^\infty(\mathbb{R}^n \times \mathbb{R}, \mathbb{R})$ with $\alpha(0,0) \neq 0$.

First, let $v_{0,yy} > 0$ and $v_{1,yy} > 0$. For $x_1 \leq 0$ we obtain locally $v_0(x,y) \geq 0$. If $x_1 > 0$, then for the feasibility of x one needs $v_1(x, \sqrt{x_1}) \leq 0$ and $v_1(x, -\sqrt{x_1}) \leq 0$. Then, define

$$\begin{aligned} M^9(x_1,x_2,x_3) := \{x \in U \mid x_1 \leq 0\} \cup \\ \cup \left\{ x \in U \;\middle|\; \begin{array}{r} x_1 > 0,\ \alpha(x,\sqrt{x_1})(\sqrt{x_1}-x_2)^2 + x_3 \leq 0 \\ \alpha(x,-\sqrt{x_1})(-\sqrt{x_1}-x_2)^2 + x_3 \leq 0 \end{array} \right\}. \end{aligned}$$

We say that $\bar{y}$ is of *Type* 9, and we have $\operatorname{rank}(\bar{y}) = 3$.

Now let $v_{0,yy} > 0$ and $v_{1,yy} < 0$. As in the latter case, locally we obtain $v_0(x,y) \geq 0$ for $x_1 \leq 0$. If $x_3 < 0$, then $Y(x)$ is empty locally around $\bar{y}$. If $x_1 > 0$ and $x_3 \geq 0$, then for the feasibility of x one needs

$$\begin{aligned} v_1(x,\sqrt{x_1}) \leq 0 \qquad &\text{and} \quad v_{1,y}(x,\sqrt{x_1}) \geq 0 \quad \text{or} \\ v_1(x,-\sqrt{x_1}) \leq 0 \qquad &\text{and} \quad v_{1,y}(x,-\sqrt{x_1}) \leq 0. \end{aligned}$$

Therefore, define

$$\begin{aligned} M^{10}(x_1,x_2,x_3) := \{x \in U \mid x_1 \leq 0\} \cup \{x \in U \mid x_3 < 0\} \\ \cup \left\{ x \in U \;\middle|\; \begin{array}{l} x_1 > 0,\ x_3 \geq 0,\ \alpha(x,\sqrt{x_1})(\sqrt{x_1}-x_2)^2 + x_3 \leq 0 \\ \dfrac{\partial}{\partial y}\left(\alpha(x,y)(y-x_2)^2 + x_3\right)|_{y=\sqrt{x_1}} \geq 0 \end{array} \right\} \\ \cup \left\{ x \in U \;\middle|\; \begin{array}{l} x_1 > 0,\ x_3 \geq 0,\ \alpha(x,-\sqrt{x_1})(-\sqrt{x_1}-x_2)^2 + x_3 \leq 0 \\ \dfrac{\partial}{\partial y}\left(\alpha(x,y)(y-x_2)^2 + x_3\right)|_{y=-\sqrt{x_1}} \leq 0 \end{array} \right\}. \end{aligned}$$

We say that $\bar{y}$ is of *Type* 10, and we have $\operatorname{rank}(\bar{y}) = 3$.

Now let $v_{0,yy} < 0$ and $v_{1,yy} < 0$. Again, $x_3 < 0$ implies that $Y(x)$ is empty locally around $\bar{y}$. If $x_3 \geq 0$, then for the feasibility of x one needs $x_1 \geq 0$ as well as $v_1(x,\sqrt{x_1}) \leq 0$, $v_{1,y}(x,\sqrt{x_1}) \leq 0$, $v_1(x,-\sqrt{x_1}) \leq 0$, and $v_{1,y}(x,-\sqrt{x_1}) \geq 0$. Define

$$\begin{aligned} M^{11}(x_1,x_2,x_3) := \{x \in U \mid x_3 < 0\} \\ \cup \left\{ x \in U \;\middle|\; \begin{array}{l} x_1 \geq 0,\ x_3 \geq 0,\ \alpha(x,\sqrt{x_1})(\sqrt{x_1}-x_2)^2 + x_3 \leq 0 \\ \alpha(x,-\sqrt{x_1})(-\sqrt{x_1}-x_2)^2 + x_3 \leq 0 \\ \dfrac{\partial}{\partial y}\left(\alpha(x,y)(y-x_2)^2 + x_3\right)|_{y=\sqrt{x_1}} \leq 0 \\ \dfrac{\partial}{\partial y}\left(\alpha(x,y)(y-x_2)^2 + x_3\right)|_{y=-\sqrt{x_1}} \geq 0 \end{array} \right\}. \end{aligned}$$

We say that $\bar{y}$ is of *Type* 11, and we have $\operatorname{rank}(\bar{y}) = 3$.

If $v_{0,yy} < 0$ and $v_{1,yy} > 0$, then $\bar{x} \in \operatorname{int} \complement M$.

Note that the sets M^i, $i = 1, .., 8$ are of cylindric type. They depend only on a subset of the coordinates $x_1, \ldots, x_n$. However, the sets M^9, M^{10}, and M^{11} depend

on the whole vector $x \in \mathbb{R}^n$, since the functions α depend on all coordinates $x_1, ., , x_n$. This completes the analysis of the four cases.

Remark 9. As already mentioned in section 1, there exist some topological phenomena of the feasible set M of a GSIP (nonclosedness, disjunctive structure), which do not appear in standard SIP. The reason for the existence of these phenomena and for the much more complicated analysis of the local structure of M is that the index set $Y(x)$ depends on the state variable x. In standard SIP we have the constant index set $Y(x) = Y$, and a local description of the feasible set takes into account only the zero-points of v_0 (see [8] for a complete discussion of the standard SIP case).

Remark 10. In the discussion of Cases A and B we already referred to the pictures in [8], which are related to the sets M^1 and M^2. In the following we illustrate for some of the subcases of the Cases C and D, where the above-mentioned topological phenomena (nonclosedness, disjunctive structure) of M appear.

Consider Subcase C1 with $I_0 = \{1, 2, 3\}$, $I_0^- = \{1, 2\}$, and $I_0^+ = \{3\}$. Then, the set

$$M^4(x_1, x_2) := \{x \in U \mid x_1 < 0\} \cup \{x \in U \mid x_2 < 0\}$$

is not closed and has a disjunctive structure.

Consider Subcase C2 with $I_0 = I_0^+$ and $v_{0,y} > 0$. Then, the set

$$M^5(x_1, \ldots, x_k) := \{x \in U \mid \max\{x_1, \ldots, x_k\} \geq 0\}$$

has a disjunctive structure (an analogous result holds for M^6).

Consider Subcase D1. Then, the set

$$M^7(x_1) := \{x \in U \mid x_1 < 0\}$$

is not closed.

Consider Subcase D2. Then, the set

$$M^8(x_1, x_2) := \{x \in U \mid x_1 < 0\} \cup \{x \in U \mid x_1 \geq 0,\ x_2 \leq -\sqrt{x_1}\}$$

is not closed.

The sets in Subcase D4 depend on the function $\alpha(x, y)$.

3.2. The decomposition theorem. Throughout this subsection let $\bar{x} \in \partial M$ and assume the following conditions:

(A3) The set $Y_0(\bar{x}) \cup \{y \in Y(\bar{x}) \mid v_0(\bar{x}, y) < 0,\ \exists\, i \in I :\ v_i(\bar{x}, y) = 0\}$ is finite—say $\{y^1, \ldots, y^q\}$—and each y^i, $i \in Q := \{1, \ldots, q\}$ is of one of the Types j, $j = 1, \ldots, 11$.

(A4) For $i \in Q$ let $\operatorname{rank}(y^i) =: r_i$, and let y^i be of Type m_i. For the sets $S(y^i)$ of linearly independent vectors let

$$\dim \operatorname{span} \bigcup_{i \in Q} S(y^i) = \sum_{i \in Q} r_i. \tag{9}$$

We note that (9) is a multitransversality condition; it guarantees the stability of the local description of the feasible set under sufficiently small smooth perturbations of the functions $v_i, i = 0, \ldots, p$. In fact, the conditions describing the Types 1–11 remain fulfilled under small perturbations up to the corresponding order of the zero-point $\bar{y}$ for $v_i(\bar{x}, .)$, $i \in I^*$.

THEOREM 11. *Assume* (A3) *and* (A4). *Let* t_i, $i = 0, \ldots, q$, *recursively be defined as follows:* $t_0 := 0$, $t_i := t_{i-1} + r_i$, $i \in Q$. *Then, there exist an open neighborhood* V

of $\bar{x}$, an open neighborhood U of $0 \in \mathbb{R}^n$, and a local C^∞-coordinate transformation $\phi : V \to U$, sending $\bar{x}$ to $0 \in \mathbb{R}^n$ such that

$$\phi(M \cap V) = \bigcap_{i \in Q} M^{m_i}(x_{t_{i-1}+1}, \ldots, x_{t_i}). \tag{10}$$

Proof. Choose for each $\bar{y} \in Y(\bar{x}) \setminus \{y^i \mid i \in Q\}$ the open neighborhoods $U_{\bar{y}}$ of $\bar{y}$ and $U_{\bar{y}}(\bar{x})$ of $\bar{x}$ such that $v_0(x, y) > 0$ for all $x \in U_{\bar{y}}(\bar{x})$ and all $y \in U_{\bar{y}} \cap Y(x)$. For y^i, $i \in Q$, choose the open neighborhoods U_{y^i} of y^i and $U_{y^i}(\bar{x})$ of $\bar{x}$ such that locally around $(\bar{x}, y^i)$ the local C^∞-coordinate transformation for the Type m_i can be applied for all $(x, y) \in U_{y^i}(\bar{x}) \times U_{y^i}$. Then, select a finite covering of $Y(\bar{x})$ from $\{U_y,\ y \in Y(\bar{x})\}$ which, by (A3), also covers $Y(x)$ for x sufficiently near $\bar{x}$. We obtain a neighborhood V of $\bar{x}$ as the intersection of the finitely many neighborhoods of $\bar{x}$, which appear in this covering. Recall that the sets $S(y^i)$ are closely related to the Jacobian $D_x\xi^1$ of the corresponding local admissible coordinates (ξ^1, ξ^2). Therefore, the multitransversality condition (9) allows the choice of new coordinates $x_1, \ldots, x_{r_1=t_1}, x_{t_1+1}, \ldots, x_{t_2}, x_{t_2+1}, \ldots, x_{t_q}$ simultaneously. After that, the functions α appearing in the sets of Types 9, 10, and 11 take their final form, and we obtain (10). ∎

Acknowledgments. The authors would like to thank the referees for their careful reading of the manuscript and constructive comments. The improved version of the paper has benefited from their valuable suggestions and detailed remarks.

REFERENCES

[1] C. Berge, *Topological Spaces*, Oliver and Boyd, London, 1963.
[2] M. A. Goberna and M. A. López, eds., *Semi-Infinite Programming—Recent Advances*, Kluwer Academic Publishers, Norwell, MA, 2001.
[3] H. Günzel, H. Th. Jongen, and O. Stein, *On the closure of the feasible set in generalized semi-infinite programming*, CEJOR, Cent. Eur. J. Oper. Res., 15 (2007), pp. 271–280.
[4] R. Hettich and G. Still, *Second order optimality conditions for generalized semi-infinite programming problems*, Optimization, 34 (1995), pp. 195–211.
[5] M. W. Hirsch, *Differential Topology*, Springer, New York, 1976.
[6] H. Th. Jongen, J.–J. Rückmann, and O. Stein, *Disjunctive optimization: Critical point theory*, J. Optim. Theory Appl., 93 (1997), pp. 321–336.
[7] H. Th. Jongen, J.–J. Rückmann, and O. Stein, *Generalized semi-infinite optimization: A first order optimality condition and examples*, Math. Program., 83 (1998), pp. 145–158.
[8] H. Th. Jongen and G. Zwier, *On the local structure of the feasible set in semi-infinite optimization*, in Parametric Optimization and Approximation, B. Brosowski and F. and Deutsch, eds., Internat. Ser. Numer. Math. 72, Birkhäuser-Verlag, Basel, Switzerland, 1985, pp. 185–202.
[9] D. Klatte, *Stability of stationary solutions in semi-infinite optimization via the reduction approach*, in Advances in Optimization, W. Oettli and D. Pallaschke, eds., Lecture Notes in Econom. and Math. Systems 382, Springer, New York, 1992, pp. 155–170.
[10] R. Reemtsen and J.–J. Rückmann, eds., *Semi-Infinite Programming*, Kluwer Academic Publishers, Norwell, MA, 1998.
[11] O. Stein, *Bi-level Strategies in Semi-Infinite Programming*, Kluwer Academic Publishers, Norwell, MA, 2003.

SIAM J. OPTIM.
Vol. 19, No. 2, pp. 655–673

FURTHER RELAXATIONS OF THE SEMIDEFINITE PROGRAMMING APPROACH TO SENSOR NETWORK LOCALIZATION*

ZIZHUO WANG†, SONG ZHENG‡, YINYU YE§, AND STEPHEN BOYD¶

Abstract. Recently, a semidefinite programming (SDP) relaxation approach has been proposed to solve the sensor network localization problem. Although it achieves high accuracy in estimating the sensor locations, the speed of the SDP approach is not satisfactory for practical applications. In this paper we propose methods to further relax the SDP relaxation, more precisely, to relax the single semidefinite matrix cone into a set of small-size semidefinite submatrix cones, which we call a sub-SDP (SSDP) approach. We present two such relaxations. Although they are weaker than the original SDP relaxation, they retain the key theoretical property, and numerical experiments show that they are both efficient and accurate. The speed of the SSDP is even faster than that of other approaches based on weaker relaxations. The SSDP approach may also pave a way to efficiently solving general SDP problems without sacrificing the solution quality.

Key words. sensor network localization, semidefinite programming, second-order cone programming, principal submatrix, chordal graph

AMS subject classifications. 90C22, 49M20, 65K05

DOI. 10.1137/060669395

1. Introduction.

There has been an increase in the use of ad hoc wireless sensor networks for monitoring environmental information (e.g., temperature, sound levels, and light) across an entire physical space, where the sensor network localization problem has received considerable attention recently. Typical networks of this type consist of a large number of densely deployed sensor nodes which gather local data and communicate with other nearby nodes. The sensor data from these nodes are relevant only if we know to what location they refer. Therefore knowledge of the node positions becomes imperative. The use of a GPS system could be a very expensive or otherwise impossible approach to this requirement. This problem is also related to other practical distance geometry problems.

The mathematical model of the problem can be described as follows. There are n distinct sensor points in R^d, whose locations are to be determined, and m other fixed points (called the anchor points), whose locations $a_1, a_2, \ldots, a_m$ are known. The Euclidean distance d_{ij} between the ith and jth sensor points is known if $(i,j) \in N_x$, and the distance $\bar{d}_{ik}$ between the ith sensor and kth anchor points is known if $(i,k) \in N_a$. Usually, $N_x = \{(i,j) : \|x_i - x_j\| = d_{ij} \le r_d\}$ and $N_a = \{(i,k) : \|x_i - a_k\| = \bar{d}_{ik} \le r_d\}$, where r_d is a fixed parameter called the radio range. The sensor network

*Received by the editors September 8, 2006; accepted for publication (in revised form) February 13, 2008; published electronically July 2, 2008. Part of this research was done when the first two authors were students and the third author was a visiting professor at the Department of Mathematical Sciences, Tsinghua University, Beijing, China.
http://www.siam.org/journals/siopt/19-2/66939.html

†Department of Management Science and Engineering, Stanford University, Stanford, CA 94305 (zzwang@stanford.edu).

‡Department of Industrial Engineering and Logistics Management, Hong Kong University of Science and Technology, Hong Kong SAR (zhengs@ust.hk).

§Department of Management Science and Engineering and, by courtesy, Electrical Engineering, Stanford University, Stanford, CA 94305 (yinyu-ye@stanford.edu).

¶Department of Electrical Engineering, Stanford University, Stanford, CA 94305 (boyd@stanford.edu).

localization problem is to find $x_i \in R^d$, $i = 1, 2, \ldots, n$, for which

$$\begin{aligned} \|x_i - x_j\|^2 &= d_{ij}^2 \qquad \forall\ (i,j) \in N_x, \\ \|x_i - a_k\|^2 &= \bar{d}_{ik}^2 \qquad \forall\ (i,k) \in N_a. \end{aligned}$$

Unfortunately, this problem is hard to solve in general even for $d = 1$; see, e.g., [15, 35].

For simplicity, we restrict ourselves to $d = 2$ in this paper. Many relaxations have been developed to tackle this and other related problems; see, e.g., [1, 4, 3, 5, 23, 6, 7, 28, 29, 33, 25, 16, 21, 12, 18, 20, 22, 26, 2, 31, 30, 19]. Among them, the work of [1, 4, 3, 5, 23, 16, 20, 19] used a Euclidean distance matrix-based approach, where no anchor was needed or used to compute the unknown portions of the distance matrix [36]; [12, 22] developed a global optimization approach; [21, 30] constructed a second-order cone relaxation; [26, 18] adapted the sum-of-squares (SOS) approach; [33] modeled a problem similar to the dual of the distance completion problem; and [6, 27] considered bounds on the solution rank of a semidefinite programming (SDP) problem. Recently, an SDP relaxation (see, e.g., [7, 28, 29, 25, 2, 31]) which explicitly used the anchors' positions as the first-order information, was applied to solving a class of sensor network localization problems. Their relaxation model can be represented by a standard SDP model

$$\begin{aligned} &\text{minimize} && \mathbf{0} \bullet Z \\ &\text{subject to} && Z_{(1,2)} = I, \\ &&& (\mathbf{0}; e_i - e_j)(\mathbf{0}; e_i - e_j)^T \bullet Z = d_{ij}^2 \quad \forall\ (i,j) \in N_x, \\ &&& (-a_k; e_i)(-a_k; e_i)^T \bullet Z = \bar{d}_{ik}^2 \quad \forall\ (i,k) \in N_a, \\ &&& Z \succeq 0. \end{aligned} \tag{1.1}$$

Here I is the 2-dimensional identity matrix and $Z_{(1,2)}$ is the upper-left 2×2 principal submatrix of Z, $\mathbf{0}$ is a vector or matrix of all zeros, and e_i is the vector of all zeros, except for a one in the ith position. If a solution

$$Z = \begin{pmatrix} I & X \\ X^T & Y \end{pmatrix}$$

to (1.1) is of rank 2, or, equivalently, $Y = X^T X$, then $X = [x_1, \ldots, x_n] \in R^{2 \times n}$ is a solution to the sensor network localization problem. Note that the SDP variable matrix has two parts: the first-order part X (positions) and the second-order part of Y (position inner products). Both parts give valuable information about the estimation and confidence measure of the final localization solution.

As the size of the SDP problem increases, the dimension of the matrix cone increases and the number of variables increases quadratically, no matter how sparse N_x and N_a might be. It is also known that the arithmetic operation complexity of the SDP is at least $O(n^3)$ to obtain an approximate solution. This complexity bound prevents solving large-size problems. Therefore, it would be very beneficial to further relax the full SDP problem by exploiting the sparsity of N_x and N_a at the relaxation modeling level.

Throughout the paper, R^d denotes d-dimensional Euclidean space, S^n denotes the space of $n \times n$ symmetric matrices, and $\mathrm{Rank}(A)$ denotes the rank of A. For $A \in S^n$, A_{ij} denotes the (i,j) entry of A, and $A_{(i_1,\ldots,i_k)}$ denotes the principal submatrix from the rows and columns indexed by $i_1, \ldots, i_k$. For $A, B \in S^n$, $A \succeq B$ means that $A - B$ is positive semidefinite, and $A \bullet B$ denotes the inner product, i.e., $A \bullet B = \mathrm{Tr}(AB)$.

2. Further relaxations of the SDP model. We will give two such relaxations. The first is a node-based relaxation, which we call the NSDP relaxation:

$$\begin{array}{ll} \text{minimize} & \mathbf{0} \bullet Z \\ \text{subject to} & Z_{(1,2)} = I, \\ & (\mathbf{0}; e_i - e_j)(\mathbf{0}; e_i - e_j)^T \bullet Z = d_{ij}^2 \quad \forall\ (i,j) \in N_x, \\ & (-a_k; e_i)(-a_k; e_i)^T \bullet Z = \bar{d}_{ik}^2 \quad \forall\ (i,k) \in N_a, \\ & Z^i = Z_{(1,2,i,N_i)} \succeq 0 \quad \forall i, \end{array} \tag{2.1}$$

where $N_i = \{j : \ (i,j) \in N_x\}$ is the sensor-i-connected point set. Here the single $(2+n)$-dimensional matrix cone is replaced by n smaller $3+|N_i|$-dimensional matrix cones, each of which is a principal submatrix of Z. We should mention that a similar idea was proposed in [24] for solving general SDP problems.

The second relaxation is an edge-based relaxation, which we call the ESDP relaxation:

$$\begin{array}{ll} \text{minimize} & \mathbf{0} \bullet Z \\ \text{subject to} & Z_{(1,2)} = I, \\ & (\mathbf{0}; e_i - e_j)(\mathbf{0}; e_i - e_j)^T \bullet Z = d_{ij}^2 \quad \forall\ (i,j) \in N_x, \\ & (-a_k; e_i)(-a_k; e_i)^T \bullet Z = \bar{d}_{ik}^2 \quad \forall\ (i,k) \in N_a, \\ & Z_{(1,2,i,j)} \succeq 0 \quad \forall (i,j) \in N_x. \end{array} \tag{2.2}$$

Here the single $(2+n)$-dimensional matrix cone is replaced by $|N_x|$ smaller 4-dimensional matrix cones, each of which is a principal submatrix of Z. If a solution

$$Z = \begin{pmatrix} I & X \\ X^T & Y \end{pmatrix}$$

to (2.2) satisfies $\text{Rank}(Z_{(1,2,i,j)} = 2$ for all $(i,j) \in N_x$, then $X = [x_1, \ldots, x_n]$ is a localization for the localization problem. An edge-based decomposition was also used for the SOS approach to localization in [26].

In practice, the distances may be corrupted by random measurement errors. In this case the ESDP model can be adjusted by forming a suitable objective. For example, if there is a random Laplacian noise added to each d_{ij}^2 and $\bar{d}_{ik}^2$, then we solve

$$\begin{array}{ll} \text{minimize} & \displaystyle\sum_{(i,j)\in N_x} |(0, e_i - e_j)(0, e_i - e_j)^T \bullet Z - d_{ij}^2| \\ & \quad + \displaystyle\sum_{(i,k)\in N_a} |(-a_k, e_i)(-a_k, e_i)^T \bullet Z - \bar{d}_{ik}^2| \\ \text{subject to} & Z_{(1,2)} = I, \\ & Z_{(1,2,i,j)} \succeq 0 \quad \forall (i,j) \in N_x, \end{array}$$

which can be written as an SDP:

$$\begin{array}{ll} \text{minimize} & \displaystyle\sum_{(i,j)\in N_x} (u_{ij} + v_{ij}) + \sum_{(i,k)\in N_a} (u_{ik} + v_{ik}) \\ \text{subject to} & Z_{(1,2)} = I, \\ & (\mathbf{0}; e_i - e_j)(\mathbf{0}; e_i - e_j)^T \bullet Z - u_{ij} + v_{ij} = d_{ij}^2 \quad \forall\ (i,j) \in N_x, \\ & (-a_k; e_i)(-a_k; e_i)^T \bullet Z - u_{ik} + v_{ik} = \bar{d}_{ik}^2 \quad \forall\ (i,k) \in N_a, \\ & Z_{(1,2,i,j)} \succeq 0, \quad u_{ij}, v_{ij} \geq 0 \quad \forall (i,j) \in N_x, \\ & u_{ik}, v_{ik} \geq 0 \quad \forall (i,k) \in N_a. \end{array} \tag{2.3}$$

Similarly, NSDP can be reformulated as

$$
\begin{aligned}
\text{minimize} \quad & \sum_{(i,j)\in N_x} (u_{ij}+v_{ij}) + \sum_{(i,k)\in N_a} (u_{ik}+v_{ik}) \\
\text{subject to} \quad & Z_{(1,2)} = I, \\
& (\mathbf{0}; e_i - e_j)(\mathbf{0}; e_i - e_j)^T \bullet Z - u_{ij} + v_{ij} = d_{ij}^2 \quad \forall\ (i,j) \in N_x, \\
& (-a_k; e_i)(-a_k; e_i)^T \bullet Z - u_{ik} + v_{ik} = \bar{d}_{ik}^2 \quad \forall\ (i,k) \in N_a, \\
& Z^i = Z_{(1,2,i,N_i)} \succeq 0 \quad \forall i, \\
& u_{ij}, v_{ij} \geq 0 \quad \forall (i,j) \in N_x, \quad u_{ik}, v_{ik} \geq 0 \quad \forall (i,k) \in N_a.
\end{aligned} \tag{2.4}
$$

For simplicity, we focus on the feasibility models of (1.1), (2.1), and (2.2) in the rest of this paper.

Obviously, (2.1) is a relaxation of (1.1), and (2.2) is a relaxation of (2.1). The following proposition will formalize these relations.

PROPOSITION 2.1. *If*

$$Z_{SDP}^* = \begin{pmatrix} I & X \\ X^T & Y \end{pmatrix}$$

is a solution to (1.1)*, then* Z_{SDP}^**, after removing the unspecified variables, is a solution to relaxation* (2.1)*; if*

$$Z_{NSDP}^* = \begin{pmatrix} I & X \\ X^T & Y \end{pmatrix}$$

is a solution to (2.1)*, then* Z_{NSDP}^**, after removing the unspecified variables, is a solution to relaxation* (2.2)*. Hence*

$$F^{SDP} \subset F^{NSDP} \subset F^{ESDP},$$

where $F^{\cdot}$ *represents the solution set of the corresponding SDP relaxation.*

We notice that (1.1) has $(n+2)^2$ variables and $|N_x| + |N_a|$ equality constraints, (2.1) has at most $4+2n+\sum_i |N_i|^2$ variables and $|N_x|+|N_a|$ equality constraints, and (2.2) has $4+3n+|N_x|$ variables and also $|N_x| + |N_a|$ equality constraints. Usually, $4+3n+|N_x|$ is much smaller than $(n+2)^2$, so that (2.2) has a much smaller number of variables than (1.1); hence, the NSDP or ESDP relaxation has the potential to be solved much faster than (1.1). Our computational results will confirm this fact.

But how good is the NSDP or ESDP relaxation? How do these relaxations perform? In the rest of the paper, we will prove that, although they are weaker than the SDP relaxation, the NSDP and ESDP relaxations share some of the same desired theoretical properties possessed by the full SDP relaxation, including the trace criterion for accuracy. We develop a sufficient condition when NSDP coincides with SDP. We also show that the ESDP relaxation is stronger than the second-order cone programming (SOCP) relaxation. Furthermore, we will present computational results and compare our method with the full SDP, SOS, SOCP relaxation, and domain-decomposition methods. One will see that our method is among the fastest methods, and its localization quality is comparable or superior to that of other methods.

3. Theoretical analyses of NSDP. We make the following basic assumption: G, the undirected graph of a sensor network consisting of all sensors and anchors, with edge sets N_x and N_a, is connected and contains at least three anchors. Before

we present our results, we recall three basic concepts: the d-uniquely localizable graph, the chordal graph, and the partial positive semidefinite matrix.

The definition of a d-uniquely localizable graph is given by [2].

DEFINITION 3.1. *A sensor localization problem is d-uniquely localizable if there is a unique localization $\bar{X} \in R^{d\times n}$ and there is no $x_i \in R^h,\ i = 1,\dots,n$, where $h > d$, such that:*

$$\begin{aligned} &\|x_i - x_j\|^2 = d_{ij}^2 && \forall\ (i,j) \in N_x,\\ &\|(a_k;\mathbf{0}) - x_i\|^2 = \bar{d}_{ik}^2 && \forall\ (i,k) \in N_a,\\ &x_i \neq (\bar{x}_i;\mathbf{0}) && \text{for some } i \in \{1,\dots,n\}. \end{aligned}$$

The latter says that the problem cannot have a nontrivial localization in some higher-dimensional space R^h (i.e., a localization different from the one obtained by simply setting $x_i = (\bar{x}_i;\mathbf{0})$, where anchor points are augmented to $(a_k;\mathbf{0}) \in R^h$).

The condition of a d-unique localizability has been proved to be the necessary and sufficient condition for the SDP relaxation to compute a solution in R^d; see [2]. For the case of $d = 2$, if a graph is 2-uniquely localizable, then the SDP relaxation (1.1) produces a unique solution Z with rank 2, and $X = [x_1,\dots,x_n] \in R^{2\times n}$ of Z is the unique localization of a localization problem in R^2.

DEFINITION 3.2. *An undirected graph is a chordal graph if every cycle of length greater than three has a chord; see, e.g.,* [8].

The chordal graph has been used for solving sparse SDP problems or reducing the number of high-order variables in SOS relaxations; see, e.g., [24, 17, 18].

DEFINITION 3.3. *A square matrix, possibly containing some unspecified entries, is called partial symmetric if whenever the (i,j) entry of the matrix is specified, then so is the (j,i) entry, and the two are equal. A partial semidefinite matrix is a partial symmetric matrix for which every fully specified principal submatrix is positive semidefinite.*

The concept of a partial positive semidefinite matrix can be found, e.g., in [9, 13, 14].

The following result was proved in [9, 13].

LEMMA 3.4. *Every partial positive semidefinite matrix with undirected graph G has positive semidefinite completion if and only if G is chordal.*

Although the NSDP model is weaker than the SDP relaxation in general, the following theorem implies that they are equivalent under the chordal condition.

THEOREM 3.5. *Let the undirected graph of sensor nodes with edge set N_x be chordal. Then*

$$F^{SDP} = F^{NSDP}.$$

Proof. We need only to prove that any solution to (2.1) can be completed to a solution of (1.1). Let

$$Z = \begin{pmatrix} I & X \\ X^T & Y \end{pmatrix}$$

be a solution to (2.1). Then all entries of Z are specified except those Y_{ij} such that $(i,j) \notin N_x$. The conic constraints of (2.1) indicate that every fully specified principal submatrix of Z is positive semidefinite, since it is a principal submatrix of Z^i in (2.1). Thus, Z is a partial semidefinite matrix.

We are also given that the undirected graph induced by Y in Z is chordal. We now prove that the undirected graph induced by Z is also chordal. Notice that the graph of Z has a total of $n+2$ nodes, and every specified entry represents an edge. Let

nodes D_1 and D_2 represent the first two rows (columns) of Z, respectively. Then each of the two nodes has edges to all other nodes in the graph. Now consider any cycle in the graph of Z. If the cycle contains D_1 or D_2 or both, then it must have a chord since each of D_1 and D_2 connect to every other node; if the cycle contains neither D_1 nor D_2, then it still contains a chord since the graph of Y is chordal. Therefore, Z has a positive semidefinite completion, say, $\bar{Z}$, from Lemma 3.4, and $\bar{Z}$ must be a solution to (1.1), since (2.1) and (1.1) share the same constraints involving only the specified entries. □

Under the condition of 2-unique localizability, we further have the following.

COROLLARY 3.6. *If a sensor network is 2-uniquely localizable and its undirected graph of sensor nodes with edge set N_x is chordal, then the solution of* (2.1) *is a unique localization for the sensor network.*

4. Theoretical analyses of ESDP. We now focus on our second relaxation, the ESDP relaxation of (2.2).

4.1. Relation between ESDP and SDP. In the SDP relaxation model, let

$$Z_{SDP} = \begin{pmatrix} I & X \\ X^T & Y \end{pmatrix}$$

be a solution to (1.1). Then it is shown that the individual traces or the diagonal entries of $Y - X^T X$ represent confidence measures in the accuracy of the corresponding sensor's location; see [7, 2]. We will show that the ESDP model retains this very desired property. More precisely, if

$$Z_{ESDP} = \begin{pmatrix} I & X \\ X^T & Y \end{pmatrix}$$

is a solution to (2.2), then the individual traces of $Y - X^T X$ also represent confidence measures in the accuracy of the corresponding sensor's location.

First, we introduce a lemma involving the rank of SDP solutions.

LEMMA 4.1. *Consider the following SDP:*

$$\begin{aligned} \text{minimize} \quad & \sum_i C_i \bullet X_i \\ \text{subject to} \quad & \sum_i A_{ij} \bullet X_i = b_j \quad \forall j, \\ & X_i \succeq 0 \quad \forall i. \end{aligned} \tag{4.1}$$

Then applying the path-following interior-point method will produce a max-rank (relative interior) solution for each X_i, i.e., if X^1 and X^2 are two different optimal solutions satisfying

$$\operatorname{Rank}(X^1_{\bar{i}}) < \operatorname{Rank}(X^2_{\bar{i}}) \quad \textit{for at least one } \bar{i}.$$

Then solving (4.1) *by applying the path-following interior-point method will not yield solution X^1.*

Proof. Problem (4.1) can be reformulated into

$$\begin{aligned} \text{minimize} \quad & \overline{C} \bullet X \\ \text{subject to} \quad & \overline{A}_j \bullet X = b_j \quad \forall j, \\ & X = \begin{pmatrix} X_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & X_n \end{pmatrix} \succeq 0, \end{aligned}$$

where $\overline{C} = \mathrm{diag}(C_i)_{i=1}^n$ and $\overline{A}_j = \mathrm{diag}(A_{ij})_{i=1}^n$. This can also be written as

$$\begin{array}{ll} \text{minimize} & \overline{C} \bullet X \\ \text{subject to} & \overline{A_j} \bullet X = b_j \ \forall j, \\ & E_{ij} \bullet X = 0 \ \forall (i,j) \notin D, \\ & X \succeq 0, \end{array}$$

where D denotes those positions that do not belong to any diagonal block of X.

Thus, the path-following algorithm will return a max-rank solution to the problem; see, e.g., [10, 11]. In other words, if X^* is a solution calculated by the path-following method, then $\sum_{i=1}^n \mathrm{Rank}(X_i^*)$ is maximal among all solutions; hence, for every i, $\mathrm{Rank}(X_i^*)$ must be maximal among all solutions to (4.1). Thus, X^1 cannot be a solution generated by the interior-point method. □

By applying this lemma, we have the following result which provides a justification for using the individual traces to measure the accuracy of computed sensor locations.

THEOREM 4.2. *Let*

$$Z = \begin{pmatrix} I & X \\ X^T & Y \end{pmatrix}$$

be a max-rank solution of (2.2). *If the diagonal entry or individual trace*

$$(Y - X^T X)_{\bar{i}\bar{i}} = 0, \tag{4.2}$$

then the $\bar{i}$th column of X, $x_{\bar{i}}$, must be the true location of the $\bar{i}$th sensor, and $x_{\bar{i}}$ is invariant over all solutions Z for (2.2).

Proof. Our proof is by contradiction. Without losing generality, we assume that $(Y - X^T X)_{jj} > 0$ for all $j \neq \bar{i}$.

Note that the constraints in (2.2) ensured that $Z_{(1,2,\bar{i},j)} \succeq 0$ for all $(\bar{i}, j) \in N_x$. Thus, $(Y - X^T X)_{\bar{i}\bar{i}} = 0$ implies that $(Y - X^T X)_{\bar{i}j} = 0$ for all $(\bar{i}, j) \in N_x$, i.e., $Z_{(1,2,\bar{i},j)}$ has rank 3 for all $(\bar{i}, j) \in N_x$. Moreover, from Lemma 4.1, the max-rank of $Z_{(1,2,\bar{i},j)}$ is at most 3 for all solutions to (2.2).

Denote by $\bar{Z}$ a true localization for (2.2), that is, $\bar{Z}_{(1,2,i,j)}$ has rank 2 for all $(i,j) \in N_x$, where

$$\bar{Z}_{(1,2,i,j)} = \begin{pmatrix} I & \bar{x}_i & \bar{x}_j \\ \bar{x}_i^T & \bar{Y}_{ii} & \bar{Y}_{ij} \\ \bar{x}_j^T & \bar{Y}_{ji} & \bar{Y}_{jj} \end{pmatrix} = \begin{pmatrix} I & \bar{x}_i & \bar{x}_j \\ \bar{x}_i^T & \|\bar{x}_i\|^2 & \bar{x}_i^T \bar{x}_j \\ \bar{x}_j^T & \bar{x}_j^T \bar{x}_i & \|\bar{x}_j\|^2 \end{pmatrix}.$$

Suppose that $\bar{x}_{\bar{i}} \neq x_{\bar{i}}$. Since the solution set is convex, then

$$Z^\alpha = \alpha \bar{Z} + (1 - \alpha) Z, \quad 0 \leq \alpha \leq 1,$$

is also a solution to (2.2). By taking α sufficiently small but strictly positive, we will get another solution Z^α which satisfies

$$\mathrm{Rank}(Z^\alpha_{(1,2,i,j)}) \geq \mathrm{Rank}(Z_{(1,2,i,j)}) \ \forall (i,j) \in N_x,$$

and the *strict inequality* holds for $i = \bar{i}$. This is because for $(\bar{i}, j) \in N_x$

$$\begin{aligned} & Y^\alpha_{(\bar{i},j)} - [x^\alpha_{\bar{i}}, x^\alpha_j]^T [x^\alpha_{\bar{i}}, x^\alpha_j] \\ &= \alpha \bar{Y}_{(\bar{i},j)} + (1-\alpha) Y_{(\bar{i},j)} - (\alpha[\bar{x}_{\bar{i}}, \bar{x}_j] + (1-\alpha)[x_{\bar{i}}, x_j])^T (\alpha[\bar{x}_{\bar{i}}, \bar{x}_j] + (1-\alpha)[x_{\bar{i}}, x_j]) \\ &= (1-\alpha)(Y_{(\bar{i},j)} - [x_{\bar{i}}, x_j]^T [x_{\bar{i}}, x_j]) + \alpha(1-\alpha)([x_{\bar{i}}, x_j] - [\bar{x}_{\bar{i}}, \bar{x}_j])^T ([x_{\bar{i}}, x_j] - [\bar{x}_{\bar{i}}, \bar{x}_j]). \end{aligned}$$

Since $(Y - X^T X)_{\bar{i}\bar{i}} = (Y - X^T X)_{\bar{i}j} = (Y - X^T X)_{j\bar{i}} = 0$,

$$Y_{(\bar{i},j)} - [x_{\bar{i}}, x_j]^T [x_{\bar{i}}, x_j] = \begin{pmatrix} 0 & 0 \\ 0 & \gamma \end{pmatrix}$$

for some $\gamma > 0$.

Also we are given that $\bar{x}_{\bar{i}} \neq x_{\bar{i}}$, so that $([x_{\bar{i}}, x_j] - [\bar{x}_{\bar{i}}, \bar{x}_j])^T([x_{\bar{i}}, x_j] - [\bar{x}_{\bar{i}}, \bar{x}_j])$ is a positive semidefinite matrix whose first element is positive, which implies that

$$\det\left[(1-\alpha)\begin{pmatrix} 0 & 0 \\ 0 & \gamma \end{pmatrix} + \alpha(1-\alpha)([x_{\bar{i}}, x_j] - [\bar{x}_{\bar{i}}, \bar{x}_j])^T([x_{\bar{i}}, x_j] - [\bar{x}_{\bar{i}}, \bar{x}_j])\right] > 0.$$

That is, $Z^{\alpha}_{(1,2,\bar{i},j)}$ is a solution to (2.2) with rank 4, which is a contradiction.

Therefore, we proved that $\bar{x}_{\bar{i}}$ must be the true location of the $\bar{i}$th sensor and $\bar{x}_{\bar{i}}$ is invariant over all solutions to (2.2). □

Theorem 4.2 is related to Proposition 2 of [30]. Moreover, the desired invariance property of $x_{\bar{i}}$ extends to the case with noises, which can also be seen from the proof in [30]. In summary, we have the following.

COROLLARY 4.3. *Let*

$$Z = \begin{pmatrix} I & X \\ X^T & Y \end{pmatrix}$$

be a solution to (2.2) *and condition* (4.2) *hold for all* i. *Then the ESDP model* (2.2) *produces a unique solution for the sensor network in* R^2.

Next we enhance Proposition 2.1 by the following theorem.

THEOREM 4.4. *Let*

$$Z = \begin{pmatrix} I & X \\ X^T & Y \end{pmatrix}$$

be a solution to (2.2), *and let*

$$\bar{Z} = \begin{pmatrix} I & \bar{X} \\ \bar{X}^T & \bar{Y} \end{pmatrix}$$

be any solution to (1.1); *both are calculated by the path-following method. If condition* (4.2) *holds for* Z, *so it does for* $\bar{Z}$.

Proof. Our proof is again by contradiction. If (4.2) holds for Z but not for $\bar{Z}$, e.g., $(\bar{Y} - \bar{X}^T \bar{X})_{ii} > 0$. Since, for $0 \leq \alpha \leq 1$,

$$Z_\alpha = (1-\alpha)Z + \alpha\bar{Z}$$

is always a solution to (2.2), by taking α sufficiently small, we will get a solution with a higher rank than Z, and this fact contradicts Lemma 4.1. □

Theorem 4.4 says that if the ESDP relaxation can accurately locate a certain sensor, so can the SDP relaxation. This implies that the ESDP relaxation is weaker than the SDP relaxation. We illustrate this by using an example.

Example 1. Consider the following graph with 3 sensors and 3 anchors. The 3 anchors are located at $(-0.4, 0)$, $(0.4, 0)$, and $(0, 0.4)$, and the 3 sensors are located at $(-0.05, 0.3)$, $(-0.08, 0.2)$, and $(0.2, 0.3)$, respectively. We set the radio range to be 0.50; see Figure 4.1.

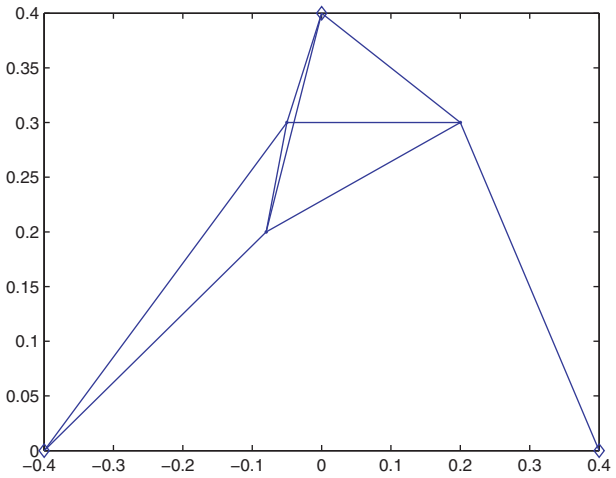


FIG. 4.1. *The locations of sensors, anchors, and connection edges in Example* 1.

In Figure 4.1 (and throughout this paper), we use diamonds to represent the anchor positions. We use a solid line to connect two points (sensors and/or anchors) when their Euclidean distance is smaller than the radio range, so that the length of the line segment is known.

First, we use full SDP relaxation (1.1) to solve this sensor localization problem, where the result is accurate (see Figure 4.2(a)). In Figure 4.2(a) (and throughout this paper), a circle denotes the true location of a sensor (they are not known to the SDP models), and a star denotes the location of a sensor computed by the SDP model. If we use the quantity of the root mean square deviance (RMSD) to measure the deviance of the computed result:

$$RMSD = \left(\frac{1}{n} \sum_{i=1}^{n} \|x_i - \bar{x}_i\|_2^2 \right)^{\frac{1}{2}}, \tag{4.3}$$

where x_i is the position vector of sensor i computed by the algorithm and $\bar{x}_i$ is its true position vector, then the RMSD of the full SDP localization is about $1e-7$. Note that the NSDP model (2.1) returns the exactly same localization of the full SDP from Theorem 3.5, since N_x is a chordal graph.

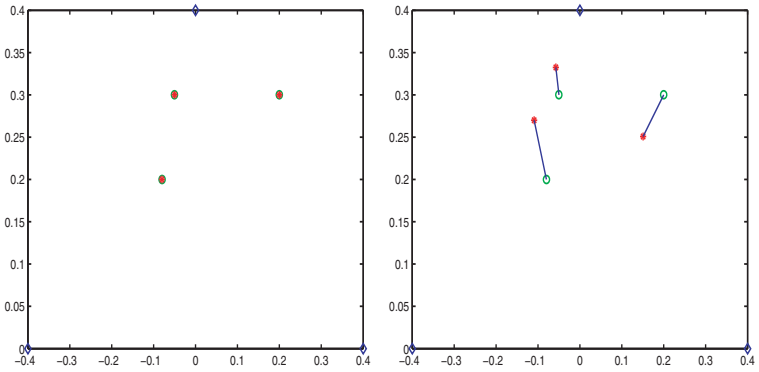
Next we use the ESDP model (2.2) to solve the problem, and this time the result is inaccurate with the RMSD at 0.048; see Figure 4.2(b), where every true sensor location and its computed corresponding position are connected by a solid line.

Now we illustrate why this error happened. In SDP model (1.1), the solution matrix Z^* is required to be positive semidefinite. If we write

$$Z^*_{SDP} = \begin{pmatrix} I & X \\ X^T & Y \end{pmatrix},$$

then the matrix $Y - X^T X$ is required to be positive semidefinite. But in model (2.2),

$$Z^*_{ESDP} = \begin{pmatrix} I & \bar{X} \\ \bar{X}^T & \bar{Y} \end{pmatrix},$$



(a) Graphical localization result of the SDP model

(b) Graphical localization result of the ESDP model

FIG. 4.2. *Comparision of graphical localization results generated by the SDP and ESDP in Example* 1.

where we just require that each 2×2 principal submatrix of $\bar{Y} - \bar{X}^T\bar{X}$ be positive semidefinite. This does not imply that the entire matrix is positive semidefinite. In fact, the solution calculated by the ESDP model (2.2) is

$$Z^*_{ESDP} = \begin{pmatrix} 1 & 0 & -0.07278 & -0.13467 & 0.14884 \\ 0 & 1 & 0.32778 & 0.25467 & 0.24884 \\ -0.07278 & 0.32778 & 0.11072 & 0.09498 & 0.06865 \\ -0.13467 & 0.25467 & 0.09498 & 0.09014 & 0.04540 \\ 0.14884 & 0.24884 & 0.06865 & 0.04540 & 0.08907 \end{pmatrix}.$$

It can be verified that Z^*_{ESDP} satisfies all constraints in (2.2) as well as in (1.1), and each 2×2 principal matrix of $\bar{Y} - \bar{X}^T\bar{X}$ is positive semidefinite. But the three eigenvalues of $\bar{Y} - \bar{X}^T\bar{X}$ are $(-0.00048, 0.0048, 0.0091)$, so that the entire matrix of $\bar{Y} - \bar{X}^T\bar{X}$ is indefinite, and this is the cause of the difference between the two relaxations.

4.2. Relation between ESDP and SOCP. A SOCP relaxation for the sensor network localization problem has been proposed (see, e.g., [21, 30]):

$$\begin{aligned} \text{minimize} \quad & \sum_{(i,j)\in N_x} (u_{ij} + v_{ij}) + \sum_{(i,k)\in N_a} (u_{ik} + v_{ik}) \\ \text{subject to} \quad & x_i - x_j - w_{ij} = 0 \quad \forall (i,j) \in N_x, \quad x_i - a_k - w_{ik} = 0 \quad \forall (i,k) \in N_a, \\ & y_{ij} - u_{ij} + v_{ij} = d_{ij}^2 \ \forall (i,j) \in N_x, y_{ik} - u_{ik} + v_{ik} = \bar{d}_{ik}^2 \ \forall (i,k) \in N_a, \\ & u_{ij} \geq 0, v_{ij} \geq 0, (y_{ij} + \tfrac{1}{4}, y_{ij} - \tfrac{1}{4}, w_{ij}) \in SOC \ \forall (i,j) \in N_x, \\ & u_{ik} \geq 0, v_{ik} \geq 0, (y_{ik} + \tfrac{1}{4}, y_{ik} - \tfrac{1}{4}, w_{ik}) \in SOC \ \forall (i,k) \in N_a. \end{aligned} \tag{4.4}$$

The SOCP relaxation can be also viewed as a further relaxation of the SDP relaxation, and it was proved to be faster than the SDP method and to serve as a useful preprocessor of the actual problem. In this section, we will show that the ESDP model is stronger than the SOCP relaxation. Our proof refers to Proposition 1 of [30].

THEOREM 4.5. *If*

$$Z = \begin{pmatrix} I & X \\ X^T & Y \end{pmatrix}$$

is an optimal solution to (2.3)*, then the ith column of* X*,* x_i*,* $i = 1, \ldots, n$*, and*

$$y_{ij} = \begin{cases} Y_{ii} + Y_{jj} - 2Y_{ij}, & (i,j) \in N_x, \\ \|a_k^2\| - 2a_k^T x_i + Y_{ii}, & (i,k) \in N_a \end{cases}$$

form a feasible solution for (4.4) *with the same objective value.*

Proof. Since Z is a feasible solution to (2.3), we have $Z_{(1,2,i,j),(1,2,i,j)} \succeq 0$ for all $(i,j) \in N_x$. So, for each $(i,j) \in N_x$, we have

$$\begin{pmatrix} Y_{ii} - \|x_i^2\| & Y_{ij} - x_i^T x_j \\ Y_{ij} - x_i^T x_j & Y_{jj} - \|x_j^2\| \end{pmatrix} \succeq 0.$$

This implies that $Y_{ii} - \|x_i^2\| \geq 0$, $Y_{jj} - \|x_j^2\| \geq 0$, and $(Y_{ii} - \|x_i^2\|)(Y_{jj} - \|x_j^2\|) \geq (Y_{ij} - x_i^T x_j)^2$.

Hence $(Y_{ii} - \|x_i^2\| + Y_{jj} - \|x_j^2\|)^2 \geq 4(Y_{ij} - x_i^T x_j)^2$, i.e.,

$$Y_{ii} + Y_{jj} - 2Y_{ij} \geq \|x_i^2\| + \|x_j^2\| - 2x_i^T x_j,$$

and the theorem follows. □

COROLLARY 4.6. *If* x_i *is invariant over all of the solutions of* (4.4)*, then it is also invariant over all of the ESDP solutions. That is, if SOCP relaxation can return the true location for a sensor, so can ESDP relaxation.*

The above theorem and corollary indicate that one can always derive the same SOCP relaxation solution from an ESDP relaxation solution; that is, the solution set of the ESDP relaxation is smaller than that of the SOCP relaxation. Thus, the ESDP relaxation is stronger than the SOCP relaxation. The following example shows that the reverse is not true.

Example 2. Consider the following problem with 3 anchors and 2 sensors. The true locations of 3 anchors are $a_1 = (-0.4, 0), a_2 = (0, 0.5)$, and $a_3 = (0.4, 0)$, and the true locations of the 2 sensors are $x_1 = (0, -0.3)$ and $x_2 = (0.4, 0.2)$ with radio range 0.7 (see Figure 4.3).

Since there are only two sensors, the ESDP relaxation is the same with the full SDP relaxation, and it is known that this graph is strongly localizable (see [2]), so we know that the ESDP relaxation will give the *unique* solution Z where X is the accurate positions of the sensors. However, for SOCP relaxation, since the graph is 2-realizable, its optimal value of (4.4) is 0 so that the optimal solution must satisfy $y_{ij} = d_{ij}^2$ and $y_{ik} = \bar{d}_{ik}^2$. Thus, any $(\bar{x}_1, \bar{x}_2)$ that satisfies

$$\begin{aligned} \|\bar{x}_1 - \bar{x}_2\|^2 &\leq 0.4^2 + 0.5^2 = 0.41, \\ \|\bar{x}_1 - a_1\|^2 &\leq 0.3^2 + 0.4^2 = 0.25, \\ \|\bar{x}_1 - a_3\|^2 &\leq 0.3^2 + 0.4^2 = 0.25, \\ \|\bar{x}_2 - a_2\|^2 &\leq 0.4^2 + 0.3^2 = 0.25, \\ \|\bar{x}_2 - a_3\|^2 &\leq 0^2 + 0.2^2 = 0.04 \end{aligned}$$

must be also optimal to (4.4).

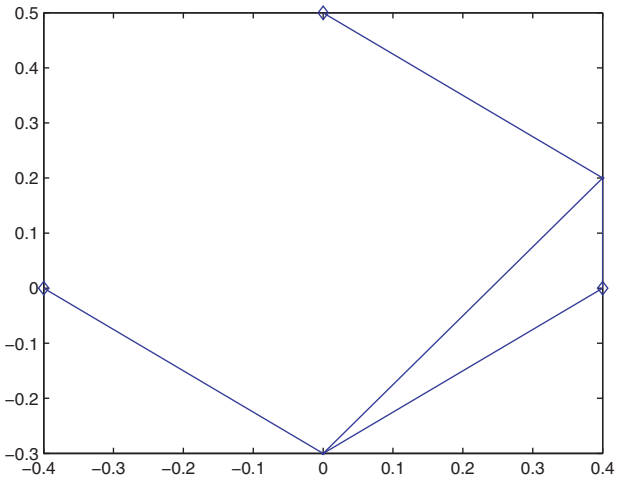


FIG. 4.3. *The locations of sensors, anchors, and connection edges in Example* 2.

Now let $\bar{x}_1 = (0, 0) \neq x_1$ and $\bar{x}_2 = (0.3, 0.15) \neq x_2$. Then it is easy to verify that the above inequalities hold, so that $(\bar{x}_1, \bar{x}_2)$ is also an optimal solution to (4.4). But we know that the interior-point method would always maximize the potential function (see [11, 10])

$$P(x, y) = \sum_{(i,j)\in N_x} \log\left(y_{ij} - \|x_i - x_j\|^2\right) + \sum_{(i,k)\in N_a} \log\left(y_{ik} - \|x_i - a_k\|^2\right)$$

in the optimal solution set; and it is obvious that $P(\bar{x}_1, \bar{x}_2) > P(x_1, x_2)$. Therefore the SOCP relaxation model (4.4) will not give the true solution x_1 and x_2, and, thereby, the ESDP relaxation is *strictly* stronger than the SOCP relaxation for this example.

4.3. The dual problem of ESDP. For a conic programming problem, it is important to consider its dual problem. In many cases, the dual problem can give much important information about the primal problem as well as many useful applications. Here we will present the dual problem of (2.2) and list some basic properties between the primal and dual.

Consider a general conic programming problem:

$$\begin{array}{ll} \text{minimize} & C \bullet X \\ \text{subject to} & A_j \bullet X = b_j \quad \forall j, \\ & X_{(N_i)} \succeq 0 \quad \forall i, \end{array} \tag{4.5}$$

where $X \in S^n$ and N_i is an index subset of $\{1, 2, \ldots, n\}$. Then the dual to the problem is

$$\begin{array}{ll} \text{maximize} & \displaystyle\sum_j b_j y_j \\ \text{subject to} & \displaystyle\sum_j y_j A_j + \sum_i S^i = C, \\ & S^i_{(N_i)} \succeq 0, \text{ and } S^i_{kj} = 0 \quad \forall k \notin N_i \text{ or } j \notin N_i; \quad \forall i. \end{array} \tag{4.6}$$

In other words, S^i is an S^n matrix, and its entries are zero outside the principal submatrix of S_{N_i,N_i}.

For the ESDP model (2.2), the dual problem is

$$
\begin{aligned}
\text{maximize} \quad & \sum_{(i,j)\in N_x} \omega_{ij} d_{ij}^2 + \sum_{(i,k)\in N_a} \omega_{ik} \bar{d}_{ik}^2 + u_{11} + 2u_{12} + u_{22} \\
\text{subject to} \quad & \sum_{(i,j)\in N_x} \omega_{ij} (\mathbf{0}; e_i - e_j)^T (\mathbf{0}; e_i - e_j) + \sum_{(i,k)\in N_a} \omega_{ik} (-a_k; e_i)^T (-a_k; e_i) \\
& + \begin{pmatrix} u_{11} + u_{12} & u_{12} & \mathbf{0} \\ u_{12} & u_{22} + u_{12} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} + \sum_{(i,j)\in N_x} S^{(i,j)} = \mathbf{0}, \\
& S^{(i,j)}_{(1,2,i,j)} \succeq 0, \text{ and } S^{(i,j)}_{kl} = 0 \ \forall k \notin \{i,j\} \text{ or } \ l \notin \{i,j\}, \ \forall (i,j) \in N_x.
\end{aligned} \tag{4.7}
$$

We have the following complementarity result.

PROPOSITION 4.7. *Let Z be a solution to* (2.2) *and $\{S^{(i,j)}\}$ be an optimal solution to the dual. Then*

$$S^{(i,j)}_{(1,2,i,j)} \bullet Z_{(1,2,i,j)} = 0 \ \forall (i,j) \in N_x.$$

In particular, if $\text{Rank}(S^{(i,j)}_{(1,2,i,j)})$ *is* 2 *for all* $(i,j) \in N_x$, *then* $\text{Rank}\left(Z_{(1,2,i,j)}\right)$ *is* 2 *for all* $(i,j) \in N_x$ *so that* (2.2) *produces a unique localization for the sensor network in* R^2.

By using duality we can solve the dual problem and simultaneously yield a primal solution from the complementarity proposition. We demonstrate in the next section that the solution speed of solving the dual is about twice as fast as solving the primal problem, which was originally observed in [34].

5. Computational results and comparison to other approaches. Now we address the question: Will the improvement in the speed of the ESDP relaxation compensate the loss in relaxation quality? In this section, we first present some computational results of the ESDP relaxation model. Then we compare the model with different kinds of approaches, including the full SDP approach (1.1) of [7], the SOCP approach [30], the SOS approach [26], and the domain-decomposition approach of [25, 29].

5.1. Computational results of the ESDP relaxation. In our numerical simulation, we follow [7]. We randomly generate the true positions of n points in a square of 1 by 1, then randomly select m points to be anchors, and compute every edge length $\bar{d}_{ij}$. We select only those edges whose edge length is less than the given radio range rd and add a multiplicative random noise to every selected edge length,

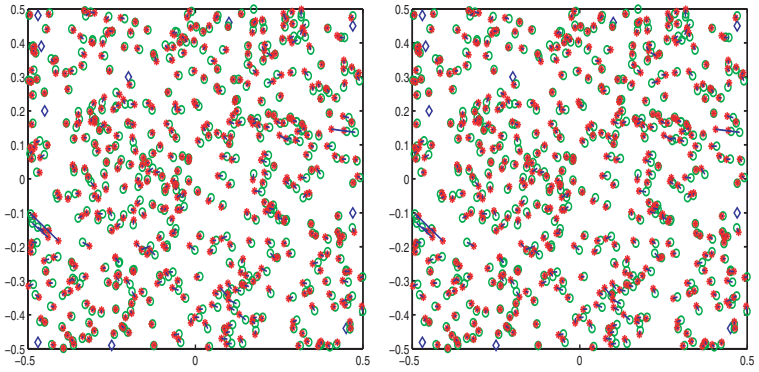
$$d_{ij} = \bar{d}_{ij}(1 + nf \cdot randn(1)),$$

as the distance input data to the SDP models. Here nf is a specified noisy factor, and $randn(1)$ is a standard Gaussian random variable. There may still be many points within the radio range for a sensor or anchor. Thus, in order to maintain the sparsity of the graph, we set a limit 7 on the number of selected edges connected to every sensor or anchor, and they are *randomly* chosen.

In our computational experiments we also implement the steepest-descent local search refinement proposed in [28, 29] for solving noisy problems. All test problems are solved by SeDuMi 1.05 [32] of Matlab7.0 on a DELL D420 laptop with 1.99 GB memory and 1.06 GHz CPU.

TABLE 5.1
Noisy test problems and the SDP solution time comparison.

Noisy problem #	n	m	rd	Full SDP time	ESDP time	Dual ESDP time
1	50	5	0.35	1.33	1.5	1.22
2	100	5	0.3	4.94	3.22	1.91
3	200	5	0.25	35.21	7.64	4.19
4	400	10	0.2	358.8	18.2	8.98
5	800	20	0.12	*	44.67	18.58
6	1600	40	0.07	*	120.58	43.91
7	3200	80	0.04	*	287.39	104.36
8	5000	100	0.03	*	426.85	192.08
9	6400	160	0.025	*	603.16	250.97



(a) Graphic result of the full SDP model

(b) Graphic result of the dual ESDP model

FIG. 5.1. *Comparision of graphical localization results generated by the full SDP and dual ESDP on a* 10% *noisy problem.*

The first set of test problems has noisy factor $nf = 0.1$ throughout. Table 5.1 contains a computational comparison of ESDP to the full SDP relaxation [7]. Here three models, the full SDP model (up to 400 points), the ESDP model, and the dual of the ESDP model, are all solved by SeDuMi 1.05. In order to see the efficiency of the ESDP model itself, the solution time (in seconds) in Table 5.1 includes only the SeDuMi solver time; that is, the data input/preparation time is excluded.

As we can see, while the full SDP solution time increases cubically in size, the SDP solver times of both ESDP and dual ESDP increase little faster than *linearity*. While this speedup was remarkable, how about the localization quality? Figure 5.1 shows two graphical results generated by full SDP and dual ESDP on solving a smaller problem, where one can barely see much difference. Here diamonds represent the anchor positions, circles represent sensor's true positions, and stars represent the computed sensor positions. (The codes and a few test problems have been placed on the public site [37]. We welcome the reader to test them and draw their own conclusions.)

Next we compare our approach to the SOS approach, the SOCP approach, and the domain-decomposition approach. We will use the same examples presented in these papers.

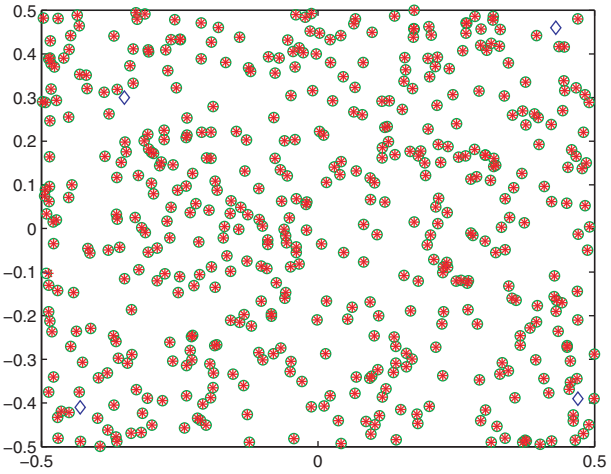


FIG. 5.2. *Graphical localization result of the ESDP model on the problem of Nie* [26]*, 500 sensors, 4 anchors,* $rd = 0.3$*,* $nf = 0$*, and RMSD* $= 1e-6$.

5.2. Computational comparison with the SOS method. The SOS method is an SDP relaxation which applies to solving the problem

$$\min \quad f(x) = \sum_{(i,j)\in N_x} (\|x_i - x_j\|_2^2 - d_{ij}^2)^2 + \sum_{(i,k)\in N_a} (\|x_i - a_k\|_2^2 - \bar{d}_{ik}^2)^2, \tag{5.1}$$

where the objective function is a polynomial.

Recent study [26] has shown that by exploiting the sparsity in SOS relaxation one can get faster computing speed than the SDP relaxation (1.1) and sometimes higher accuracy as well. The author demonstrated that this structure can help save computation time significantly. In [26], the author used the model of 500 sensors and 4 anchors with a radio range of 0.3 and no noises in distance measurements.

The author of [26] reported that it took totally about 1 hour and 25 minutes on a 0.98 GB RAM and 1.46 GHz CPU computer to get a result with RMSD$=2.9e-6$. However, with the same parameters, our approach needs only 30 seconds to get the result with RMSD$=1e-6$. Thus, the ESDP approach is much faster than the SOS approach in this case, and the solution quality is comparable to that of the SOS method; see Figure 5.2.

5.3. Computational comparison with the SOCP method. The SOCP model performs best with a large fraction of anchors and a low noise. Thus, we test (primal) ESDP on the same set of problems reported in [30], where $m = 0.1n$ (10% of points are anchors) and $nf \le 0.01$ (less than 1% noise), and the results are shown in Table 5.2. To solve the SOCP relaxation model, two methods are proposed in [30]: one directly uses Matlab SeDuMi, and the other uses a smoothing coordinate gradient descent (SCGD) method coded in FORTRAN 77. The latter is highly parallelizable, similar to the distributed methods of [25, 29].

From Table 5.2, we see that the ESDP approach is much faster than the SOCP approach when both use Matlab SeDuMi, and it is slower than the tailored and FORTRAN-coded SCGD method. On the other hand, the localization quality (see RMSD in Table 5.3) of ESDP is much better than that reported in [30] for both SeDuMi of SOCP and SCGD of SOCP. Figure 5.3 shows the graphical result of test

TABLE 5.2
*ESDP times are taken on DELL D*420 *(*1.99 *GB and* 1.06 *GHz), and SOCP times are reported from* [30] *on a HP DL*360 *(*1 *G memory and* 3 *GHz).*

Test problem #	n	nf	rd	ESDP time	SeDuMi of SOCP	SCGD of SOCP
1	1000	0	0.06	59.60 sec	3.6 min	0.2 min
2	1000	0.001	0.06	57.55 sec	3.2 min	0.4 min
3	1000	0.01	0.06	53.60 sec	3.9 min	1.6 min
4	4000	0	0.035	653.7 sec	202.5 min	1.6 min
5	4000	0.001	0.035	668.3 sec	193.8 min	5.1 min
6	4000	0.01	0.035	615.9 sec	196.3 min	6.2 min

TABLE 5.3
Input parameters for the test problems, the corresponding ESDP dimensions, and ESDP computational results.

Test problem #	n	nf	rd	SeDuMi SDP dim	CPU time	obj	RMSD
1	1000	0	0.06	20321×29195	59.60	3e-3	2e-3
2	1000	0.001	0.06	20321×29195	57.55	5e-4	3e-3
3	1000	0.01	0.06	20321×29195	53.66	4e-2	2e-2
4	4000	0	0.035	93727×133285	653.7	3e-3	1e-3
5	4000	0.001	0.035	93727×133285	668.3	7e-3	8e-4
6	4000	0.01	0.035	93727×133285	615.9	2e-2	3e-2

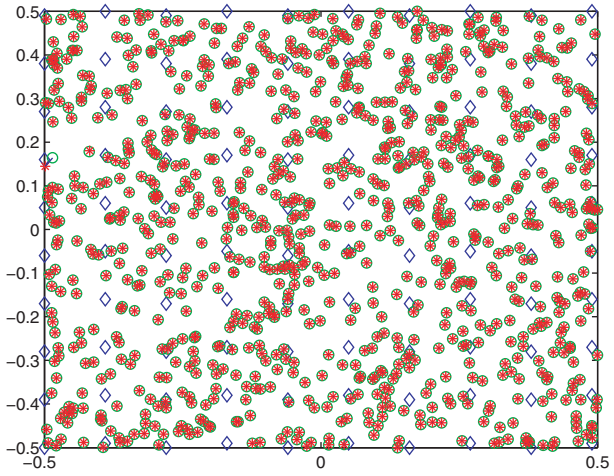


FIG. 5.3. *Graphical localization result of the ESDP model on test problem* 2 *in Table* 5.2.

problem 2 (900 sensors, 100 anchors, $nf = 0.001$, and $rd = 0.06$), where the localization of ESDP is quite accurate compared with the graphical result on the same problem reported in [30].

In Table 5.2, "ESDP time" denotes the *total* solution running time, including Matlab data preparation and SeDuMi input setup time. By comparing Tables 5.1 and 5.2, one can see that, for ESDP, the Matlab data input and SeDuMi setup time is considerable. This is because Matlab is notoriously slow on matrix loops and data inputs. This problem should go away when the algorithm is coded in C or FORTRAN 77.

Table 5.3 contains more detailed statistical results on this test, where "SeDuMi SDP dim" represents problem dimensions solved by SeDuMi, "CPU time" denotes

the total ESDP solution time in seconds (including Matlab data preparation and SeDuMi input setup time), "obj" denotes the SDP objective value, and RMSD is the localization quality defined by (4.3).

5.4. Computational comparison with the decomposition method. There are other earlier approaches to speed up the SDP solution time. The domain-decomposition method of [29] and SpaceLoc of [25] are both based on breaking the localization problem into many geographically partitioned and smaller-sized localization problems, since each smaller SDP problem can be solved much faster and more accurately. Thus, they work quite well when many anchors are uniformly distributed in the region so that one is able to partition the network into many smaller domains; and, as a result, each of them contains enough anchors and forms its own independent localization problem. However, when the quantity of anchors is small or most of them are located on the boundary, such as the problems in Table 5.1, these approaches would fail at the beginning, simply because they are reduced to solving a nearly full-size SDP problem.

In contrast, our new approach does not depend on the quantity and location of anchors, since it is designed to improve the efficiency of solving a full-size SDP problem. In fact, any improvement on solving an individual SDP problem would complement the domain-decomposition approaches, since it would be possible to handle much larger-sized subproblems.

6. Future directions. From the computational results, we can see that the sub-SDP approaches indeed have a great potential to save computation time in solving sensor network localization problems, and the efficiency of the model is considerable. At the same time, they retain some of the most important theoretical features of the original SDP relaxation and achieve high localization quality.

There are many directions for future research. First, although our ESDP relaxation performs very well in localization quality, we still lack some powerful theorems to illustrate why the model works. This is a major issue that needs to be answered. Second, since, in our ESDP model, the decision matrix has its special structure, applying a tailored interior-point method (such as SCGD for the SOCP approach) may save more computational time. We also see that the NSDP relaxation has its own merit, both in theory and in practice. Therefore, further research about the NSDP model is also worth perusing. In fact, we have experimented with the NSDP model for solving the Max-Cut problem and will discuss its behavior and performance in another report. Finally, we plan to investigate the applicability of the SSDP relaxation idea for solving general SDP problems.

REFERENCES

[1] A. Y. Alfakih, *Graph rigidity via Euclidean distance matrices*, Linear Algebra Appl., 310 (2000), pp. 149–165.

[2] A. M.-C So and Y. Ye, *Theory of semidefinite programming for sensor network localization*, Math. Program., 109 (2007), pp. 367–384.

[3] A. Y. Alfakih, A. Khandani, and H. Wolkowicz, *Solving Euclidean distance matrix completion problems via semidefinite programming*, Comput. Optim. Appl., 12 (1999), pp. 13–30.

[4] A. Y. Alfakih, *On rigidity and realizability of weighted graphs*, Linear Algebra Appl., 325 (2001), pp. 57–70.

[5] M. Bâdoiu, *Approximation algorithm for embedding metrics into a two-dimensional space*, in SODA '03: Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, Philadelphia, 2003, pp. 434–443.

[6] A. I. Barvinokdoi, *Problems of distance geometry and convex properties of quadratic maps*, Discrete Comput. Geom., 13 (1995), pp. 189–202.
[7] P. Biswas and Y. Ye, *Semidefinite programming for ad hoc wireless sensor network localization*, in IPSN 2004 Proceedings of the Third International Symposium on Information Processing in Sensor Networks, ACM, New York, 2004, pp. 46–54.
[8] J. Blair and B. Peyton, *An introduction to chordal graphs and clique trees*, Inst. Math. Appl., 56 (1993), pp. 1–30.
[9] B. Grone, C. R. Johnson, E. M. Sa, and H. Wolkowicz, *Positive definite completions of partial Hermitian matrices*, Linear Algebra Appl., 58 (1984), pp. 109–124.
[10] D. Goldfarb and K. Scheinberg, *Interior point trajectories in semidefinite programming*, SIAM J. Optim., 8 (1998), pp. 871–886.
[11] O. Güler and Y. Ye, *Convergence behavior of interior-point algorithms*, Math. Program., 60 (1993), pp. 215–228.
[12] B. Hendrickson, *The molecule problem: Exploiting structure in global optimization*, SIAM J. Optim., 5 (1995), pp. 835–857.
[13] L. Hogben, *Graph theoretic methods for matrix completion problems*, Linear Algebra Appl., 328 (2001), pp. 161–202.
[14] C. R. Johnson and R. L. Smith, *The positive definite completion problem relative to a subspace*, Linear Algebra Appl., 307 (2000), pp. 1–14(14).
[15] J. Aspnes, D. Goldenberg, and Y. R. Yang, *On the computational complexity of sensor network localization*, in Algorithmic Aspects of Wireless Sensor Networks: First International Workshop, ALGOSENSORS 2004, Turku, Finland, July 16, 2004. Proceedings, Lecture Notes in Comput. Sci. 3121, Springer-Verlag, New York, 2004, pp. 32–44.
[16] J. A. Costa, N. Patwari, and A. O. Hero, III, *Distributed weighted-multidimensional scaling for node localization in sensor networks*, ACM Trans. Sen. Netw., 2 (2006), pp. 39–64.
[17] J. Dahl, V. Roychowdhury, and L. Vandenberghe, *Maximum-Likelihood Estimation of Multivariate Normal, Graphical Models: Large-Scale Numerical Implementation and Topology*, Working Paper, Electrical Engineering Department, UCLA, 2006.
[18] M. Kojima, S. Kim, and H. Waki, *Sparsity in sums of squares of polynomials*, Math. Program., 103 (2005), pp. 45–62.
[19] K. Q. Weinberger, F. Sha, and L. K. Saul, *Learning a kernel matrix for nonlinear dimensionality reduction*, in ICML '04: Proceedings of the Twenty-First International Conference on Machine Learning, ACM, New York, 2004, p. 106.
[20] M. Laurent, *Matrix completion problems*, Encyclopedia Optim., 3 (2001), pp. 221–229.
[21] L. Doherty, K. S. J. Pister, and L. El Ghaoui, *Convex position estimation in wireless sensor networks*, in INFOCOM 2001 Proceedings of the Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies, 2001, pp. 1655–1663.
[22] J. J. Moré and Z. Wu, *Global continuation for distance geometry problems*, SIAM J. Optim., 7 (1997), pp. 814–836.
[23] M. Bâdoiu, E. D. Demaine, M. T. Hajiaghayi, and P. Indyk, *Low-dimensional embedding with extra information*, Discrete Comput. Geom., 36 (2006), pp. 609–632.
[24] M. Fukuda, M. Kojima, K. Murota, and K. Nakata, *Exploiting sparsity in semidefinite programming via matrix completion* I: *General framework*, SIAM J. Optim., 11 (2000), pp. 647–674.
[25] M. W. Carter, H. H. Jin, M. A. Saunders, and Y. Ye, *SpaseLoc: An adaptive subproblem algorithm for scalable wireless sensor network localization*, SIAM J. Optim., 17 (2006), pp. 1102–1128.
[26] J. Nie, *Sum of squares method for sensor network localization*, Comput. Optim. Appl., to appear.
[27] G. Pataki, *On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eignevalues*, Math. Oper. Res., 23 (1998), pp. 339–358.
[28] P. Biswas, T.-C. Liang, K.-C. Toh, Y. Ye, and T.-C. Wang, *Semidefinite programming approaches for sensor network localization with noisy distance measurements*, IEEE Trans. Automat. Sci. Eng., 3 (2006), pp. 360–371.
[29] P. Biswas, T.-C. Lian, T.-C. Wang, and Y. Ye, *Semidefinite programming based algorithms for sensor network localization*, ACM Trans. Sen. Netw., 2 (2006), pp. 188–220.
[30] P. Tseng, *Second-order cone programming relaxation of sensor network localization*, SIAM J. Optim., 18 (2007), pp. 156–185.
[31] A. M.-C. So and Y. Ye, *A semidefinite programming approach to tensegrity theory and realizability of graphs*, in SODA '06: Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm, ACM, New York, 2006, pp. 766–775.

[32] J. F. Sturm, *Using SeDuMi* 1.02, *a MATLAB toolbox for optimization over symmetric cones*, Optim. Methods Softw., 11–12 (1999), pp. 625–653.
[33] S. Boyd, P. Diaconis, and L. Xiao, *Fastest mixing Markov chain on a graph*, SIAM Rev., 46 (2004), pp. 667–689.
[34] S. J. Benson, Y. Ye, and X. Zhang, *Solving large-scale sparse semidefinite programs for combinatorial optimization*, SIAM J. Optim., 10 (1999), pp. 443–461.
[35] T. Eren, O. K. Goldenberg, W. Whiteley, Y. R. Yang, A. S. Morse, B. D. O. Anderson, and P. N. Belhumeur, *Rigidity, computation, and randomization in network localization*, in INFOCOM 2004 Proceedings of the Twenty-Third Annual Joint Conference of the IEEE Computer and Communications Societies, vol. 4, 2004, pp. 2673–2684.
[36] Y. Ding, N. Krislock, J. Qian, and H. Wolkowicz, *Sensor Network Localization, Euclidean Distance Matrix. Completions, and Graph Realization*, Research report CORR 2006-23, Department of Combinatorics and Optimization, University of Waterloo, Waterloo, ON N2L 3G1, Canada.
[37] http://www.stanford.edu/˜yyye/

SIAM J. OPTIM.
Vol. 19, No. 2, pp. 674–699

A SAMPLE APPROXIMATION APPROACH FOR OPTIMIZATION WITH PROBABILISTIC CONSTRAINTS*

JAMES LUEDTKE† AND SHABBIR AHMED†

Abstract. We study approximations of optimization problems with probabilistic constraints in which the original distribution of the underlying random vector is replaced with an empirical distribution obtained from a random sample. We show that such a sample approximation problem with a risk level larger than the required risk level will yield a lower bound to the true optimal value with probability approaching one exponentially fast. This leads to an a priori estimate of the sample size required to have high confidence that the sample approximation will yield a lower bound. We then provide conditions under which solving a sample approximation problem with a risk level smaller than the required risk level will yield feasible solutions to the original problem with high probability. Once again, we obtain a priori estimates on the sample size required to obtain high confidence that the sample approximation problem will yield a feasible solution to the original problem. Finally, we present numerical illustrations of how these results can be used to obtain feasible solutions and optimality bounds for optimization problems with probabilistic constraints.

Key words. probabilistic constraints, chance constraints, Monte Carlo, stochastic programming, large deviation

AMS subject classification. 90C15

DOI. 10.1137/070702928

1. Introduction.

We consider optimization problems with probabilistic constraints (also known as chance constraints) of the form

$$\text{(PCP)} \qquad \min \{ f(x) : x \in X,\ \Pr \{ G(x, \xi) \leq \mathbf{0} \} \geq 1 - \epsilon \},$$

where $X \subset \mathbf{R}^n$ represents a deterministic feasible region, $f : \mathbf{R}^n \to \mathbf{R}$ represents the objective to be minimized, ξ is a random vector with suport $\Xi \subseteq \mathbf{R}^d$, $G : \mathbf{R}^n \times \mathbf{R}^d \to \mathbf{R}^m$ is a given constraint mapping, and ϵ is a risk parameter chosen by the decision maker, typically near zero, e.g., $\epsilon = 0.01$ or $\epsilon = 0.05$. Such problems are sometimes called probabilistic programs. In (PCP) a single probabilistic constraint is enforced over *all* rows in the constraints $G(x, \xi) \leq \mathbf{0}$ rather than requiring that each row independently be satisfied with high probability. Such a constraint is known as a *joint probabilistic constraint* and is appropriate in a context in which it is important to have all constraints satisfied simultaneously and there may be dependence between random variables in different rows.

Problems with joint probabilistic constraints have been extensively studied; see [25] for a background and an extensive list of references. Probabilistic constraints have been used in various applications including supply chain management [17], production planning [21], optimization of chemical processes [13, 14], and surface water quality management [30].

Unfortunately, probabilistic programs are still largely intractable except for a few special cases. There are two primary reasons for this intractability. First, in general, for a given $x \in X$, the quantity $\Pr\{G(x, \xi) \leq \mathbf{0}\}$ is hard to compute, as it requires

*Received by the editors September 15, 2007; accepted for publication (in revised form) February 20, 2008; published electronically July 2, 2008. This research has been supported in part by the National Science Foundation under grants DMI-0133943 and DMI-0522485.

http://www.siam.org/journals/siopt/19-2/70292.html

†H. Milton Stewart School of Industrial & Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332 (jrluedt1@wisc.edu, sahmed@isye.gatech.edu).

multidimensional integration, and hence just checking the feasibility of a solution is difficult. Second, the feasible region defined by a probabilistic constraint generally is not convex. In this paper, we study how the difficulty in checking feasibility can be addressed by solving a *sample approximation* problem based on a Monte Carlo sample of ξ. In particular, we study how this approximation can be used to generate feasible solutions and optimality bounds for general probabilistic programs.

The sample approximation that we study is a probabilistic program in which the original distribution of the random vector ξ is replaced with the empirical distribution obtained from the random sample. We show that such a sample approximation problem with a risk level larger than the nominal risk level ϵ will yield a lower bound to the true optimal value with probability approaching one exponentially fast. This leads to an a priori estimate of the sample size required to have high confidence that the sample approximation will yield a lower bound. We also discuss alternative means of generating lower bounds, which can be used regardless of the sample size used. We then provide conditions under which solving a sample approximation problem with a risk level smaller than ϵ will yield feasible solutions to the original problem with high probability. Once again, we obtain a priori estimates on the sample size required to obtain high confidence that the sample approximation problem will yield a feasible solution to the original problem.

Recently, a number of approaches have been proposed to find approximate solutions to probabilistic programs; the common theme among these is that they all seek "safe" or conservative approximations which can be solved efficiently. That is, they propose approximation problems which are convex and yield solutions which are feasible, or at least highly likely to be feasible, to the original probabilistic program. Approaches of this type include: the scenario approximation method studied by Calafiore and Campi [7, 8] and extended by Nemirovski and Shapiro [22]; the Bernstein approximation scheme of Nemirovski and Shapiro [23]; and robust optimization, e.g., [4, 6, 11]. The conservative approximations, when applicable, are attractive because they allow efficient generation of feasible solutions. In particular, they can yield feasible solutions when the probabilistic constraint is "hard," that is, with ϵ very small, such as $\epsilon = 10^{-6}$ or even $\epsilon = 10^{-12}$. However, in a context in which ϵ is not so small, such as $\epsilon = 0.05$ or $\epsilon = 0.01$, the probabilistic constraint is more likely to represent a "soft" constraint, one which the decision-maker would like to have satisfied but is willing to allow a nontrivial chance that it will be violated if doing so would sufficiently decrease the cost of the implemented solution. In this latter context, it would be desirable to obtain solutions which are feasible to the probabilistic constraint *along with* an assurance that the solutions are not much more costly than the lowest-cost solution attaining the same risk level. In this way, the decision-maker can be confident that they are choosing from solutions on the efficient frontier between the competing objectives of cost and risk. Unfortunately, the recently proposed conservative approximations say very little in terms of how conservative the solutions are. In particular, it is generally not possible to make a statement about how much worse the objective is relative to the optimal value at a fixed risk level ϵ.

The scenario approximation methods are most similar to the sample approach that we study in that they solve an approximation problem based on an independent Monte Carlo sample of the random vector. For example, the scenario approximation of [7, 8] takes a sample $\xi^1, \ldots, \xi^N$ and solves the problem

$$\min_{x \in X} \{ f(x) : G(x, \xi^i) \leq \mathbf{0}, \quad i = 1, \ldots, N \}. \tag{1}$$

That is, the scenario approximation enforces *all* of the constraints corresponding to the sample taken. When the nominal problem is convex (that is, $X \subseteq \mathbf{R}^n$ is a convex set, f is convex, and G is convex in x for each ξ), they show that the scenario approximation problem will yield a feasible solution to (PCP) with probability of at least $1 - \delta$ for

$$N \geq \frac{2}{\epsilon} \log\left(\frac{1}{\delta}\right) + 2n + \frac{2n}{\epsilon} \log\left(\frac{2}{\epsilon}\right). \tag{2}$$

In addition, under the stated convexity assumptions, the scenario approximation problem remains a convex program. An advantage of this approach relative to the approximations [4, 6, 11, 23] is that the only assumption that is made on the distribution of ξ is that it can be sampled from.

The key difference between the sample approximation that we study and scenario approximation is that we allow the risk level in the sample approximation problem to be positive; that is, we do not require that all sampled constraint sets be satisfied. Instead, the constraint sets which will be satisfied can be chosen optimally. The disadvantage of this scheme is that the sample approximation problem with a positive risk level has a nonconvex feasible region and hence may be difficult to solve despite having a simplified probabilistic structure. Specifically, if we allow k of the N sampled constraint sets to be violated, then we must choose a set of k constraint sets which will not be enforced, and there are $\binom{N}{k}$ possible sets from which to choose. Choosing the optimal set is an NP-hard problem even in a very special case [20]. However, in some special cases, such as when randomness appears only in the right-hand side of the constraints, the sample approximation problem may be relatively tractable to solve with integer programming techniques; see [20, 19]. In addition, for generating feasible solutions to (PCP), our analysis indicates that with appropriately chosen parameters *any* feasible solution to the sample approximation problem will be feasible to the original problem with high probability, so that it is sufficient to generate heuristic solutions. Similarly, to obtain a lower bound for (PCP), it is sufficient to obtain a lower bound for the appropriate sample approximation problem.

In the context of generating feasible solutions for (PCP), our sample approximation scheme includes as a special case the scenario approximation of [7, 8] in which the constraints corresponding to all sampled vectors ξ^i are enforced. In this special case, we obtain results very similar to those in [8] in terms of how many samples should be used to yield a solution feasible to (PCP) with high probability. However, our analysis is quite different from the analysis of [8] and, in particular, requires a significantly different set of assumptions. In some cases our assumptions are more stringent, but there are also a number of cases in which our assumptions apply and those of [8] do not, most notably if the feasible region X is not convex, as in the case of a mixed-integer program. Thus, our results complement those of [8] in two ways: First we show that sample approximations with positive risk levels can be used to yield feasible solutions to (PCP), and second we relax the convexity assumptions. Another closely related work is [9], in which the authors consider a sample approximation problem in which some of the sampled constraints are allowed to be violated. When the nominal problem is convex and a nondegeneracy assumption holds, they present an estimate on the sample size needed to obtain a feasible solution with high probability when a fixed number of sampled constraint sets are discarded optimally. Under these assumptions, their results for generating feasible solutions are very similar to the results that we present. The unique contributions of the present paper are (1) we use assumptions which are significantly different from the convexity and nondegeneracy assumptions

used in [9] (neither set of assumptions implies the other), (2) we analyze a method for generating *lower* bounds on the optimal value (which is useful for validating the quality of a given solution), (3) we prove that the sample approximation yields an *exact* optimal solution with high probability when X is finite (as in the case of an integer program), and (4) we conduct extensive numerical experiments on practical size problems indicating the potential of the approach.

The sample approximation problem that we study can be thought of as a variation of the well-studied *sample average approximation* (SAA) approach; see, e.g., [1, 10, 16, 29]. The difference is that the approximation that we study enforces a sample average constraint involving expectations of indicator functions, whereas the SAA approach typically optimizes a sample average objective. Shapiro [28] and Wang [32] have considered SAA approximation for expected value constraints. However, in these works, the function taken under expectation in the constraints is assumed to be continuous, and hence these results cannot be directly applied because of the discontinuity of indicator functions. In [2] a model with expected value constraints in which the function taken under expectation is not necessarily continuous is considered, and hence their analysis does apply to the case of probabilistic constraints. However, they consider only the case in which the feasible region is finite, and they discuss only the theoretical rate of convergence. In contrast, we begin with a similar analysis for the finite feasible region case but then extend the analysis to a number of significantly more general settings. In addition, we separate the analysis of when the sample approximation will be likely to yield a lower bound and when it will be likely to yield feasible solutions. This separate analysis allows for the development of methods which yield optimality statements which hold with high probability.

Finally, we mention the work of Vogel [31], which considers convergence properties of the sample approximation we use for probabilistic programs. When only the right-hand side is random with continuous distribution, it is shown that the probability that the distance between the sample feasible region and the true feasible region is larger than any positive threshold decreases exponentially fast with the size of the sample. However, the convergence rate has poor dependence on the dimension of the random vector, implying that the number of samples required to yield a reasonable approximation would have to grow exponentially in this dimension. Better convergence is demonstrated for the case of random right-hand side with discrete distribution. For the general case, linear convergence is demonstrated in the case of continuous distributions. Our analysis of the sample approximation problem extends these results by improving on the convergence rates and by analyzing what happens when the sample approximation problem is allowed to have a different risk level than the nominal risk level ϵ. This allows the sample approximation problem to be used to generate feasible solutions and optimality bounds.

The remainder of this paper is organized as follows. In section 2 we present and analyze the sample approximation scheme. We present results of a preliminary computational study of the use of the sample approximation scheme in section 3. We close with concluding remarks and directions for future research in section 4.

2. Analysis of sample approximation. We now study how Monte Carlo sampling can be used to generate probabilistically constrained problems with finite distribution which can be used to approximate problems with general distributions. Let us restate (PCP) as

$$(\mathrm{P}_\epsilon) \qquad z_\epsilon^* = \min\{f(x) : x \in X_\epsilon\},$$

where

$$X_\epsilon = \Big\{ x \in X : \Pr\{G(x,\xi) \leq \mathbf{0}\} \geq 1-\epsilon \Big\}.$$

We assume that z_ϵ^* exists and is finite. For example, if X is compact and $G(x,\xi)$ is affine in x for each $\xi \in \Xi$, then X_ϵ is closed [12] and hence compact, and so if $f(x)$ is continuous, then an optimal solution exists whenever $X_\epsilon \neq \emptyset$. Furthermore, we take as an assumption the measurability of any event S taken under probability, such as the event $\{G(x,\xi) \leq \mathbf{0}\}$ for each $x \in X$.

If X is a polyhedron, $f(x) = cx$, $G(x,\xi) = \xi - Tx$ $(d = m)$, then we obtain the probabilistically constrained linear program with random right-hand side

$$\min\Big\{ cx : x \in X, \Pr\{Tx \geq \xi\} \geq 1-\epsilon \Big\}.$$

We can also model a two-stage problem in which we make a decision x and wish to guarantee that with probability at least $1-\epsilon$ there is a feasible recourse decision y satisfying $Wy \geq H(x,\xi)$, where W is an m by l matrix and $H : \mathbf{R}^n \times \mathbf{R}^d \to \mathbf{R}^m$. This is accomplished by letting $G : \mathbf{R}^n \times \mathbf{R}^d \to \mathbf{R}$ be defined by

$$G(x,\xi) = \min_{\mu,y} \{\mu : Wy + \mu\mathbf{e} \geq H(x,\xi), \mu \geq -1\},$$

where $\mathbf{e} \in \mathbf{R}^m$ is a vector of all ones. Indeed, $G(x,\xi) \leq \mathbf{0}$ if and only if there exists $y \in \mathbf{R}^l$ and $\mu \leq 0$ such that $Wy + \mu\mathbf{e} \geq H(x,\xi)$, which occurs if and only if there exists $y \in \mathbf{R}^l$ such that $Wy \geq H(x,\xi)$.

Due to the general difficulty in calculating $\Pr\{G(x,\xi) \leq \mathbf{0}\}$ for a given $x \in X$, we seek to approximate (P_ϵ) by solving a sample approximation problem. We let $\xi^1, \ldots, \xi^N$ be an independent Monte Carlo sample of the random vector ξ. Then, for fixed $\alpha \in [0,1)$, the sample approximation problem is defined to be

$$(\mathrm{P}_\alpha^N) \qquad \hat{z}_\alpha^N = \min\big\{ f(x) : x \in X_\alpha^N \big\},$$

where

$$X_\alpha^N = \left\{ x \in X : \frac{1}{N} \sum_{i=1}^N \mathbb{I}\big(G(x,\xi^i) \leq \mathbf{0}\big) \geq 1-\alpha \right\},$$

where $\mathbb{I}(\cdot)$ is the indicator function which takes value one when $\cdot$ is true and zero otherwise. We adopt the convention that if $X_\alpha^N = \emptyset$, then $\hat{z}_\alpha^N = +\infty$, whereas if (P_α^N) is unbounded, we take $\hat{z}_\alpha^N = -\infty$. We assume that, except for these two cases, (P_α^N) has an optimal solution. This assumption is satisfied, for example, if X is compact, $f(x)$ is continuous, and $G(x,\xi)$ is continuous in x for each $\xi \in \Xi$, since then X_α^N is the union of finitely many compact sets (in this case $\hat{z}_\alpha^N = -\infty$ is also not possible). If $\alpha = 0$, the sample approximation problem (P_0^N) corresponds to the *scenario approximation* of probabilistic constraints, studied in [8, 22]. Our goal is to establish statistical relationships between problems (P_ϵ) and (P_α^N) for $\alpha \geq 0$. We first consider when (P_α^N) yields lower bounds for (P_ϵ) and then consider when (P_α^N) yields feasible solutions for (P_ϵ).

2.1. Lower bounds. We now establish a bound on the probability that (P_α^N) yields a lower bound for (P_ϵ). Let

$$\rho(\alpha, \epsilon, N) = \sum_{i=0}^{\lfloor \alpha N \rfloor} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}.$$

$\rho(\alpha, \epsilon, N)$ represents the probability of having at most $\lfloor \alpha N \rfloor$ "successes" in N independent trials, in which the probability of a success in each trial is ϵ.

LEMMA 1. *Assume that* (P_ϵ) *has an optimal solution. Then*

$$\Pr\left\{\hat{z}_\alpha^N \le z_\epsilon^*\right\} \ge \rho(\alpha, \epsilon, N).$$

Proof. Let $x^* \in X_\epsilon$ be an optimal solution to (P_ϵ). Then $\Pr\{G(x^*, \xi^i) \not\le \mathbf{0}\} \le \epsilon$ for each i. Hence, if we call the event $\{G(x^*, \xi^i) \not\le \mathbf{0}\}$ a success, then the probability of a success in trial i is $\bar{\phi}(x^*) := \Pr\{G(x^*, \xi^i) \not\le \mathbf{0}\} \le \epsilon$. By the definition of X_α^N, $x^* \in X_\alpha^N$ if and only if

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^N \mathbb{I}\big(G(x^*, \xi^i) \le \mathbf{0}\big) \ge 1-\alpha \quad &\Leftrightarrow \quad \frac{1}{N}\sum_{i=1}^N \mathbb{I}\big(G(x^*, \xi^i) \not\le \mathbf{0}\big) \le \alpha \\ &\Leftrightarrow \quad \sum_{i=1}^N \mathbb{I}\big(G(x^*, \xi^i) \not\le \mathbf{0}\big) \le \lfloor \alpha N \rfloor. \end{aligned}$$

Hence, $\Pr\{x^* \in X_\alpha^N\}$ is the probability of having at most $\lfloor \alpha N \rfloor$ successes in N trials. Also, if $x^* \in X_\alpha^N$, then $\hat{z}_\alpha^N \le z_\epsilon^*$. Thus,

$$\Pr\left\{\hat{z}_\alpha^N \le z_\epsilon^*\right\} \ge \Pr\left\{x^* \in X_\alpha^N\right\} = \rho(\alpha, \bar{\phi}(x^*), N) \ge \rho(\alpha, \epsilon, N)$$

since $\rho(\alpha, \epsilon, N)$ is decreasing in ϵ. ∎

For example, if $\alpha = 0$ as in the previously studied scenario approximation [8, 22], then we obtain $\Pr\{\hat{z}_\alpha^N \le z_\epsilon^*\} \ge \rho(0, \epsilon, N) = (1-\epsilon)^N$. For this choice of α, it becomes very unlikely that the sample approximation (P_α^N) will yield a lower bound as N gets large. For $\alpha > \epsilon$ we see different behavior: the sample approximation yields a lower bound with probability approaching one exponentially fast as N increases. The proof is based on Hoeffding's inequality.

THEOREM 2 (Hoeffding's inequality [15]). *Let* $Y_1, \ldots, Y_N$ *be independent random variables, with* $\Pr\{Y_i \in [a_i, b_i]\} = 1$*, where* $a_i \le b_i$ *for* $i = 1, \ldots, N$*. Then if* $t > 0$*,*

$$\Pr\left\{\sum_{i=1}^N (Y_i - \mathbb{E}[Y_i]) \ge tN\right\} \le \exp\left\{-\frac{2N^2t^2}{\sum_{i=1}^N (b_i - a_i)^2}\right\}.$$

THEOREM 3. *Let* $\alpha > \epsilon$*, and assume that* (P_ϵ) *has an optimal solution. Then*

$$\Pr\left\{\hat{z}_\alpha^N \le z_\epsilon^*\right\} \ge 1 - \exp\big\{-2N(\alpha-\epsilon)^2\big\}.$$

Proof. Let x^* be an optimal solution to (P_ϵ). As in the proof of Lemma 1, if $x^* \in X_\alpha^N$, then $\hat{z}_\alpha^N \le z_\epsilon^*$. For $i = 1, \ldots, N$ let Y_i be a random variable taking value 1

if $G(x^*, \xi^i) \not\leq \mathbf{0}$ and 0 otherwise. Then $\Pr\{Y_i \in [0,1]\} = 1$ and $E[Y_i] \leq \epsilon$. Hence,

$$\begin{aligned}\Pr\left\{\hat{z}_\alpha^N > z_\epsilon^*\right\} \leq \Pr\left\{x^* \notin X_\alpha^N\right\} &= \Pr\left\{\frac{1}{N}\sum_{i=1}^N Y_i > \alpha\right\} \\ &\leq \Pr\left\{\frac{1}{N}\sum_{i=1}^N (Y_i - E[Y_i]) > \alpha - \epsilon\right\} \\ &\leq \exp\Big\{-\frac{2N^2(\alpha-\epsilon)^2}{N}\Big\} = \exp\big\{-2N(\alpha-\epsilon)^2\big\},\end{aligned}$$

where the first inequality follows since $E[Y_i] \leq \epsilon$ and the second inequality follows from Hoeffding's inequality. □

Theorem 3 states that, by taking a risk parameter $\alpha > \epsilon$ in our sample approximation problem, we will obtain a lower bound to the true optimal value with probability approaching one exponentially fast as N increases. Stated another way, suppose that we solve a sample approximation problem (P_α^N) with $\alpha = \epsilon$. Then for any $\gamma > 0$ such that $\gamma < \epsilon$, the optimal value of this problem, $\hat{z}_\epsilon^N$, will be a lower bound to the optimal value of $\mathrm{P}_{\epsilon-\gamma}$ with probability approaching one exponentially fast with N. If γ is small, this states that the optimal solution to the sample problem will have cost no worse than any solution that is "slightly less risky" than the nominal risk level ϵ.

Theorem 3 immediately yields a method for generating lower bounds with specified confidence $1-\delta$, where $\delta \in (0,1)$. If we select $\alpha > \epsilon$ and

$$N \geq \frac{1}{2(\alpha-\epsilon)^2}\log\Big(\frac{1}{\delta}\Big),$$

then Theorem 3 ensures that $\hat{z}_\alpha^N \leq z_\epsilon^*$ with probability of at least $1-\delta$. Indeed, with this choice of α and N, we have

$$\Pr\left\{\hat{z}_\alpha^N > z_\epsilon^*\right\} \leq \exp\big\{-2N(\alpha-\epsilon)^2\big\} \leq \exp\Big\{-\log\Big(\frac{1}{\delta}\Big)\Big\} = \delta.$$

Because $1/\delta$ is taken under logarithm, we can obtain a lower bound with high confidence, i.e., with δ very small, without significantly increasing the required sample size N. On the other hand, the required sample size grows quadratically with $1/(\alpha-\epsilon)$ and hence will be large for α very close to ϵ.

Lemma 1 can also be used to obtain lower bounds with specified confidence, by using the bounding procedure proposed by Nemirovski and Shapiro [23]. They restrict $\alpha = 0$ in the sample approximation, but the technique can be applied in exactly the same way when $\alpha > 0$, and it is likely that this can make the bounding technique significantly more powerful. The idea is as follows. Take M sets of N independent samples of ξ, given by $\xi^{i,j}$ for $j = 1, \ldots, M$ and $i = 1, \ldots, N$, and for each j solve the associated sample approximation problem

$$\hat{z}_{\alpha,j}^N = \min\big\{f(x) : x \in X_{\alpha,j}^N\big\},$$

where

$$X_{\alpha,j}^N = \left\{x \in X : \frac{1}{N}\sum_{i=1}^N \mathbb{I}\big(G(x,\xi^{i,j}) \leq \mathbf{0}\big) \geq 1-\alpha\right\}.$$

We then rearrange the values $\{\hat{z}^N_{\alpha,j}\}_{j=1}^M$ to obtain the *order statistics* $\hat{z}^N_{\alpha,[j]}$ for $j = 1,\ldots,M$ satisfying $\hat{z}^N_{\alpha,[1]} \le \cdots \le \hat{z}^N_{\alpha,[M]}$. Then a lower bound which is valid with specified confidence $1-\delta$ can be obtained as follows.

THEOREM 4. *Let $\delta \in (0,1)$, $\alpha \in [0,1)$, and N, L, and M be positive integers such that $L \le M$ and*

$$\sum_{i=0}^{L-1} \binom{M}{i} \rho(\alpha,\epsilon,N)^i \big(1-\rho(\alpha,\epsilon,N)\big)^{M-i} \le \delta. \tag{3}$$

Then

$$\Pr\left\{\hat{z}^N_{\alpha,[L]} \le z^*_\epsilon\right\} \ge 1-\delta.$$

Proof. We show that $\Pr\{\hat{z}^N_{\alpha,[L]} > z^*_\epsilon\} \le \delta$. Note that $\hat{z}^N_{\alpha,[L]} > z^*_\epsilon$ if and only if less than L of the values $\hat{z}^N_{\alpha,j}$ satisfy $\hat{z}^N_{\alpha,j} \le z^*_\epsilon$. Thus, calling the event $\{\hat{z}^N_{\alpha,j} \le z^*_\epsilon\}$ a success, the event $\hat{z}^N_{\alpha,[L]} > z^*_\epsilon$ occurs if and only if there are fewer than L successes in M trials, in which the probability of a success is $\eta := \Pr\{\hat{z}^N_{\alpha,j} \le z^*_\epsilon\}$. The result then follows since $\eta \ge \rho(\alpha,\epsilon,N)$ by Lemma 1 and so

$$\sum_{i=0}^{L-1} \binom{M}{i} \eta^i (1-\eta)^{M-i} \le \sum_{i=0}^{L-1} \binom{M}{i} \rho(\alpha,\epsilon,N)^i \big(1-\rho(\alpha,\epsilon,N)\big)^{M-i} \le \delta$$

by (3). □

An interesting special case of Theorem 4 is obtained by taking $L = 1$. In this case, we are taking as our lower bound the minimum of the optimal values obtained from solving the M sample approximation problems. To have confidence $1-\delta$ that the lower bound is truly a lower bound, we should choose M such that

$$\big(1-\rho(\alpha,\epsilon,N)\big)^M \le \delta. \tag{4}$$

With the choice of $L = 1$, let us consider how large M should be with $\alpha = 0$ and with $\alpha = \epsilon$. With $\alpha = 0$, we obtain $\rho(0,\epsilon,N) = (1-\epsilon)^N$. Hence, to have confidence $1-\delta$ to obtain a lower bound, we should take

$$M \ge \log\left(\frac{1}{\delta}\right) / \log\left(\frac{1}{1-(1-\epsilon)^N}\right). \tag{5}$$

By using the inequality $\log(1+x) \le x$ for $x > 0$, we have

$$\log\left(\frac{1}{1-(1-\epsilon)^N}\right) = \log\left(1 + \frac{(1-\epsilon)^N}{1-(1-\epsilon)^N}\right) \le \frac{(1-\epsilon)^N}{1-(1-\epsilon)^N}.$$

Hence, when $\alpha = 0$, we should take

$$M \ge \log\left(\frac{1}{\delta}\right) \frac{1-(1-\epsilon)^N}{(1-\epsilon)^N}.$$

Thus, for fixed $\epsilon \in (0,1)$, the required M grows exponentially in N. For example, by using (5), if $\delta = 0.001$ and $\epsilon = 0.01$, then for $N = 250$ we need $M \ge 82$, for $N = 500$ we need $M \ge 1048$, and for $N = 750$ we need $M \ge 12967$. If $\delta = 0.001$ and $\epsilon = 0.05$,

then for $N = 50$ we should take $M \geq 87$, for $N = 100$ we should take $M \geq 1160$, and for $N = 150$ we must already have $M \geq 15157$! Thus, to keep M reasonably small, we must keep N small, but this will weaken the lower bound obtained in each sample.

Now suppose that we take $L = 1$ and $\alpha = \epsilon$. Then, for N "large enough" (e.g., $N\epsilon \geq 10$), we have $\rho(\epsilon, \epsilon, N) \approx 1/2$. Indeed, $\rho(\epsilon, \epsilon, N)$ is the probability that a binomial random variable with success probability ϵ and N trials is at most $\lfloor \epsilon N \rfloor$. With N large enough relative to ϵ, this probability can be approximated by the probability that a random variable with normal distribution having mean ϵN does not exceed $\lfloor \epsilon N \rfloor$. Because the median of the normal distribution equals the mean, we obtain $\rho(\epsilon, \epsilon, N) \gtrsim 1/2$. Thus, with $L = 1$ and $\alpha = \epsilon$, we should choose M such that $(1/2)^M \leq \delta$ or

$$M \geq \log_2 \left(\frac{1}{\delta} \right).$$

Note that this bound is *independent of* N *and* ϵ. For example, for $\delta = 0.001$, we should take $M \geq 10$. The independence of N has the advantage that we can take N to be as large as is computationally tractable, which will tend to make each of the optimal values $\hat{z}^N_{\epsilon,j}$ closer to the true optimal z^*_ϵ and hence make the lower bound $\min_j \{\hat{z}^N_{\epsilon,j}\}$ tighter.

We close this section by commenting that, although our results have been stated in terms of the *exact* optimal solution $\hat{z}^N_\alpha$ of the sample approximation problem, it is not necessary to calculate this value exactly to use the results. All of the results about lower bounds for z^*_ϵ will be valid if $\hat{z}^N_\alpha$ is replaced with a lower bound of $\hat{z}^N_\alpha$, at the expense, of course, of weakening the lower bound.

2.2. Feasible solutions. We now consider conditions under which an optimal solution to (P^N_α), if one exists, is feasible to (P_ϵ). The idea is that if we take the risk parameter α in (P^N_α) to be smaller than ϵ, then for N large enough the feasible region of (P^N_α) will be a subset of the feasible region of (P_ϵ), so that any optimal solution to (P^N_α) must be feasible to (P_ϵ). Unlike the case for lower bounds, we will need to make additional assumptions to assure that (P^N_α) yields a feasible solution with high probability.

We begin by assuming that the feasible region X is finite. Note, however, that $|X|$ may be exponentially large; for example, X could be the feasible region of a bounded integer program. We then show how this assumption can be relaxed and replaced with some milder assumptions.

2.2.1. Finite X.

THEOREM 5. *Suppose that X is finite and $\alpha \in [0, \epsilon)$. Then*

$$\Pr\left\{X^N_\alpha \subseteq X_\epsilon\right\} \geq 1 - |X \setminus X_\epsilon| \exp\left\{-2N(\epsilon - \alpha)^2\right\}.$$

Proof. Consider any $x \in X \setminus X_\epsilon$, i.e., $x \in X$ with $\Pr\{G(x, \xi) \leq \mathbf{0}\} < 1 - \epsilon$. We want to estimate the probability that $x \in X^N_\alpha$. For $i = 1, \ldots, N$ define the random variable Y_i by $Y_i = 1$ if $G(x, \xi^i) \leq \mathbf{0}$ and $Y_i = 0$ otherwise. Then $E[Y_i] = \Pr\{G(x, \xi^i) \leq \mathbf{0}\} < 1 - \epsilon$ and $\Pr\{Y_i \in [0, 1]\} = 1$. By observing that $x \in X^N_\alpha$ if and only if $(1/N) \sum_{i=1}^N Y_i \geq 1 - \alpha$ and applying Hoeffding's inequality, we obtain

$$\begin{aligned}\Pr\left\{x \in X^N_\alpha\right\} = \Pr\left\{ \frac{1}{N} \sum_{i=1}^N Y_i \geq 1 - \alpha \right\} &\leq \Pr\left\{ \sum_{i=1}^N (Y_i - E[Y_i]) \geq N(\epsilon - \alpha) \right\} \\ &\leq \exp\left\{-2N(\epsilon - \alpha)^2\right\}.\end{aligned}$$

Then

$$\begin{aligned}\Pr\left\{X_\alpha^N \not\subseteq X_\epsilon\right\} &= \Pr\left\{\exists x \in X_\alpha^N \text{ such that } \Pr\left\{G(x,\xi) \leq \mathbf{0}\right\} < 1-\epsilon\right\} \\ &\leq \sum_{x \in X \setminus X_\epsilon} \Pr\left\{x \in X_\alpha^N\right\} \\ &\leq |X \setminus X_\epsilon| \exp\left\{-2N(\epsilon-\alpha)^2\right\}. \qquad \square\end{aligned}$$

For fixed $\alpha < \epsilon$ and $\delta \in (0,1)$, Theorem 5 shows that if we take

$$N \geq \frac{1}{2(\epsilon-\alpha)^2} \log\left(\frac{|X \setminus X_\epsilon|}{\delta}\right),$$

then, if (P_α^N) is feasible, it will yield a feasible solution to (P_ϵ) with probability at least $1-\delta$. If $|X| \leq U^n$, we can take

$$N \geq \frac{1}{2(\epsilon-\alpha)^2} \log\left(\frac{1}{\delta}\right) + \frac{n}{2(\epsilon-\alpha)^2} \log(U). \tag{6}$$

Note that N grows linearly with the dimension n of the feasible region and logarithmically with $1/\delta$, so that the confidence of generating a feasible solution can be made large without requiring N to be too large. However, the quadratic dependence on $\epsilon - \alpha$ implies that this a priori estimate of how large N should be will grow quite large for α near ϵ.

Theorem 5 states that for $\alpha < \epsilon$ *every* feasible solution to the sample approximation problem will be feasible to the original problem with risk level ϵ with high probability as N gets large. This is in contrast to the results of the scenario approximation method presented in [8] in which $\alpha = 0.0$ is required, and the result is that the *optimal* solution to the sample approximation problem will be feasible to the original problem with high probability. The advantage of our approach is that one need not solve the sample approximation problem to optimality to obtain a solution to the original problem. Simple heuristics which select which sampled constraints to be satisfied, e.g., greedily or by local search, can be used to yield feasible solutions for the approximation problem, which by virtue of Theorem 5 will have high probability of being feasible to the original problem. This comment also applies to subsequent feasibility results in which we relax the assumption that the feasible region X is finite.

In this case of finite X, we can combine Theorem 5 with Theorem 3 to demonstrate that solving a sample approximation with $\alpha = \epsilon$ will yield an exact optimal solution with probability approaching one exponentially fast with N. Let X_ϵ^* be the set of optimal solutions to (P_ϵ), and define $\underline{\alpha} = \max\{\Pr\{G(x,\xi) \not\leq \mathbf{0}\} : x \in X_\epsilon^*\}$. By definition, we have $z_{\underline{\alpha}}^* = z_\epsilon^*$. Next, let $\overline{\alpha} = \min\{\Pr\{G(x,\xi) \not\leq \mathbf{0}\} : x \in X \setminus X_\epsilon\}$. By definition, we have $\overline{\alpha} > \epsilon$. Finally, define $\kappa = \min\{\epsilon - \underline{\alpha}, \overline{\alpha} - \epsilon\}$.

COROLLARY 6. *Assume that* $\underline{\alpha} < \epsilon$. *Then*

$$\Pr\left\{\hat{z}_\epsilon^N = z_\epsilon^*\right\} \geq 1 - (|X|+1)\exp\left\{-2N\kappa^2\right\}.$$

Proof. First observe that $\kappa > 0$ when $\underline{\alpha} < \epsilon$. Next, we apply Theorem 3 with $\underline{\alpha}$ in place of ϵ and ϵ in place of α to obtain $\Pr\{\hat{z}_\epsilon^N \leq z_{\underline{\alpha}}^*\} \geq 1 - \exp\{-2N(\epsilon-\underline{\alpha})^2\}$. Because $z_{\underline{\alpha}}^* = z_\epsilon^*$, this implies that $\Pr\{\hat{z}_\epsilon^N > z_\epsilon^*\} \leq \exp\{-2N(\epsilon-\underline{\alpha})^2\}$.

We next observe that the proof of Theorem 5 can be modified to show the slightly stronger result that

$$\Pr\left\{X_\alpha^N \subseteq X_\epsilon'\right\} \geq 1 - |X \setminus X_\epsilon'| \exp\left\{-2N(\epsilon-\alpha)^2\right\},$$

where $X'_\epsilon = \{x \in X : \Pr\{G(x,\xi) \leq \mathbf{0}\} > 1-\epsilon\}$. (In the proof, we consider each $x \in X \setminus X'_\epsilon$ and observe that the defined random variable Y_i satisfies $E[Y_i] \leq 1-\epsilon$. The remainder of the proof is identical with X_ϵ replaced by X'_ϵ.) By applying this result, we obtain

$$\Pr\left\{X_\epsilon^N \subseteq X'_{\overline{\alpha}}\right\} \geq 1 - |X \setminus X'_{\overline{\alpha}}| \exp\left\{-2N(\overline{\alpha}-\epsilon)^2\right\}.$$

However, if $x \in X'_{\overline{\alpha}}$, then $\Pr\{G(x,\xi) \not\leq \mathbf{0}\} < \overline{\alpha}$, and by definition of $\overline{\alpha}$ this implies that $\Pr\{G(x,\xi) \not\leq \mathbf{0}\} \leq \epsilon$ and thus $X'_{\overline{\alpha}} \subseteq X_\epsilon$. It follows that

$$\Pr\left\{\hat{z}_\epsilon^N < z_\epsilon^*\right\} \leq \Pr\left\{X_\epsilon^N \not\subseteq X_\epsilon\right\} \leq |X| \exp\left\{-2N(\overline{\alpha}-\epsilon)^2\right\}.$$

Therefore,

$$\begin{aligned}\Pr\left\{\hat{z}_\epsilon^N \neq z_\epsilon^*\right\} &\leq \Pr\left\{\hat{z}_\epsilon^N > z_\epsilon^*\right\} + \Pr\left\{\hat{z}_\epsilon^N < z_\epsilon^*\right\} \\ &\leq \exp\left\{-2N(\epsilon-\underline{\alpha})^2\right\} + |X| \exp\left\{-2N(\overline{\alpha}-\epsilon)^2\right\} \\ &\leq (1+|X|)\exp\left\{-2N\kappa^2\right\}. \qquad \square\end{aligned}$$

The assumption that $\underline{\alpha} < \epsilon$ is mild since, because X is finite, there are only finitely many values of $\epsilon \in [0,1]$ for which it is possible to have $\epsilon = \underline{\alpha}$. Stated another way, if we add a random perturbation uniformly distributed in $(-\gamma,\gamma)$ to ϵ, where γ can be arbitrarily small, then the assumption will hold with probability one. On the other hand, the number of scenarios required to guarantee a reasonably high probability of obtaining the optimal solution will be at least proportional to $(\epsilon-\underline{\alpha})^{-2}$ and hence may be very large. Thus, Corollary 6 illustrates the *qualitative* behavior of the sample approximation with $\alpha = \epsilon$ in the finite feasible region case but may not be useful for estimating the required sample size.

If we take $\alpha = 0$ in Theorem 5, we obtain improved dependence of N on ϵ.

THEOREM 7. *Suppose that X is finite and $\alpha = 0$. Then*

$$\Pr\left\{X_0^N \subseteq X_\epsilon\right\} \geq 1 - |X \setminus X_\epsilon|(1-\epsilon)^N.$$

Proof. With $\alpha = 0$, if $x \in X$ satisfies $\Pr\{G(x,\xi) \leq \mathbf{0}\} < 1-\epsilon$, then $x \in X_0^N$ if and only if $G(x,\xi^i) \leq \mathbf{0}$ for each $i = 1,\ldots,N$, and hence $\Pr\{x \in X_0^N\} < (1-\epsilon)^N$. The claim then follows just as in the proof of Theorem 5. $\square$

When $\alpha = 0$, to obtain confidence $1-\delta$ that (P_α^N) will yield a feasible solution to (P_ϵ) whenever (P_α^N) is feasible, we should take

$$N \geq \log^{-1}\left(\frac{1}{1-\epsilon}\right) \log\left(\frac{|X \setminus X_\epsilon|}{\delta}\right).$$

If $|X| \leq U^n$, then it is sufficient to take

$$N \geq \frac{1}{\epsilon}\log\left(\frac{1}{\delta}\right) + \frac{n}{\epsilon}\log U, \tag{7}$$

where we have used the inequality $\log(1/(1-\epsilon)) \geq \epsilon$. Hence, with $\alpha = 0$, the required sample size again grows linearly in n but now also linearly with $1/\epsilon$. Note the similarity between the bound (7) and the bound of Campi and Calafiore [8]

$$N \geq \frac{2}{\epsilon}\log\left(\frac{1}{\delta}\right) + 2n + \frac{2n}{\epsilon}\log\left(\frac{2}{\epsilon}\right),$$

which also exhibits linear dependence in n and (nearly) linear dependence in $1/\epsilon$. This is interesting considering the significantly different assumptions used for the analysis. In [8] it is assumed that X is a convex set and $G(x,\xi)$ is a convex function of x for every possible value of ξ. In contrast, we make the strong assumption that X is finite but require no other assumptions on the form of the random constraint $G(x,\xi) \le \mathbf{0}$.

2.2.2. Random right-hand side. We now show how the assumption that X is finite can be relaxed when the probabilistic constraint involves randomness only in the right-hand side. Thus, in this section we assume that $G(x,\xi) = \xi - g(x)$, where $g : \mathbf{R}^n \to \mathbf{R}^m$, and $\Xi \subseteq \mathbf{R}^m$. Let the cumulative distribution function of ξ be $F(y) = \Pr\{\xi \le y\}$ for $y \in \mathbf{R}^m$. Then the feasible region of the probabilistically constrained problem with a random right-hand side is

$$\bar{X}_\epsilon = \Big\{ x \in X : F(g(x)) \ge 1 - \epsilon \Big\}.$$

The feasible region of the sample approximation problem for $\alpha \in [0,1)$ is

$$\bar{X}_\alpha^N = \left\{ x \in X : \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\big(g(x) \ge \xi^i\big) \ge 1 - \alpha \right\}.$$

We first consider the case that ξ has a finite distribution, that is, $\Xi = \{\xi^1, \ldots, \xi^K\}$. Note that K may be very large, for example, $K = U^m$ for a positive integer U. Next, for $j = 1, \ldots, m$ define $\Xi_j = \{\xi_j^k : k = 1, \ldots, K\}$, and finally let $C = \prod_{j=1}^m \Xi_j$.

THEOREM 8. *Suppose that ξ has a finite distribution, and let $\alpha \in [0,\epsilon)$. Then*

$$\Pr\big\{\bar{X}_\alpha^N \subseteq \bar{X}_\epsilon\big\} \ge 1 - |C| \exp\big\{-2N(\epsilon-\alpha)^2\big\}.$$

Proof. Let $C_\epsilon = \{y \in C : F(y) \ge 1 - \epsilon\}$ and

$$C_\alpha^N = \left\{ y \in C : \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\big(y \ge \xi^i\big) \ge 1 - \alpha \right\}.$$

Because C is a finite set, we can apply Theorem 5 to obtain

$$\Pr\big\{C_\alpha^N \subseteq C_\epsilon\big\} \ge 1 - |C| \exp\big\{-2N(\epsilon-\alpha)^2\big\}. \tag{8}$$

Now, let $x \in \bar{X}_\alpha^N$, so that $x \in X$ and $\sum_{i=1}^N \mathbb{I}(g(x) \ge \xi^i) \ge N(1-\alpha)$. Define $\bar{y} \in C$ by

$$\bar{y}_j = \max\{y_j \in \Xi_j : y_j \le g_j(x)\}, \qquad j = 1, \ldots, m,$$

so that by definition $\bar{y} \le g(x)$. Next, note that if $g(x) \ge \xi^i$ for some i, then also $\bar{y} \ge \xi^i$ since $\xi^i \in C$. Hence, $\sum_{i=1}^N \mathbb{I}(\bar{y} \ge \xi^i) \ge N(1-\alpha)$ and so $\bar{y} \in C_\alpha^N$. Hence, when $C_\alpha^N \subseteq C_\epsilon$, $F(\bar{y}) \ge 1 - \epsilon$, and, because $\bar{y} \le g(x)$, also $F(g(x)) \ge 1 - \epsilon$ and so $x \in \bar{X}_\epsilon$. Since $x \in \bar{X}_\alpha^N$ was arbitrary, this shows that, when $C_\alpha^N \subseteq C_\epsilon$, $\bar{X}_\alpha^N \subseteq \bar{X}_\epsilon$, and the result follows from (8). ∎

If, for example, $|\Xi_j| \le U$ for each j, then $|C| \le U^m$, so to obtain confidence $1 - \delta$ that $\bar{X}_\alpha^N \subseteq \bar{X}_\epsilon$ it is sufficient to take

$$N \ge \frac{1}{2(\epsilon-\alpha)^2} \log\left(\frac{1}{\delta}\right) + \frac{m}{2(\epsilon-\alpha)^2} \log U. \tag{9}$$

The difference between this bound and (6) is that (9) depends linearly on m, the dimension of ξ, whereas (6) depends linearly on n, the dimension of x.

Similarly to the case of finite feasible region X, when ξ has a finite distribution, it can be shown that the sample approximation problem with $\epsilon = \alpha$ will yield an exact optimal solution with probability approaching one as N increases. The statement and proof of this result are completely analogous to those of Corollary 6 and are omitted for the sake of brevity.

As in the case of Theorem 7, if we take $\alpha = 0$, we can obtain the stronger convergence result

$$\Pr\left\{\bar{X}_0^N \subseteq \bar{X}_\epsilon\right\} \geq 1 - |C|(1-\epsilon)^N.$$

The assumption in Theorem 8 that Ξ is finite can be relaxed if we assume that $\bar{X}_\epsilon \subseteq \bar{X}(l,u) := \{x \in X : l \leq g(x) \leq u\}$ for some $l, u \in \mathbf{R}^m$. This assumption is not very strict. Indeed, if we define $l \in \mathbf{R}^m$ by

$$l_j = \min\{l \in \mathbf{R} : F_j(l) \geq 1 - \epsilon\},$$

where F_j is the marginal distribution of ξ_j for $j = 1, \ldots, m$, then $g(x) \geq l$ for any $x \in \bar{X}_\epsilon$. This holds because if $g_j(x) < l_j$ for some j, then $\Pr\{g(x) \geq \xi\} \leq \Pr\{g_j(x) \geq \xi_j\} = F_j(g_j(x)) < 1 - \epsilon$ by the definition of l_j and hence $x \notin \bar{X}_\epsilon$. Furthermore, if X is compact and $g(x)$ is continuous in x, then if we define $u \in \mathbf{R}^m$ by

$$u_j = \max\{g_j(x) : x \in X\}, \qquad j = 1, \ldots, m,$$

each u_j is finite, and, by definition, $g(x) \leq u$ for any $x \in \bar{X}$. Under the assumption that $\bar{X}_\epsilon \subseteq \bar{X}(l,u)$ the assumption that Ξ is finite can be replaced by the assumption that $\Xi \cap \{y \in \mathbf{R}^m : l \leq y \leq u\}$ is finite, leading to a result similar to Theorem 8, with a nearly identical proof.

Alternatively, when $\bar{X}_\epsilon \subseteq \bar{X}(l,u)$, we can obtain a similar result if ξ has a Lipschitz continuous cumulative distribution function F on $[l,u] = \{y \in \mathbf{R}^m : l \leq y \leq u\}$. That is, we assume that there exists $L > 0$ such that

$$|F(y) - F(y')| \leq L\|y - y'\|_\infty \qquad \forall y, y' \in [l,u],$$

where $\|y\|_\infty = \max\{|y_j| : j = 1, \ldots, m\}$. Under the assumption that $\bar{X}_\epsilon \subseteq \bar{X}(l,u)$ we add the constraints $l \leq g(x) \leq u$ to the sample approximation problem to obtain

$$\bar{X}_\alpha^N(l,u) = \left\{ x \in \bar{X}(l,u) : \frac{1}{N}\sum_{i=1}^N \mathbb{I}\big(g(x) \geq \xi^i\big) \geq 1 - \alpha \right\}.$$

We define $D = \max\{u_j - l_j : j = 1, \ldots, m\}$. Then we have the following.

THEOREM 9. *Suppose that $\bar{X}_\epsilon \subseteq \bar{X}(l,u)$ and F is Lipschitz continuous with constant L. Let $\alpha \in [0, \epsilon)$ and $\beta \in (0, \epsilon - \alpha)$. Then*

$$\Pr\left\{\bar{X}_\alpha^N(l,u) \subseteq \bar{X}_\epsilon\right\} \geq 1 - \lceil DL/\beta \rceil^m \exp\left\{-2N(\epsilon - \alpha - \beta)^2\right\}.$$

Proof. Let $K = \lceil DL/\beta \rceil$, and define $Y_j = \{l_j + (u_j - l_j)i/K : i = 1, \ldots, K\}$ for $j = 1, \ldots, m$ and $Y = \prod_{j=1}^m Y_j$, so that $|Y| = K^m$ and that for any $y \in [l,u]$ there exists $y' \in Y$ such that $y' \geq y$ and $\|y - y'\|_\infty \leq \beta/L$. Indeed, for a given $y \in [l,u]$ such a y' can be obtained by letting

$$y'_j = \min\{w \in Y_j : w \geq y_j\}, \qquad j = 1, \ldots, m.$$

With this definition of y', we have $y' \geq y$ and

$$|y'_j - y_j| = y'_j - y_j \leq (u_j - l_j)/K \leq D/K \leq \beta/L, \qquad j = 1, \ldots, m.$$

Next, let $Y_{\epsilon-\beta} = \{y \in Y : F(y) \geq 1 - \epsilon + \beta\}$ and

$$Y_\alpha^N = \left\{ y \in Y : \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\big(y \geq \xi^i\big) \geq 1 - \alpha \right\}. \tag{10}$$

Since Y is finite and $\alpha < \epsilon - \beta$, we can apply Theorem 5 to obtain

$$\Pr\left\{Y_\alpha^N \subseteq Y_{\epsilon-\beta}\right\} \geq 1 - |Y| \exp\left\{-2N(\epsilon - \alpha - \beta)^2\right\}.$$

Now, let $x \in \bar{X}_\alpha^N(l, u)$, and let $y' \in Y$ be such that $y' \geq g(x)$ and $\|y' - g(x)\|_\infty \leq \beta/L$. By Lipschitz continuity of F, this implies that

$$F(y') - F(g(x)) \leq L\|y' - g(x)\|_\infty \leq \beta. \tag{11}$$

Because x satisfies $\sum_{i=1}^N \mathbb{I}(g(x) \geq \xi^i) \geq N(1-\alpha)$ and $y' \geq g(x)$, we have $\sum_{i=1}^N \mathbb{I}(y' \geq \xi^i) \geq N(1-\alpha)$ and hence $y' \in Y_\alpha^N$. Thus, by using (11), when $Y_\alpha^N \subseteq Y_{\epsilon-\beta}$ occurs,

$$F(g(x)) \geq F(y') - \beta \geq (1 - \epsilon + \beta) - \beta = 1 - \epsilon.$$

Since $x \in \bar{X}_\alpha^N(l, u)$ was arbitrary, $Y_\alpha^N \subseteq Y_{\epsilon-\beta}$ implies that $\bar{X}_\alpha^N(l, u) \subseteq \bar{X}_\epsilon$, and the result follows from (10). □

To obtain a confidence of at least $1 - \delta$ that $\bar{X}_\alpha^N(l, u) \subseteq \bar{X}_\epsilon$, it is sufficient to take

$$N \geq \frac{1}{2(\epsilon - \alpha - \beta)^2} \log\left(\frac{1}{\delta}\right) + \frac{m}{2(\epsilon - \alpha - \beta)^2} \log\left\lceil \frac{DL}{\beta} \right\rceil.$$

Note that for fixed $\epsilon > 0$ and $\alpha \in [0, \epsilon)$, β is a free parameter which can be chosen in $(0, \epsilon - \alpha)$. If, for example, we take $\beta = (\epsilon - \alpha)/2$, we obtain

$$N \geq \frac{2}{(\epsilon - \alpha)^2} \log\left(\frac{1}{\delta}\right) + \frac{2m}{(\epsilon - \alpha)^2} \log\left\lceil \frac{2DL}{\epsilon - \alpha} \right\rceil.$$

Once again, if $\alpha = 0$, similar arguments can be used to conclude that if

$$N \geq \frac{2}{\epsilon} \log\left(\frac{1}{\delta}\right) + \frac{2m}{\epsilon} \log\left\lceil \frac{2DL}{\epsilon} \right\rceil,$$

then $\Pr\{\bar{X}_0^N(l, u) \subseteq \bar{X}_\epsilon\} \geq 1 - \delta$.

2.2.3. Lipschitz continuous *G*. We now turn to the problem of using a sample approximation problem to generate feasible solutions to (P_ϵ) when X is not necessarily finite and $G(x, \xi)$ does not necessarily have the form $G(x, \xi) = g(x) - \xi$. In this section, we assume for simplicity of exposition that G takes values in $\mathbf{R}$. This is without loss of generality, since if $\bar{G} : \mathbf{R}^n \times \mathbf{R}^d \to \mathbf{R}^m$, we can define $G : \mathbf{R}^n \times \mathbf{R}^d \to \mathbf{R}$ by $G(x, \xi) = \max\{\bar{G}_j(x, \xi) : j = 1, \ldots, m\}$ and the constraints $G(x, \xi) \leq 0$ and $\bar{G}(x, \xi) \leq \mathbf{0}$ are equivalent. In this section, we shall make the following Lipschitz continuity assumption on G.

Assumption 1. There exists $L > 0$ such that

$$|G(x, \xi) - G(x', \xi)| \leq L\|x - x'\|_\infty \qquad \forall x, x' \in X \text{ and } \forall \xi \in \Xi.$$

It is important that the Lipschitz constant L is independent of $\xi \in \Xi$, and this condition may make Assumption 1 appear rather stringent. There are, however, interesting cases in which the assumption does hold. For example, if Ξ is finite (with possibly huge cardinality) and $G(x,\xi)$ is Lipschitz continuous with Lipschitz constant $L(\xi)$ for each $\xi \in \Xi$, then Assumption 1 holds with $L = \max\{L(\xi) : \xi \in \Xi\}$. Alternatively, if Ξ is compact, $G(x,\xi) = \max\{T_j(\xi)x : j = 1,\dots,m\}$, and $T_j : \Xi \to \mathbf{R}^n$ is continuous in ξ for each j, then Assumption 1 holds with

$$L = \sup_{\xi\in\Xi}\big\{\max\{\|T_j(\xi)\|_\infty : j = 1,\dots,m\}\big\}.$$

To generate feasible solutions for this general case, we will also need to modify the sample approximation problem somewhat. In addition to taking a risk level α less than the nominal risk level ϵ, we will require that at least $(1-\alpha)N$ of the constraints be satisfied *strictly*. That is, for a fixed $\gamma > 0$, we define the sample approximation feasible region to be

$$X^N_{\alpha,\gamma} = \left\{ x \in X : \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\big(G(x,\xi) + \gamma \le 0\big) \ge 1 - \alpha \right\}.$$

Finally, we will assume that X is bounded and let $D = \sup\{\|x - x'\|_\infty : x, x' \in X\}$ be the diameter of X.

THEOREM 10. *Suppose that X is bounded with diameter D and Assumption 1 holds. Let $\alpha \in [0,\epsilon)$, $\beta \in (0, \epsilon-\alpha)$, and $\gamma > 0$. Then*

$$\Pr\big\{X^N_{\alpha,\gamma} \subseteq X_\epsilon\big\} \ge 1 - \lceil 1/\beta \rceil \lceil 2LD/\gamma \rceil^n \exp\big\{-2N(\epsilon - \alpha - \beta)^2\big\}.$$

Proof. For $x \in X$, let $\phi(x) = \Pr\{G(x,\xi) \le 0\}$. Let $J = \lceil 1/\beta \rceil$, for $j = 1,\dots,J-1$, define

$$X_j = \left\{ x \in X : \frac{j-1}{J} \le \phi(x) < \frac{j}{J} \right\},$$

and let $X_J = \{x \in X : (J-1)/J \le \phi(x) \le 1\}$. Next, we claim that for each j there exists a finite set $Z^\gamma_j \subseteq X_j$ such that $|Z^\gamma_j| \le \lceil 2LD/\gamma \rceil^n$ and for all $x \in X_j$ there exists $z \in Z^\gamma_j$ such that $\|x - z\|_\infty \le \gamma/L$. Indeed, because $X_j \subseteq X$ and X is bounded with diameter D, there exists a finite set $Y \subseteq \mathbf{R}^n$ with $|Y| \le \lceil 2LD/\gamma \rceil^n$ such that for all $x \in X$ there exists $y \in Y$ such that $\|x - y\|_\infty \le \gamma/2L$. For any $y \in \mathbf{R}^n$ and $\eta > 0$, define $B(y,\eta) = \{x \in \mathbf{R}^N : \|y - x\|_\infty \le \eta\}$. Now, let $Y'_j = \{y \in Y : X_j \cap B(y, \gamma/2L) \ne \emptyset\}$, and for $y \in Y'_j$ select an arbitrary $x_y \in X_j \cap B(y,\gamma/2L)$. Then let $Z^\gamma_j = \bigcup_{y\in Y'_j} x_y$. By definition, $Z^\gamma_j \subseteq X_j$ and $|Z^\gamma_j| \le \lceil 2LD/\gamma \rceil^n$. In addition, for any $x \in X_j$, there exists y such that $x \in B(y,\gamma/2L)$, and, because for this y, $X_j \cap B(y,\gamma/2L) \ne \emptyset$, there exists $x_y \in Z^\gamma_j$ such that $\|x_y - y\|_\infty \le \gamma/2L$. Hence,

$$\|x_y - x\|_\infty \le \|x_y - y\|_\infty + \|y - x\|_\infty \le \gamma/L.$$

Now define $Z^\gamma = \bigcup_{j=1}^J Z^\gamma_j$, and observe that $|Z^\gamma| \le J\lceil 2LD/\gamma \rceil^n$. Next, define $Z^\gamma_{\epsilon-\beta} = \big\{x \in Z^\gamma : \Pr\{G(x,\xi) \le 0\} \ge 1 - \epsilon + \beta\big\}$ and

$$Z^{\gamma,N}_\alpha = \left\{ x \in Z^\gamma : \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\big(G(x,\xi^i) \le 0\big) \right\}.$$

Since Z^γ is finite and $\alpha < \epsilon - \beta$, we can apply Theorem 5 to obtain

$$\Pr\left\{Z_\alpha^{\gamma,N} \subseteq Z_{\epsilon-\beta}^\gamma\right\} \geq 1 - \lceil 1/\beta \rceil \lceil 2LD/\gamma \rceil^n \exp\{-2N(\epsilon - \alpha - \beta)^2\}. \tag{12}$$

Now consider an arbitrary $x \in X_{\alpha,\gamma}^N$. Let $j \in \{1, \ldots, J\}$ be such that $x \in X_j$. By the definition of Z_j^γ there exists $z \in Z_j^\gamma$ such that $\|x - z\|_\infty \leq \gamma/L$. By the definition of X_j and because $Z_j^\gamma \subseteq X_j$, we have $|\phi(x) - \phi(z)| \leq \beta$. In addition, Assumption 1 implies that $|G(x, \xi^i) - G(z, \xi^i)| \leq \gamma$. Hence, if $G(x, \xi^i) + \gamma \leq 0$, then $G(z, \xi^i) \leq 0$, and, because x satisfies $\sum_{i=1}^N \mathbb{I}(G(x, \xi^i) + \gamma \leq 0) \geq N(1-\alpha)$, it follows that z satisfies $\sum_{i=1}^N \mathbb{I}(G(z, \xi^i) \leq 0) \geq N(1 - \alpha)$. Thus $z \in Z_\alpha^{\gamma,N}$, and so if $Z_\alpha^{\gamma,N} \subseteq Z_{\epsilon-\beta}^\gamma$, then $\phi(z) \geq 1 - \epsilon + \beta$. Thus, $\phi(x) \geq \phi(z) - \beta \geq 1 - \epsilon$ when $Z_\alpha^{\gamma,N} \subseteq Z_{\epsilon-\beta}^\gamma$. Since $x \in X_{\alpha,\gamma}^N$ was arbitrary, $Z_\alpha^{\gamma,N} \subseteq Z_{\epsilon-\beta}^\gamma$ implies that $X_{\alpha,\gamma}^N \subseteq X_\epsilon$, and the result follows from (12). □

Once again, for fixed ϵ and $\alpha < \epsilon$, β is a free parameter to be chosen in $(0, \epsilon - \alpha)$. If we choose, for example, $\beta = (\epsilon - \alpha)/2$, then we can assure that $X_{\alpha,\gamma}^N \subseteq X_\epsilon$ with confidence at least $1 - \delta$ by taking

$$N \geq \frac{2}{(\epsilon - \alpha)^2}\left[\log\left(\frac{1}{\delta}\right) + n \log\left\lceil \frac{2LD}{\gamma} \right\rceil + \log\left\lceil \frac{2}{\epsilon - \alpha} \right\rceil\right].$$

Additionally, if $\alpha = 0$, similar arguments show that $X_{0,\gamma}^N \subseteq X_\epsilon$ occurs with probability at least $1 - \delta$ if

$$N \geq \frac{2}{\epsilon}\left[\log\left(\frac{1}{\delta}\right) + n \log\left\lceil \frac{2LD}{\gamma} \right\rceil + \log\left\lceil \frac{2}{\epsilon} \right\rceil\right].$$

Regardless of whether $\alpha = 0$ or $\alpha > 0$, the term $1/\gamma$ is taken under log, and hence γ can be made very small without significantly increasing the required sample size, suggesting that modifying the sample approximation problem to require at least $(1 - \alpha)N$ of the sampled constraints to be satisfied with a slack of at least γ need not significantly alter the feasible region.

2.2.4. A posteriori feasibility checking. The results of sections 2.2.1–2.2.3 demonstrate that, with appropriately constructed sample approximation problems, the probability that the resulting feasible region will be a subset of the true feasible region X_ϵ approaches one exponentially fast. This gives strong theoretical support for using these sample approximations to yield solutions feasible to X_ϵ. These results yield *a priori* estimates on how large the sample size N should be to have high confidence that the sample approximation feasible region will be a subset of X_ϵ. However, these a priori estimates are likely to yield required sample sizes which are very large, and hence the sample approximation problems will still be impractical to solve. This is particularly true if $\alpha > 0$ and $\epsilon - \alpha$ is small. However, typically in sampling approximation results such as these, the a priori estimates of the required sample size are very conservative, and in fact much smaller sample sizes are sufficient. See [18] for a computational demonstration of this phenomenon for the case of SAA applied to two-stage stochastic linear programs. Thus, a natural alternative to using the sample size suggested by the a priori estimates is to solve a sample approximation problem with a smaller sample to yield a candidate solution $\hat{x} \in X$ and then conduct an *a posteriori* check to see whether $\Pr\{G(\hat{x}, \xi) \leq \mathbf{0}\} \geq 1 - \epsilon$. A simple method for conducting an a posteriori analysis of the risk of a candidate solution is to take a single

very large Monte Carlo sample $\xi^1, \ldots, \xi^{N'}$ and count how many times $G(\hat{x}, \xi^i) \leq \mathbf{0}$ holds. Bounds on the true risk $\Pr\{G(\hat{x}, \xi) \leq \mathbf{0}\}$ which hold with high confidence can then be constructed, and, if N' is very large, these bounds should be tight. This approach will not work well if the allowed risk ϵ is extremely small, but, on the other hand, we do not expect the sample approximation approach to be practical in this case anyway. Of course, if good estimates of $\Pr\{G(\hat{x}, \xi) \leq \mathbf{0}\}$ can be obtained efficiently by some other method, then this other method should be used for a posteriori feasibility checking. For example, if $G(x, \xi) = \xi - g(x)$ and the components of ξ are independent, then $\Pr\{g(x) \geq \xi\}$ can be calculated as $\prod_i \Pr\{g_i(x) \geq \xi_i\}$.

3. Numerical experiments. We conducted experiments to test the effectiveness of the sample approximation approach for yielding good feasible solutions and lower bounds. In particular, our aim is to determine whether using $\alpha > 0$ in the sample approximation can yield better solutions than when using $\alpha = 0$ as in the scenario approximation approach of [7, 22]. In addition, we test whether reasonable lower bounds which are valid with high probability can be obtained. We first conducted tests on a probabilistic version of the classical set covering problem, which has been studied recently in [5, 26, 27]. This problem has both a finite feasible region and finite distribution (although both are exponentially large) so that, for generating feasible solutions, the stronger Theorems 5 and 8 apply. These results are given in section 3.1. We also conducted tests on a probabilistic version of the transportation problem. For this problem, the feasible region is continuous, and we also use a joint normal distribution for the right-hand side vector, so that Theorem 9 applies. These results are presented in section 3.2.

Note that, although Theorem 3 provides support for using the sample approximation scheme to generate lower bounds, we will use Theorem 4 to actually obtain lower bounds which are valid with high confidence, because it can be used regardless of how large the sample size N is (with the possible drawback that using smaller N will yield weaker lower bounds). Similarly, Theorems 5, 8, and 9 support the use of sample approximation to yield feasible solutions, but we do not use these theorems to guide our choice of α and N. Indeed, the bounds implied by these theorems would suggest using N which is far too large to be able to solve the approximation problem. Instead, we experiment with different values of α and N and perform an a posteriori test on each solution generated to determine whether it is feasible (with high confidence).

3.1. Probabilistic set cover problem. The probabilistic set cover problem is given by

$$\text{(PSC)} \qquad \min\bigl\{cx : \Pr\{Ax \geq \xi\} \geq 1 - \epsilon, x \in \{0,1\}^n\bigr\},$$

where $c \in \mathbf{R}^n$ is the cost vector, A is an $m \times n$ zero-one matrix, and ξ is a random vector taking values in $\{0,1\}^m$. We conducted tests on a single instance of (PSC), with two values of ϵ: 0.05 and 0.1.

3.1.1. Test instance. Following [5], we based our tests on a deterministic set-covering instance, scp41, of the OR library [3], which has $m = 200$ rows and $n = 1000$ columns. Also following [5], the random vector ξ is assumed to consist of 20 independent subvectors, with each subvector having size $k = 10$ following the *circular* distribution. The circular distribution is defined by parameters $\lambda_j \in [0,1]$ for $j = 1, \ldots, k$. First, Bernoulli random variables Y_j for $j = 1, \ldots, k$ are generated independently, with $\Pr\{Y_j = 1\} = \lambda_i$. Then the random subvector is defined by $\xi_j = \max\{Y_j, Y_{j+1}\}$

for $j < k$ and by $\xi_k = \max\{Y_1, Y_k\}$. Because of the simple form of this distribution, given a solution x, it is possible to calculate exactly $\Pr\{Ax \geq \xi\}$. Thus, when a solution is obtained from a sample approximation problem, we test a posteriori whether it is feasible at a given risk level by exactly calculating $\Pr\{Ax \geq \xi\}$. To illustrate this calculation, we show how to calculate the probability for a single subvector, that is, $\Pr\{\xi_j \leq y_j,\ j = 1, \ldots, k\}$. Then, with $y = Ax$, the overall probability $\Pr\{Ax \geq \xi\}$ is calculated as the product of the probabilities for each subvector. Let $J = \{1 \leq j \leq k : y_j = 0\}$. Then

$$\Pr\{\xi_j \leq y_j,\ j = 1, \ldots, k\} = \Pr\{\xi_j = 0, j \in J\} = \Pr\left\{Y_j = 0, j \in J^+\right\} = \prod_{j \in J^+} (1 - \lambda_j),$$

where $J^+ = \cup_{j \in J}\{j, (j+1) \mod k\}$. Although in this test calculation of the distribution function is easy, we stress that this is not a necessary condition to use the sample approximation; it is necessary only that sampling from the distribution can be done efficiently.

3.1.2. Solving the sample approximation. To solve the sample approximation of problem (PSC), we used a mixed-integer program (MIP) formulation which is equivalent to an extended formulation studied in [20] (see also [19]). The formulation is not exactly the same, since, because the random right-hand side can take on only two values, it can be simplified somewhat. Let the scenarios obtained in the sample of size N be denoted by ξ^i for $i = 1, \ldots, N$, where each $\xi^i \in \{0,1\}^m$. Then the formulation we use is

$$\begin{aligned}
\min \quad & cx \\
\text{subject to } & Ax \geq y, \\
& y_j + z_i \geq 1 \qquad \forall i, j \text{ s.t. } \xi^i_j = 1, \qquad (13) \\
& \sum_{i=1}^{N} z_i \leq p, \qquad (14) \\
& x \in \{0,1\}^n,\ z \in \{0,1\}^N,\ y \in \{0,1\}^m,
\end{aligned}$$

where $p = \lfloor \alpha N \rfloor$. We could relax the intregrality restriction on the y variables, but we found that leaving this restriction and also placing a higher branching priority on these variables significantly improved performance when solving with CPLEX 9.0. The intuition behind this is that if we fix $y_j = 1$, then we are enforcing the constraint $A^j x \geq 1$, and, on the other hand, if we fix $y_j = 0$, then any scenario i for which $\xi^i_j = 1$ will be fixed to 1, and constraint (14) will quickly become binding. We also found that some simple preprocessing of the formulation significantly helped solution times. If, for a row j, $\sum_i \xi^i_j > p$, then we cannot have $y_j = 0$, and so we fixed $y_j = 1$, and the corresponding inequalities (13) for j were not included. After this preprocessing, for each j there will be at most p inequalities in (13), so that these inequalities add at most mp rows and $O(mp)$ nonzeros to the formulation. By using this formulation, we found that the sample approximation problems could be solved quickly, in all cases in less than ten seconds and usually much less. However, this may be due to the particular distribution used (and the simplicity of the underlying set cover instance), and thus this should not be taken as a study of the effectiveness of this formulation in general. Rather, we are interested here only in the properties of the solutions generated by the sample approximation problems.

TABLE 1
Solution results for (PSC) *sample problems with* $\epsilon = 0.05$.

		Solution risk				Feasible solutions cost				
α	N	Ave	Min	Max	σ	#	Ave	Min	Max	σ
0.00	100	0.107	0.048	0.185	0.042	1	425.0	425	425	***
	110	0.071	0.013	0.100	0.029	3	425.7	424	429	2.9
	120	0.069	0.013	0.152	0.049	4	424.8	424	427	1.5
	130	0.062	0.020	0.124	0.036	5	424.8	420	429	4.3
	140	0.042	0.018	0.080	0.017	8	425.6	421	429	2.8
	150	0.041	0.005	0.080	0.026	6	427.3	421	429	3.1
0.05	1000	0.056	0.041	0.072	0.009	2	414.0	414	414	0.0
	3000	0.044	0.041	0.055	0.005	8	414.0	414	414	0.0
	5000	0.044	0.041	0.060	0.006	8	414.0	414	414	0.0
	7500	0.041	0.041	0.041	0.000	10	414.0	414	414	0.0
	10000	0.044	0.041	0.054	0.005	8	414.0	414	414	0.0

TABLE 2
Solution results for (PSC) *sample problems with* $\epsilon = 0.1$.

		Solution risk				Feasible solutions cost				
α	N	Ave	Min	Max	σ	#	Ave	Min	Max	σ
0.0	80	0.203	0.095	0.311	0.076	1	420.0	420	420	***
	90	0.169	0.084	0.239	0.051	1	428.0	428	428	***
	100	0.107	0.048	0.185	0.042	4	426.0	423	428	500.7
	110	0.071	0.013	0.100	0.029	9	425.4	421	429	499.8
	120	0.069	0.013	0.152	0.049	7	424.6	419	428	534.3
	130	0.062	0.020	0.124	0.036	7	425.3	420	429	488.8
0.1	1000	0.111	0.095	0.141	0.015	4	401.3	400	403	1.5
	3000	0.101	0.092	0.115	0.006	6	401.0	400	402	1.1
	5000	0.101	0.092	0.108	0.005	5	401.2	400	402	1.1
	7500	0.099	0.092	0.105	0.004	7	401.1	400	402	1.1
	10000	0.097	0.088	0.103	0.004	8	401.8	400	404	1.3

3.1.3. Feasible solutions. We first tested the effectiveness of the sample approximation approach for generating feasible solutions. To do so, we varied the risk level of the approximation problem α and the sample size N. For each combination of α and N, we generated and solved 10 sample approximation problems. Table 1 gives statistics of the solutions generated for the (PSC) instance with $\epsilon = 0.05$, and Table 2 gives the same for the (PSC) instance with $\epsilon = 0.1$. For each combination of α and N, we report statistics on the risk of the generated solutions, where for a solution x the risk is $\Pr\{Ax \not\geq \xi\}$, as well as on the costs of the *feasible* solutions generated, i.e., those solutions which have risk less than 0.05 and 0.1, respectively. For the risk of the solutions, we report the average, minimum, maximum, and sample standard deviation over the 10 solutions. For the solution costs, we report first how many solutions were feasible, then report the average, minimum, maximum, and sample standard deviation of the cost taken over these solutions.

We first discuss results for the case of nominal risk level $\epsilon = 0.05$. When using $\alpha = 0$, the best results were obtained with N in the range of 100–150, and these are the results that we report. With $\alpha = 0$, as N increases, more constraints are being enforced, which leads to a smaller feasible region of the approximation and a higher likelihood that the optimal solution of the approximation is feasible at the nominal risk level. However, the smaller feasible region also causes the cost to increase, so that increasing N more would yield overly conservative solutions. We also conducted tests with $\alpha = 0.05$, and for this value of α we used significantly larger sample sizes.

TABLE 3
Lower bounds (LB) *for* (PSC) *sample problems with* $\alpha = \epsilon = 0.05$.

	LB with confidence at least:				Gap with confidence at least:			
N	0.999	0.989	0.945	0.828	0.999	0.989	0.945	0.828
1000	412	414	414	414	0.5%	0.0%	0.0%	0.0%
3000	412	414	414	414	0.5%	0.0%	0.0%	0.0%
5000	412	414	414	414	0.5%	0.0%	0.0%	0.0%
7500	414	414	414	414	0.0%	0.0%	0.0%	0.0%
10000	413	414	414	414	0.2%	0.0%	0.0%	0.0%

The best feasible solution found by using $\alpha = 0$ had cost 420, and the average cost of the feasible solutions found was significantly greater than this. When $\alpha = 0.05$, every sample size N yielded at least one feasible solution in the ten runs, and *every* feasible solution found had cost 414. Thus, using $\alpha = 0.05$ consistently yields solutions which are closer to the efficient frontier between the objectives of risk and cost.

For $\epsilon = 0.1$, we observed similar results. In this case, when using $\alpha = 0$, the best results were obtained with N in the range of 80–130. The best solution found by using $\alpha = 0$ had cost 419, whereas the best solution found by using $\alpha = 0.1$ was 400, which was obtained by one of the ten runs for *every* sample size N. In addition, observe from Table 1 that using $\alpha = 0.05$ yields solutions with a risk not exceeding 0.05 and a cost of 414, which is also less than the cost of the best solution found that had a risk not exceeding 0.1 when using $\alpha = 0$. Thus, by using $\alpha > 0$ we are able to get solutions with lower risk *and* lower cost as compared to those obtained when using $\alpha = 0$.

In terms of the variability of the risks and costs of the solutions generated, using $\alpha > 0$ and a much larger sample size yielded solutions with much lower variability than when using $\alpha = 0$ and small sample size. This is not surprising since using a larger sample size naturally should reduce variability. On the other hand, constraining the sample approximation to have $\alpha = 0$ prohibits the use of a larger sample size, as the solutions produced then become overly conservative.

3.1.4. Lower bounds. We next discuss the results for obtaining lower bounds for (PSC). We used the procedure of Theorem 4 with $\alpha = \epsilon$ and $M = 10$. We use the same 10 sample approximation problems as when generating feasible solutions. As argued after Theorem 4, with $\alpha = \epsilon$, we have $\rho(\alpha, \epsilon, N) = \rho(\epsilon, \epsilon, N) \gtrsim 1/2$. Then, if we take $L = 1$, the test of Theorem 4 yields a lower bound with confidence 0.999. Taking $L = 1$ corresponds to taking the minimum optimal value over all of the $M = 10$ runs (not just over the ones which yielded feasible solutions). More generally, we can take $L \in \{1, \ldots, 10\}$ yielding a lower bound with confidence at least

$$1 - \sum_{i=0}^{L-1} \binom{10}{i} \rho(\epsilon, \epsilon, N)^i \big(1 - \rho(\epsilon, \epsilon, N)\big)^{10-i} \gtrsim 1 - \sum_{i=0}^{L-1} \binom{10}{i} (1/2)^{10}$$

to obtain possibly "tighter" lower bounds of which we are less confident.

The results obtained by using varying values of N and $\epsilon = \alpha = 0.05$ are given in Table 3. The gaps reported are the percent by which the lower bound is below the best feasible solution (414, obtained with $\alpha = 0.05$ and any of the tested sample sizes N). Thus, for example, by solving 10 problems with sample size $N = 1000$, we obtained a feasible solution of cost 414 and a lower bound of 412, which is valid with probability at least 0.999. In addition, we obtain a lower bound of 414 which is valid

TABLE 4
Lower bounds for (PSC) *sample problems with* $\alpha = \epsilon = 0.1$.

	LB with confidence at least:				Gap with confidence at least:			
N	0.999	0.989	0.945	0.828	0.999	0.989	0.945	0.828
1000	397	397	398	398	0.8%	0.8%	0.5%	0.5%
3000	399	400	400	400	0.3%	0.0%	0.0%	0.0%
5000	400	400	400	400	0.0%	0.0%	0.0%	0.0%
7500	400	400	400	400	0.0%	0.0%	0.0%	0.0%
10000	400	400	400	400	0.0%	0.0%	0.0%	0.0%

with probability of at least 0.989. Thus, we have confidence at least 0.989 that 414 is the optimal value. Similar results were obtained with larger sample sizes.

Table 4 yields the lower bound results obtained with $\epsilon = \alpha = 0.1$ and varying sample size N. By solving 10 sample problems with $N = 1000$, we obtained a feasible solution of cost 400 and can say with confidence 0.999 that the optimal solution is at most 0.8% less costly than this solution. By using $N = 5000$ (or greater), we obtain a feasible solution of the same cost but a lower bound which states that with confidence at least 0.999 this feasible solution is optimal.

3.2. Probabilistic transportation problem. We next tested the sampling approach on a probabilistic version of the classical transportation problem, which we call the probabilistic transportation problem (PTP). In this problem, we have a set of suppliers I and a set of customers D, with $|D| = m$. The suppliers have limited capacity M_i for $i \in I$. There is a transportation cost c_{ij} for shipping a unit of product from supplier $i \in I$ to customer $j \in D$. The customer demands are random and are represented by a random vector $\tilde{d}$ taking values in $\mathbf{R}^m$. We assume that we must choose the shipment quantities before the customer demands are known. We enforce the probabilistic constraint

$$\Pr\left\{\sum_{i\in I} x_{ij} \geq \tilde{d}_j, j = 1, \ldots, m\right\} \geq 1 - \epsilon, \tag{15}$$

where $x_{ij} \geq 0$ is the amount shipped from supplier $i \in I$ to customer $j \in D$. The objective is to minimize distribution costs subject to (15) and the supply capacity constraints

$$\sum_{j\in D} x_{ij} \leq M_i \quad \forall i \in I.$$

3.2.1. Test instances. We conducted our tests on an instance with 40 suppliers and 50 customers. The supply capacities and cost coefficients were randomly generated by using normal and uniform distributions, respectively. The demand is assumed to have a joint normal distribution. The mean vector and covariance matrix were randomly generated. We considered two cases for the covariance matrix: a low variance and a high variance case. In the low variance case, the standard deviation of the one-dimensional marginal random demands is 10% of the mean on average. In the high variance case, the covariance matrix of the low variance case is multiplied by 25, yielding standard deviations of the one-dimensional marginal random demands being 50% of the mean on average. In both cases, we consider a single risk level $\epsilon = 0.05$.

We remark that, for this particular choice of distribution, the feasible region defined by the probabilistic constraint is convex [24]. However, the dimension of the

random vector $\tilde{d}$ is $m = 50$, and so evaluating $\Pr\{y \geq \tilde{d}\}$ for a single vector $y \in \mathbf{R}^m$ would present a computational challenge, whereas in our approach we merely need to generate random samples from the joint normal distribution, which is relatively easy. On the other hand, we have not conducted experiments using the convex programming approach, so we cannot comment on whether our approach works better than this. This would be an interesting future experiment. Our intention here is merely to test our approach on a problem with a continuous feasible region and distribution.

Once a sample approximation is solved yielding solution $\hat{x}$, we use a single very large sample ($N' = 250000$) to estimate $\Pr\{\hat{y} \geq \tilde{d}\}$, where $\hat{y} \in \mathbf{R}^m$ is the vector given by $\hat{y}_j = \sum_{i \in I} \hat{x}_{ij}$ for $j \in D$. Letting $d^1, \ldots, d^{N'}$ be the realizations of this large sample, we calculate $\sum_{i=1}^{N'} \mathbb{I}(\hat{y} \geq d^i)$ and use the normal approximation to the binomial distribution to construct an upper bound $\hat{\alpha}$ on the true solution risk $\Pr\{\hat{y} \geq \tilde{d}\}$, which is valid with confidence 0.999. Henceforth for this experiment, if we say a solution is feasible at risk level ϵ, we mean $\hat{\alpha} \leq \epsilon$, and so it is feasible at this risk level with confidence 0.999. We used such a large sample to get a good estimate of the true risk of the solutions generated, but we note that, because this sample was so large, generating this sample and calculating $\sum_{i=1}^{N'} \mathbb{I}(\hat{y} \geq d^i)$ often took longer than solving the sample approximation itself.

3.2.2. Solving the sample approximation. We solved the sample approximation problem by using an MIP formulation, augmented with a class of strong valid inequalities. We refer the reader to [20, 19] for details of this formulation and the valid inequalities, as well as detailed computational results for solving the sample approximation problems. However, we mention that, in contrast to the probabilistic set cover problem, solving the sample approximation problem with the largest sample size that we consider ($N = 10000$) and the largest α (0.05) takes a nontrivial amount of time, in some cases as long as 30 minutes. On the other hand, for $N = 5000$, the worst case was again $\alpha = 0.05$ and usually took less than 4 minutes to solve.

3.2.3. Low variance instance. We begin by presenting results for the instance in which the distribution of demand has relatively low variance. For generating feasible solutions, we tested $\alpha = 0$ with various sample sizes N and report the results for the sample sizes which yielded the best results. Once again, this means that we use a relatively small sample size for the case $\alpha = 0$, as compared to the cases with $\alpha > 0$. We tested several values of $\alpha > 0$ and varying sample size. In contrast to the (PSC) case, we found that taking $\alpha = \epsilon$ or even α close to ϵ did not yield feasible solutions, even with a large sample size. Thus, we report results for several different values of α in the range 0.03–0.036. The reason that we report results for this many different values of α is to illustrate that, within this range, the results are not extremely sensitive to the choice of α (results for more values of α can be found in [19]).

Table 5 gives the characteristics of the solutions generated for the different values of α and N. We observe that, as in the case of (PSC), the *average* cost of the feasible solutions obtained by using $\alpha > 0$ is always less than the *minimum* cost of the feasible solutions obtained with $\alpha = 0$. However, for this instance, the minimum cost solution obtained by using $\alpha = 0$ is not so significantly worse than the minimum cost solutions using different values of $\alpha > 0$, being between 0.40% and 0.58% more costly. As in the case of (PSC), using $\alpha > 0$ and large N significantly reduced the variability of the risk and cost of the solutions generated.

We next investigated the quality of the lower bounds that can be obtained for PTP by solving sample approximation problems. As in the case of (PSC), we obtained

TABLE 5
Solution results for low variance PTP sample problems with $\epsilon = 0.05$.

		Solution risk				Feasible solutions cost				
α	N	Ave	Min	Max	σ	#	Ave	Min	Max	σ
0.000	900	0.048	0.036	0.066	0.011	7	2.0266	2.0199	2.0320	0.0045
	950	0.047	0.039	0.055	0.005	6	2.0244	**2.0185**	2.0291	0.0041
	1000	0.045	0.040	0.051	0.004	8	2.0253	2.0185	2.0300	0.0039
	1500	0.033	0.025	0.043	0.005	10	2.0336	2.0245	2.0406	0.0053
0.030	5000	0.049	0.045	0.050	0.002	6	2.0098	2.0075	2.0114	0.0013
	7500	0.045	0.041	0.047	0.002	10	2.0112	2.0094	2.0136	0.0015
	10000	0.042	0.041	0.044	0.001	10	2.0129	2.0112	2.0145	0.0010
0.033	5000	0.052	0.049	0.054	0.002	2	2.0080	2.0073	2.0088	0.0011
	7500	0.048	0.045	0.051	0.002	7	2.0092	2.0075	2.0107	0.0012
	10000	0.045	0.044	0.047	0.001	10	2.0103	2.0089	2.0118	0.0009
0.036	5000	0.055	0.053	0.057	0.002	0	***	***	***	***
	7500	0.052	0.049	0.054	0.002	2	2.0079	2.0077	2.0080	0.0002
	10000	0.049	0.047	0.051	0.001	8	2.0080	**2.0066**	2.0093	0.0008

TABLE 6
Lower bounds for low variance PTP sample problems with $\alpha = \epsilon = 0.05$.

	LB with confidence at least:				Gap with confidence at least:			
N	0.999	0.989	0.945	0.828	0.999	0.989	0.945	0.828
1000	1.9755	1.9757	1.9775	1.9782	1.55%	1.54%	1.45%	1.42%
3000	1.9879	1.9892	1.9892	1.9910	0.93%	0.87%	0.87%	0.78%
5000	1.9940	1.9943	1.9948	1.9951	0.63%	0.62%	0.59%	0.57%
7500	1.9954	1.9956	1.9959	1.9963	0.56%	0.55%	0.54%	0.52%
10000	1.9974	1.9977	1.9980	1.9981	0.46%	0.45%	0.43%	0.42%

lower bounds by generating and solving 10 sample approximation problems with $\alpha = \epsilon = 0.05$. By taking the lowest value of all of the optimal values, we obtain a lower bound valid with confidence 0.999, and taking the second smallest yields a lower bound which is valid with confidence 0.989, etc. The results for different values of N are given in Table 6. For reference, the percentage gap between these lower bounds and the best feasible solution found (with cost 2.0066) is also given. By using $N \geq 3000$ we obtain lower bounds that are valid with confidence 0.999 and are within one percent of the best feasible solution, indicating that, for this low variance instance, the lower bounding scheme yields good evidence that the solutions that we have found are good quality.

3.2.4. High variance instance. Table 7 gives the characteristics of the solutions generated for the high variance instance. In this case, the *maximum* cost of a feasible solution generated by using any combination of $\alpha > 0$ and N was less than the *minimum* cost of any feasible solution generated by using $\alpha = 0$. The minimum cost feasible solution generated with $\alpha = 0$ was between 0.87% and 1.6% more costly than the best feasible solution generated for the different combinations of $\alpha > 0$ and N. Thus, it appears that, for the high variance instance, using $\alpha > 0$ in a sample approximation is more important for generating good feasible solutions than for the low variance instance.

Table 8 gives the lower bounds for different confidence levels and sample sizes, as well as the gaps between these lower bounds and the best feasible solution found. In this case, solving 10 instances with sample size $N = 1000$ yields a lower bound that is not very tight, 5.11% from the best solution cost at confidence level 0.999. Increasing the sample size improves the lower bound, but even with $N = 10000$ the gap between

TABLE 7
Solution results for high variance PTP sample problems with $\epsilon = 0.05$.

		Solution risk				Feasible solutions cost				
α	N	Ave	Min	Max	σ	#	Ave	Min	Max	σ
0.000	900	0.050	0.035	0.066	0.010	4	3.5068	3.4672	3.5488	0.0334
	950	0.050	0.041	0.058	0.006	6	3.4688	**3.4403**	3.4917	0.0191
	1000	0.045	0.041	0.052	0.004	9	3.4895	3.4569	3.5167	0.0234
	1500	0.030	0.022	0.035	0.005	10	3.5494	3.5205	3.6341	0.0368
0.030	5000	0.050	0.045	0.053	0.002	4	3.4014	3.3897	3.4144	0.0101
	7500	0.046	0.043	0.050	0.002	9	3.4060	3.3920	3.4235	0.0098
	10000	0.043	0.041	0.046	0.001	10	3.4139	3.4001	3.4181	0.0055
0.033	5000	0.053	0.046	0.057	0.003	1	3.4107	3.4107	3.4107	***
	7500	0.049	0.046	0.054	0.002	7	3.3928	**3.3865**	3.4020	0.0062
	10000	0.046	0.042	0.049	0.002	10	3.3982	3.3885	3.4139	0.0086
0.036	5000	0.057	0.049	0.060	0.003	1	3.3979	3.3979	3.3979	***
	7500	0.053	0.050	0.057	0.002	0	***	***	***	***
	10000	0.050	0.046	0.053	0.002	4	3.3927	3.3859	3.3986	0.0054

TABLE 8
Lower bounds for high variance PTP sample problems with $\alpha = \epsilon = 0.05$.

	LB with confidence at least:				Gap with confidence at least:			
N	0.999	0.989	0.945	0.828	0.999	0.989	0.945	0.828
1000	3.2089	3.2158	3.2178	3.2264	5.11%	4.91%	4.85%	4.59%
3000	3.2761	3.2775	3.2909	3.2912	3.12%	3.08%	2.69%	2.68%
5000	3.3060	3.3075	3.3077	3.3094	2.24%	2.19%	2.19%	2.14%
7500	3.3083	3.3159	3.3165	3.3169	2.17%	1.95%	1.93%	1.92%
10000	3.3200	3.3242	3.3284	3.3299	1.83%	1.70%	1.58%	1.53%

the lower bound at confidence 0.999 and the best solution found is 1.83%. Thus, it appears that, for the high variance instance, the sample approximation scheme exhibits considerably slower convergence, in terms of the lower bounds, the feasible solutions generated, or both.

4. Concluding remarks. We have studied a sample approximation scheme for probabilistically constrained optimization problems and demonstrated how this scheme can be used to generate optimality bounds and feasible solutions for very general optimization problems with probabilistic constraints. We have also conducted a preliminary computational study of this approach. This study demonstrates that using sample approximation problems that allow a choice of which sampled constraints to satisfy can yield good quality feasible solutions. In addition, the sample approximation scheme can be used to obtain lower bounds which are valid with high confidence. We found that good lower bounds could be found in the case of a finite (but possibly exponential) feasible region and distribution and also in the case of a continuous feasible region and distribution, provided the distribution has a reasonably low variance. With a continuous feasible region and distribution, if the distribution has a high variance, the lower bounds were relatively weak. Future work in this area will include conducting more extensive computational tests and also extending the theory to allow generation of samples which are not necessarily independent and identically distributed. For example, the use of variance-reduction techniques such as Latin hypercube sampling or quasi-Monte Carlo sampling may yield significantly faster convergence. In addition, to apply the results of this paper to more general probabilistic programs, such as mixed-integer programming with a random constraint matrix, it will be necessary to study how to solve the nonconvex sample approximation problem in these cases.

Acknowledgment. The authors express thanks to Alexander Shapiro for his helpful comments and suggestions related to this work.

REFERENCES

[1] S. Ahmed and A. Shapiro, *The Sample Average Approximation Method for Stochastic Programs with Integer Recourse*, preprint available from www.optimization-online.org, February 2002.

[2] J. Atlason, M. A. Epelman, and S. G. Henderson, *Call center staffing with simulation and cutting plane methods*, Ann. Oper. Res., 127 (2004), pp. 333–358.

[3] J. E. Beasley, *OR-library: Distributing test problems by electronic mail*, J. Oper. Res. Soc., 41 (1990), pp. 1069–1072.

[4] A. Ben-Tal and A. Nemirovski, *Robust convex optimization*, Math. Oper. Res., 23 (1998), pp. 769–805.

[5] P. Beraldi and A. Ruszczyński, *The probabilistic set-covering problem*, Oper. Res., 50 (2002), pp. 956–967.

[6] D. Bertsimas and M. Sim, *The price of robustness*, Oper. Res., 52 (2004), pp. 35–53.

[7] G. C. Calafiore and M. C. Campi, *Uncertain convex programs: Randomized solutions and confidence levels*, Math. Program., 102 (2005), pp. 25–46.

[8] G. C. Calafiore and M. C. Campi, *The scenario approach to robust control design*, IEEE Trans. Automat. Control, 51 (2006), pp. 742–753.

[9] M. C. Campi, G. Calafiore, and S. Garatti, *New results on the identification of interval predictor models*, in Proceedings of the 16th IFAC World Congress, Prague, 2005, Elsevier, New York, 2005.

[10] L. Dai, H. Chen, and J. R. Birge, *Convergence properties of two-stage stochastic programming*, J. Optim. Theory Appl., 106 (2000), pp. 489–509.

[11] L. El Ghaoui and H. Lebret, *Robust solutions to least-squares problems with uncertain data*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 1035–1064.

[12] W.-R. Heilmann, *A note on chance-constrained programming*, J. Oper. Res. Soc., 34 (1983), pp. 533–537.

[13] R. Henrion, P. Li, A. Möller, M. C. Steinbach, M. Wendt, and G. Wozny, *Stochastic optimization for operating chemical processes under uncertainty*, in Online Optimization of Large Scale Systems, M. Grötschel, S. O. Krunke, and J. Rambau, eds., Springer, New York, 2001, pp. 457–478.

[14] R. Henrion and A. Möller, *Optimization of a continuous distillation process under random inflow rate*, Comput. Math. Appl., 45 (2003), pp. 247–262.

[15] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, J. Amer. Statist. Assoc., 58 (1963), pp. 13–30.

[16] A. J. Kleywegt, A. Shapiro, and T. Homem-de-Mello, *The sample average approximation method for stochastic discrete optimization*, SIAM J. Optim., 12 (2001), pp. 479–502.

[17] M. A. Lejeune and A. Ruszczyński, *An efficient trajectory method for probabilistic inventory-production-distribution problems*, Oper. Res., 55 (2007), pp. 378–394.

[18] J. Linderoth, A. Shapiro, and S. Wright, *The empirical behavior of sampling methods for stochastic programming*, Ann. Oper. Res., 142 (2006), pp. 215–241.

[19] J. Luedtke, *Integer Programming Approaches to Some Non-convex and Stochatic Optimization Problems*, Ph.D. thesis, Georgia Institute of Technology, Atlanta, GA, 2007.

[20] J. Luedtke, S. Ahmed, and G. Nemhauser, *An Integer Programming Approach for Linear Programs with Probabilistic Constraints*, in IPCO 2007, Lecture Notes in Comput. Sci., M. Fischetti and D.P. Williamson, eds., Springer, Berlin, 2007, pp. 410–423; full version available online from www.optimization-online.org.

[21] M. R. Murr and A. Prékopa, *Solution of a product substitution problem using stochastic programming*, in Probabilistic Constrained Optimization: Methodology and Applications, S. P. Uryasev, ed., Kluwer Academic, Norwell, MA, 2000, pp. 252–271.

[22] A. Nemirovski and A. Shapiro, *Scenario approximation of chance constraints*, in Probabilistic and Randomized Methods for Design Under Uncertainty, G. Calafiore and F. Dabbene, eds., Springer, London, 2005, pp. 3–48.

[23] A. Nemirovski and A. Shapiro, *Convex approximations of chance constrained programs*, SIAM J. Optim., 17 (2006), pp. 969–996.

[24] A. Prékopa, *On probabilistic constrained programmming*, in Proceedings of the Princeton Symposium on Mathematical Programming, H. W. Kuhn, ed., Princeton University Press, Princeton, NJ, 1970, pp. 113–138.

[25] A. Prékopa, *Probabilistic programming*, in Stochastic Programming, Handbooks Oper. Res. Management Sci. 10, A. Ruszczyński and A. Shapiro, eds., Elsevier, New York, 2003, pp. 267–351.

[26] A. Saxena, *A Short Note on the Probabilistic Set Covering Problem*, available online from www.optimization-online.org, March 2007.

[27] A. Saxena, V. Goyal, and M. Lejeune, *MIP Reformulations of the Probabilistic Set Covering Problem*, available online from www.optimization-online.org, February 2007.

[28] A. Shapiro, *Asymptotic behavior of optimal solutions in stochastic programming*, Math. Oper. Res., 18 (1993), pp. 829–845.

[29] A. Shapiro and T. Homem-de-Mello, *On the rate of convergence of optimal solutions of Monte Carlo approximations of stochastic programs*, SIAM J. Optim., 11 (2000), pp. 70–86.

[30] A. K. Takyi and B. J. Lence, *Surface water quality management using a multiple-realization chance constraint method*, Water Resour. Res., 35 (1999), pp. 1657–1670.

[31] S. Vogel, *Stability results for stochastic programming problems*, Optimization, 19 (1988), pp. 269–288.

[32] W. Wang, *Sample Average Approximation of Risk-Averse Stochastic Programs*, Ph.D. thesis, Georgia Institute of Technology, Atlanta, GA, 2007.

SIAM J. OPTIM.
Vol. 19, No. 2, pp. 700–713

OUTER SEMICONTINUITY OF POSITIVE HULL MAPPINGS WITH APPLICATION TO SEMI-INFINITE AND STOCHASTIC PROGRAMMING*

DINH THE LUC† AND ROGER J.-B. WETS‡

Abstract. Let W be an arbitrary subset of $\mathbb{R}^n$ and pos$W$ the positive hull of W. We are concerned with conditions under which one can guarantee continuity properties for pos$W$ as a function of W. The results are then applied in the context of semi-infinite linear programs and stochastic programs with recourse.

Key words. positive hull, set limits, horizon limits

AMS subject classifications. 49J53, 90C15, 90C34

DOI. 10.1137/070687505

1. Introduction. Let $\{W; W^\nu, \nu \in \mathbb{N}\}$ be a collection of nonempty subsets of $\mathbb{R}^m$ and $\{\text{pos}W; \text{pos}W^\nu, \nu \in \mathbb{N}\}$ the positive hulls generated by these sets, i.e.,

$$\text{pos}W = \left\{ t = \sum_{j=1}^{q} t^j x_j | t^j \in W, x_j \geq 0, \qquad \text{with} \quad q \quad \text{finite} \right\},$$

and posW^ν is defined similarly for each W^ν. Our overall concern is with the continuity of the positive hull mapping $W \mapsto \text{pos}W$, but, more specifically, we are interested in finding conditions under which

$$\limsup_{\nu \to \infty} \text{pos}W^\nu \subset \text{pos}W$$

when the W^ν "converge" to W; here and throughout, $\limsup_\nu C^\nu$ designates the *outer limit*, in the sense of Painlevé-Kuratowski, of the sets C^ν, that is, the set of all limits of subsequences $\lim_{\nu_i \to \infty} t^{\nu_i}$, with $t^{\nu_i} \in C^{\nu_i}$; cf. [11, Chapter 4].

One can view this work as an extension, in various directions, of a result of Walkup and Wets [12, Theorem 2] where the sets W and W^ν were of constant finite cardinality, or, more simply, they consist of the points identified by the columns of constant size matrices. Questions of this type occur in a variety of variational problems. For example, in the analysis of the stability of the solutions of linear programs, semi-infinite linear programs [3], linear complementarity problems [7], [8, section 4], equilibrium and quasi-equilibrium problems [9], generalized linear programs [5, Chapter 22], and stochastic programming problems [2, 10]. Two of these applications are further analyzed in the last two sections.

2. Outer semicontinuity. Our major objective is to obtain the inclusion $\limsup_\nu \text{pos}W^\nu \subset \text{pos}W$ that can be viewed as an outer-semicontinuity result. We

*Received by the editors April 6, 2007; accepted for publication (in revised form) February 22, 2008; published electronically July 2, 2008. This research was supported in part by a grant of the National Science Foundation.

http://www.siam.org/journals/siopt/19-2/68750.html

†Department of Mathematics, University of Avignon, 33 rue Louis Pasteur, Avignon, 84000, France (dtluc@univ-avignon.fr).

‡Department of Mathematics, University of California, Davis, CA 95616-8633 (rjbwets@ucdavis.edu).

begin with a characterization of the outer limit of a sequence of sets that could be deduced from the results in [11, Chapter 4, section H] but isn't readily available in the literature. A (closed) ball centered at x and radius ρ is denoted by $\mathbb{B}(x,\rho)$ and the unit ball simply by $\mathbb{B}$, and so, for some positive scalar η, $\eta\mathbb{B} = \mathbb{B}(0,\eta)$.

PROPOSITION 2.1 (outer limit of sets). *A closed set $C \subset \mathbb{R}^m$ is the outer limit of a sequence of sets $\{C^\nu \subset \mathbb{R}^m,\ \nu \in \mathbb{N}\}$ if and only if given any $\epsilon > 0, \rho > 0$, there is $\nu_{\epsilon,\rho}$ such that*

$$\forall\, \nu \geq \nu_{\epsilon,\rho}: \quad C^\nu \cap \rho\mathbb{B} \subset C + \epsilon\mathbb{B}.$$

Proof. We rely on the criterion provided by [11, Proposition 4.5(b)], namely, that $C \supset \limsup_\nu C^\nu$ if and only if whenever $C \cap B = \emptyset$ for a compact set B, then also $C^\nu \cap B = \emptyset$ for ν large enough. The proof, in both directions, proceeds by contradiction.

When the asserted inclusion is not satisfied for some pair $\epsilon > 0, \rho > 0$, then for a countable collection of indexes, say, for $\nu \in N^\sharp$, there is a collection

$$\left\{x^\nu \in (C^\nu \cap \rho\mathbb{B}) \setminus (C + \epsilon\mathbb{B}),\ \nu \in N^\sharp\right\}$$

converging to some $\bar{x}$; all these points belong to the compact set $\rho\mathbb{B}$. Thus, $C \cap B = \emptyset$ for the compact set $B = \{\bar{x}; x^\nu,\ \nu \in \mathbb{N}\}$, whereas there is no ν_B arbitrarily large so that $C^\nu \cap B$ is empty for all $\nu \geq \nu_B$. Hence, C can't contain the outer limit of the C^ν.

On the other hand, if there is a compact set B such that $C \cap B = \emptyset$ but for some countable collection of indexes, say, $\nu \in N^\sharp$, $C^\nu \cap B$ fails to be empty, choose $\rho > 0$ such that $B \subset \rho\mathbb{B}$ and $\epsilon > 0$ such that $(C + \epsilon\mathbb{B}) \cap B = \emptyset$. Then, for all $\nu \in N^\sharp$, $C^\nu \cap \rho\mathbb{B} \not\subset C + \epsilon\mathbb{B}$; i.e., there is no $\nu_{\epsilon,\rho}$ such that the asserted inclusion holds for all $\nu \geq \nu_{\epsilon,\rho}$. ∎

3. The core cone. To obtain the outer semicontinuity of the positive hulls, our conditions will involve two limit cones associated with a collection of sets $\{C^\nu,\ \nu \in \mathbb{N}\}$. The first one, generated by the directions at the "horizon," is the *horizon outer limit* defined as

$$\limsup_\nu^\infty C^\nu = \{0\} \cup \left\{t = \lim_{\nu\in N} \lambda_\nu t^\nu,\ t^\nu \in C^\nu,\ \lambda_\nu \downarrow 0\right\},$$

where $N \subset \mathbb{N}$ indicates that the limit is conceivable with respect to a subsequence. This set is a closed cone whose properties are detailed in [11, Chapter 4, section F]. The notation

$$C^\infty = \{t | \exists\, t^\nu \in C,\ \lambda_\nu \downarrow 0,\ \text{with } \lambda_\nu t^\nu \to t\}$$

is reserved for the *horizon cone* associated with a set $C \neq \emptyset$ [11, Chapter 3, section B] (see also [1, Chapter 2]).

The second one, believed to be new, can be interpreted as the "inverse" of the horizon limit cone, as will be seen below. It's defined by

$$\limsup_\nu^o C^\nu = \left\{t = \lim_{\nu\in N} \lambda_\nu^{-1} t^\nu,\ t^\nu \in C^\nu,\ \lambda_\nu \downarrow 0\right\},$$

where, again, $N \subset \mathbb{N}$ suggests that the limit may involve only a subsequence of indexes. We refer to this limiting set as the *core outer limit* of the sequence of nonempty sets

$\{C^\nu, \nu \in \mathbb{N}\}$. So, rather than the direction points at the horizon, it identifies the "direction points" at the origin. It's immediate, from the definition, that this (core) limit set is also a *closed cone*. We recall that given a set C, the tangent cone of C at 0 is the cone $\limsup_{\lambda \downarrow 0} \lambda^{-1} C$. Here are some simple rules of calculus of the core outer limit.

PROPOSITION 3.1. *Let $\{C^\nu \subset \mathbb{R}^m, \nu \in \mathbb{N}\}$ and $\{D^\nu \subset \mathbb{R}^m, \nu \in \mathbb{N}\}$ be two collections of sets. Then*

(i) *the core outer limit of C^ν coincides with the intersection of the tangent cones $\limsup_{\lambda \downarrow 0} \lambda^{-1}(\cup_{\nu \geq k} C^\nu)$ for all $k \geq 1$. In particular, when the sets C^ν are constant, say, equal to C, the core outer limit is exactly the tangent cone of C at 0;*

(ii) $\limsup^o_\nu C^\nu \subseteq \limsup^o_\nu D^\nu$ *if* $C^\nu \subseteq D^\nu$ *for all* ν*;*

(iii) $\limsup^o_\nu C^\nu \cup D^\nu = \limsup^o_\nu C^\nu \cup \limsup^o_\nu D^\nu$*;*

(iv) $\limsup^o_\nu C^\nu \cap D^\nu \subseteq \limsup^o_\nu C^\nu \cap \limsup^o_\nu D^\nu$*;*

(v) $\limsup^o_\nu C^\nu + \limsup^o_\nu D^\nu \subseteq co \limsup^o_\nu (C^\nu + D^\nu)$ *provided that C^ν and D^ν contain the origin;*

(vi) $\limsup^o_\nu C^\nu + \limsup^o_\nu D^\nu \supseteq \limsup^o_\nu (C^\nu + D^\nu)$ *provided that the core outer limit of C^ν and that of $-D^\nu$ have only the zero vector in common.*

Proof. To prove the first assertion let $t = \lim_{\nu_i} \lambda^{-1}_{\nu_i} t^{\nu_i}$, with $t^{\nu_i} \in C^{\nu_i}$. Then, for each k, $t^{\nu_i} \in \cup_{\nu \geq k} C^\nu$ whenever $\nu_i \geq k$. Thus, t belongs to the tangent cone of $\cup_{\nu \geq k} C^\nu$ for every k. Conversely, let t be a vector such that, for every $k \geq 1$, there is a sequence of positive numbers $\lambda_{k,i}$ converging to 0 as i tends to ∞ and $t^{\nu_i}_k \in \cup_{\nu \geq k} C^\nu$ such that $t = \lim_i \lambda^{-1}_{k,i} t^{\nu_i}_k$. For each k, choose $i(k)$, with $\lambda_{k,i(k)} \leq \frac{1}{k}$, and $\nu(k) \geq k$, with $t^{\nu_{i(k)}}_k \in C^{\nu(k)}$. By taking a subsequence if necessary, one may assume that $\nu(k) > \nu(k-1)$ for every $k > 1$. Then t is the limit of $\lambda^{-1}_{k,i(k)} t^{\nu_{i(k)}}_k$, with $\lambda_{k,i(k)}$ tending to 0, and hence belongs to the core outer limit of C^ν.

The three assertions that follow are straightforward. The assertion (v) is obtained from (ii) and the fact that C^ν and D^ν are contained in the sum $C^\nu + D^\nu$. For the last assertion, let $v = \lim_{\lambda_i \downarrow 0} \lambda_i (t^{\nu_i} + s^{\nu_i})$, $v \neq 0$, with $t^{\nu_i} \in C^{\nu_i}$ and $s^{\nu_i} \in D^{\nu_i}$. Consider the sequences $\{\lambda_i t^{\nu_i}\}_i$ and $\{\lambda_i s^{\nu_i}\}_i$. They are both either bounded or unbounded. In the first case we may assume that they converge, respectively, to some vector v_1 of the core outer limit of C^ν and some v_2 of the core outer limit of D^ν, which yields $v = v_1 + v_2$. In the other case we divide the general terms of these sequences by $\|\lambda_i t^{\nu_i}\|$ and assume that the obtained sequences converge, respectively, to some vector u_1 of the core outer limit of C^ν and some u_2 of the core outer limit of D^ν. Then, $u_1 + u_2 = 0$, which contradicts the hypothesis. ∎

Example 3.2. The inclusion of (v) and the containment of (vi) may be strict, and, without appropriate assumptions, they may fail to hold.

Detail. In $\mathbb{R}^2$, set $C^\nu = \{(\frac{1}{\nu}, \frac{1}{\nu^2}), (0,0)\}$ and $D^\nu = \{(-\frac{1}{\nu}, \frac{1}{\nu^2}), (0,0)\}$. Then the core outer limit of C^ν consists of the vectors $(x, 0)$, with $x \geq 0$, and the core outer limit of D^ν consists of the vectors $(-x, 0)$, with $x \geq 0$. The core outer limit of the sum $C^\nu + D^\nu$ contains the above-mentioned vectors and the vectors $(0, y)$, with $y \geq 0$, as well. This shows that the inclusion of (v) is strict.

By setting $C^\nu_0 = \{(\frac{1}{\nu}, \frac{1}{\nu^2})\}$ and $D^\nu_0 = \{(-\frac{1}{\nu}, \frac{1}{\nu^2})\}$ we see that the hypothesis of (v) is violated for these families. The convex hull of the core outer limit of their sum is the half-space $\{(0, y) | y \geq 0\}$, which contains neither the core outer limit of C^ν_0 nor the core outer limit of D^ν_0.

For the families C^ν and D^ν above, the hypothesis of (vi) does not hold, and the containment is not true. For the families $A^\nu = \{(\frac{1}{\nu}, \frac{1}{\nu})\}$ and $B^\nu = \{(-\frac{1}{\nu}, \frac{1}{\nu})\}$, direct

calculation shows that the core outer limit of $A^\nu + B^\nu$ is the set of the vectors $(0, y)$, with $y \geq 0$, the core outer limit of A^ν is the ray (x, x), with $x \geq 0$, and the core outer limit of B^ν is the ray $(-x, x)$, with $x \geq 0$. Hence, the sum of the latter core outer limits contains the core outer limit of $A^\nu + B^\nu$ as a proper subset.

To see that the core outer limit set can be interpreted as an inverse of a horizon outer limit set, let's introduce the mapping

$$t \mapsto t^- = t/|t|^2 \qquad \text{for} \qquad t \in \mathbb{R}^m, t \neq 0,$$

with $0^- = 0$. For a set $C \subset \mathbb{R}^m$, by definition, $C^- = \{t^- | t \in C \subset \mathbb{R}^m\}$, and obviously $(C^-)^- = C$. The mapping $t \mapsto t^-$ has the following properties:

(a) It's a homeomorphism on $\mathbb{R}^m \setminus \{0\}$.
(b) C is bounded if and only if $\mathrm{cl}\,(C \setminus \{0\})^-$ does not contain the origin.
(c) With $C^o = \{0\} \cup \{t = \lim_\nu \lambda_\nu^{-1} t^\nu | t^\nu \in C,\ \lambda_\nu \downarrow 0\}$, the *core cone* associated with C, one has $C^o = (C^-)^\infty$ and $C^\infty = (C^-)^o$.
(d) If C is itself a cone, then $C^- = C$.

PROPOSITION 3.3. *For a sequence of nonempty sets* $\{C^\nu \subset \mathbb{R}^m,\ \nu \in \mathbb{N}\}$,

$$\limsup{}^o_\nu C^\nu = \limsup{}^\infty_\nu (C^\nu)^-$$

and

$$\limsup{}^\infty_\nu C^\nu = \limsup{}^o_\nu (C^\nu)^-.$$

Proof. Indeed, let $t \neq 0$ belong to the core outer limit. Without loss of generality, one can assume that $|t| = 1$ and is the limit of some sequence $\{t^\nu/|t^\nu|\}_{\nu=1}^\infty$, with $t^\nu \in C^\nu$ and $\lim_\nu t^\nu = 0$. Then $(t^\nu)^- = t^\nu/|t^\nu|^2 \in (C^\nu)^-$, with $|(t^\nu)^-| \to \infty$. Clearly, t is the limit of the sequence $(t^\nu)^-/|(t^\nu)^-|$, and hence t belongs to the horizon outer limit of the sets C^ν. The converse is obtained by the same argument. The second equality follows from the first one via the identity $(C^-)^- = C$. □

4. The outer semicontinuity of pos$W$. In addition to the limiting cones introduced in the previous section, our conditions also involve $\mathrm{lil}\, K$, the *lineality space* of a convex cone K; it's the maximal linear subspace contained in K.

THEOREM 4.1 (the outer semicontinuity of positive hulls). *The taking of positive hulls is outer semicontinuous, more precisely: Given* $\{W; W^\nu,\ \nu \in \mathbb{N}\}$, *a collection of nonempty subsets of* $\mathbb{R}^m$,

$$\limsup{}_\nu \mathrm{pos} W^\nu \subset \mathrm{pos} W$$

under the following hypotheses:

(a) W *includes the outer limit of the sets* W^ν,
(b) pos$W$ *includes the horizon outer limit of the sets* W^ν,
(c) pos$W$ *includes the core outer limit of the sets* W^ν,
(d) *when* $0 \neq t \in \mathrm{lil}(\mathrm{pos} W)$ *is a cluster point of a sequence* $\{t^\nu/|t^\nu|,\ \nu \in \mathbb{N}\}$, *where* $t^\nu \in \mathrm{pos} W^\nu \setminus \{0\}$, *then* $t^\nu \in \mathrm{lil}(\mathrm{pos} W^\nu)$ *for* ν *sufficiently large and* $\mathrm{lil}(\mathrm{pos} W) \supset \limsup_\nu \mathrm{lil}(\mathrm{pos} W^\nu)$.

Proof. Let $\{t^\nu \in \mathrm{pos} W^\nu\}_{\nu=1}^\infty$ converge to $t \in \mathbb{R}^m$. One needs to show that $t \in \mathrm{pos} W$. According to Carathéodory's theorem [11, Theorem 2.29], one can always express t^ν as

$$t^\nu = \sum_{i=1}^m \lambda_{\nu,i} w^{\nu,i}, \text{ with } \lambda_{\nu,i} \geq 0,\ w^{\nu,i} \in W^\nu,$$

i.e., as a nonnegative linear combination of no more than m vectors in W^ν. Fix an index $i \in \{1, \ldots, m\}$, and consider the sequence $\{\lambda_{\nu,i} w^{\nu,i}\}_{\nu=1}^{\infty}$.

CLAIM 1. *Any cluster point, say,* $\hat{t}^i$*, of* $\{\lambda_{\nu,i} w^{\nu,i},\ \nu \in \mathbb{N}\}$ *belongs to* pos$W$.

In all of the arguments that follow, it's taken for granted that one passes to a subsequence whenever that's required or appropriate. Certainly, if $\hat{t}^i = 0$, it belongs to pos$W$. When $\hat{t}^i \neq 0$, one has to consider three possibilities: (i) $\lim_\nu \lambda_{\nu,i} = 0$ (i.e., the limit of some subsequence is 0), (ii) $\lim_\nu \lambda_{\nu,i} = \lambda_i > 0$ is finite, or (iii) $\lim_\nu \lambda_{\nu,i} = \infty$. In case (i), the sequence $\{w^{\nu,i}\}$ must be unbounded, and all of its "cluster points" belong to the horizon outer limit $\limsup_\nu^\infty W^\nu$. From (b) it follows that they also belong to pos$W$. In case (ii), the (sub)sequence $\{w^{\nu,i},\ \nu \in \mathbb{N}\}$ must be bounded, and hence $\hat{t}^i = \lambda_i w^i$ for some $w^i \in \limsup_\nu W^\nu$. From (a), it then follows that every such cluster point $\hat{t}^i$ also belongs to pos$W$. In the third case (iii), i.e., $\lim_\nu \lambda_{\nu,i} = \infty$, $\hat{t}^i$ then belongs to the core outer limit of the sets W^ν, in which case $\hat{t}^i \in$ pos$W$ by (c). This completes the proof of the assertion.

Let i_0 be an index such that

$$|\lambda_{\nu,i_0} w^{\nu,i_0}| = \max_{i=1,\ldots,m} |\lambda_{\nu,i} w^{\nu,i}|,$$

and consider the sequence $\{\lambda_{\nu,i_0} w^{\nu,i_0}\}$; i_0 is assumed to be common for all ν, again passing to a subsequence if required. If it is bounded, then all of the sequences $\{\lambda_{\nu,i} w^{\nu,i}\}$, $i = 1, \ldots, m$, are bounded, and one can assume that, for each i, they converge to (cluster at) some $\hat{t}^i \in \mathbb{R}^m$. In view of the earlier claim, these limits belong to pos$W$. Thus, also $t = \hat{t}^1 + \cdots + \hat{t}^m$ belongs to pos$W$. There remains only to consider the case when the sequence $\{\lambda_{\nu,i_0} w^{\nu,i_0}\}$ is unbounded, say, $\lim_\nu |\lambda_{\nu,i_0} w^{\nu,i_0}| = \infty$. For all $i = 1, \ldots, m$, the sequences $\{\lambda_{\nu,i} w^{\nu,i} / |\lambda_{\nu,i_0} w^{\nu,i_0}|\}$ are bounded, and one can assume that, for each i, they converge to (equivalently, cluster at) some $u^1, \ldots, u^m$. By Claim 1, these u^i belong to pos$W$.

CLAIM 2. *These limit points* $u^1, \ldots, u^m$ *belong to* lil(pos$W$)*. Consequently, if* u^i *is nonzero, then, for* ν *sufficiently large,* $w^{\nu,i}$ *belong to* lilposW^ν.

Since $|\lambda_{\nu,i_0} w^{\nu,i_0}| \to \infty$, dividing t^ν by $|\lambda_{\nu,i_0} w^{\nu,i_0}|$ and passing to the limit when ν tends to ∞, one obtains $u^1 + \cdots + u^m = 0$. Since pos$W$ is a convex cone, we conclude that $u^1, \ldots, u^m$ must belong to the lineality space of pos$W$. The second part of the assertion now follows directly from condition (d), and thus Claim 2 is verified.

Let I_0 denote the set of all indexes i such that $u^i = 0$ and J_0 the set of remaining indexes. Note that i_0 belongs to J_0 since $|u^{i_0}| = 1$. According to Claim 2, if the index set I_0 is empty, then t^ν belongs to lilposW^ν for ν large enough and, consequently, so does $t \in$ pos$W$. Thus, one has only to consider the case when I_0 is nonempty. Let's write t^ν as the sum of two following terms:

$$t^\nu = z^\nu + y^\nu, \quad \text{where} \quad z^\nu = \sum_{j \in J_0} \lambda_{\nu,j} w^{\nu,j},\ y^\nu = \sum_{i \in I_0} \lambda_{\nu,i} w^{\nu,i}.$$

Now, consider the sequence $\{z^\nu\}_{\nu=1}^{\infty}$. It's bounded or not.

Bounded case: The sequence $\{z^\nu\}$ is bounded. Let z be a cluster point of this sequence that necessarily belongs to pos$W$. Then the corresponding (sub)sequence $\{y^\nu\}$ converges to $t - z$. There are two possibilities as far as this latter sequence is concerned. The first one is when all sequences $\{\lambda_{\nu,i} w^{\nu,i}\}$ are bounded; hence, one may assume that they converge to some w^i, with $i \in I_0$. In view of Claim 1, these limits w^i belong to pos$W$, the limit $t - z = \sum_{i \in I_0} w^i$ also belongs to pos$W$, and one may conclude that $t \in$ pos$W$. When all of the sequences $\{\lambda_{\nu,i} w^{\nu,i}\}$ are not necessarily

bounded, there is a sequence, again possibly passing to a subsequence, $\{\lambda_{\nu,i_1} w^{\nu,i_1}\}$ such that

$$|\lambda_{\nu,i_1} w^{\nu,i_1}| = \max_{i\in I_0} |\lambda_{\nu,i} w^{\nu,i}| \text{ and } \lim_{\nu} |\lambda_{\nu,i_1} w^{\nu,i_1}| = \infty.$$

One can appeal to the same argument as that used earlier for the sequence $\{\lambda_{\nu,i_0} w^{\nu,i_0},\ \nu \in \mathbb{N}\}$, and one can find subsets $I_1 \subset I_0$ and $J_1 = I_0 \setminus I_1$ such that

$$y^\nu = v^\nu + \sum_{i\in I_1} \lambda_{\nu,i} w^{\nu,i},$$

where $v^\nu = \sum_{j\in J_1} \lambda_{\nu,j} w^{\nu,j}$ belongs to lilposW^ν for ν sufficiently large. One can rewrite t^ν as follows:

$$t^\nu = (z^\nu + v^\nu) + \sum_{i\in I_1} \lambda_{\nu,i} w^{\nu,i}.$$

Note that the first term $(z^\nu + v^\nu)$ belongs to lilposW^ν for ν sufficiently large and that the index set I_1 has cardinality strictly smaller than that of I_0. One can proceed in this manner, one eventually exhausts all possible indexes, and one is led to conclude that $t \in \text{pos}W$.

Unbounded case: The sequence $\{z^\nu\}$ is unbounded, say, $\lim_\nu |z^\nu| = \infty$. Assume that $\lim_\nu z^\nu/|z^\nu| = v \in \text{pos}W, v \neq 0$, and

$$0 = \lim_{\nu} \frac{t^\nu}{|z^\nu|} = v + \lim \frac{y^\nu}{|z^\nu|}.$$

Set $\mu_{\nu,i} = \lambda_{\nu,i}/|z^\nu|$, and consider the sequences $\{\mu_{\nu,i} w^{\nu,i}\}_{\nu=1}^\infty$, with $i \in I_0$. By the same argument as in the "bounded case" for the sequences $\{\lambda_{\nu,i} w^{\nu,i}\}_{\nu=1}^\infty$, we come to the conclusion that either $w^{\nu,i}$ belong to lilposW^ν for ν sufficiently large and $i \in I_0$ or there is a nonempty subset $J_{1,ubdd}$ of I_0 such that

$$t^\nu = \sum_{j\in J_0\cup J_{1,ubdd}} \lambda_{\nu,j} w^{\nu,j} + \sum_{i\in I_{1,ubdd}} \lambda_{\nu,i} w^{\nu,i},$$

where the elements of the first sum belong to lilposW^ν for ν sufficiently large and $I_{1,ubdd} := I_0 \setminus J_{1,ubdd}$ is of cardinality strictly smaller than that of I_0. Continuing this procedure and remembering that the number of indexes is finite (m), we arrive at the final step in which either all terms t^ν belong to lilposW^ν for ν sufficiently large or t is a sum of the limit points that belong to pos$W$. In both cases, t is an element of pos$W$ because the latter set is a convex cone. This completes the proof of the theorem. ∎

Remark 4.2. Let's record the following observations about the hypotheses of this theorem:

- (a) When Theorem 4.1(a) holds, so does Theorem 4.1(c), trivially, when either of the following conditions is satisfied:
  - (a_1) The closure of W does not contain the origin;
  - (a_2) $[\limsup_\nu^o W^\nu] \setminus \{0\} \subset \text{int}(\text{pos}W)$.
- (b) If pos$W$ is pointed, then Theorem 4.1(d) is trivially satisfied.
- (c) In Theorem 4.1(d), the inclusion $\text{lil}(\text{pos}W) \supset \limsup_\nu \text{lil}(\text{pos}W^\nu)$ when
  - (c_1) $\text{pos}W \supset \liminf_\nu W^\nu$;
  - (c_2) $\dim \text{lil}(\text{pos}W^\nu) \leq \dim \text{lil}(\text{pos}W)$.

Clearly, condition 4.1(a) is essential. Next, we give three examples to show that each of the remaining conditions can't be neglected either.

Example 4.3 (necessity of condition 4.1(b)). Let $W = \{(0, y) \in \mathbb{R}^2 : y \geq 0\}$, and let $W^\nu = W \cup \{(\nu, \nu)\}$. Then $\text{pos}W = \{(0, y) \in \mathbb{R}^2 : y \geq 0\}$, while $\text{pos}W^\nu = \text{pos}W \cup \{(x, x) : x \geq 0\}$. All of the conditions of the theorem are fulfilled except for the second one.

Example 4.4 (necessity of condition 4.1(c)). Let $W = \{(0, y) : y \geq 0\}$, and let $W^\nu = W \cup \{(1/\nu, 1/\nu^2)\}$. Then $\text{pos}W$ coincides with W, but $\limsup_\nu \text{pos}W^\nu$ is the positive orthant of $\mathbb{R}^2$. In this example only the third condition is violated.

Example 4.5 (necessity of condition 4.1(d)). Let $W = \{(-1, 0), (1, 0)\}$, and let $W^\nu = \{(-1, 1/\nu), (1, 1/\nu)\}$. Then $\text{pos}W = \{(x, 0) : x \in \mathbb{R}\}$, while $\limsup_\nu \text{pos}W^\nu = \{(x, y) : y \geq 0\}$. In this example all of the conditions of the theorem are satisfied except for the fourth one.

When W^ν consists of the columns $a^{\nu,1}, \ldots, a^{\nu,k}$ of a constant size $m \times k$-matrix, Theorem 4.1 yields the following improvement of [12, Theorem 2].

COROLLARY 4.6 (positive hull of converging matrices). *Assume that the vectors $a^{\nu,1}, \ldots, a^{\nu,k}$ converge, respectively, to $a^i, i = 1 \ldots, k$. Set $W = \{a^1, \ldots, a^k\}$, and assume further that*

(a) *for all ν,* $\dim \text{lil}(\text{pos}W^\nu) = \dim \text{lil}(\text{pos}W)$*;*

(b) $\text{pos}W$ *includes all of the cluster points of* $\{a^{\nu,i}/|a^{\nu,i}|\}$ *when* $a^i = 0$*;*

(c) *if* $0 \neq a^i \in \text{lil}(\text{pos}W)$*, then* $a^{\nu,i} \in \text{lil}(\text{pos}W^\nu)$ *for all* ν*.*

Then $\limsup_\nu \text{pos}W^\nu \subset \text{pos}W$*.*

Proof. The proof combines the observations in Remark 4.2(c) with the assertion of Theorem 4.1. □

The condition (b) of [12, Theorem 2] requires that Corollary 4.6(c) holds even when $a^i = 0$. It is clear that this condition then implies both conditions 4.6(b) and 4.6(c), but the converse is not the case, as seen by the next example.

Example 4.7 (relaxed outer semicontinuity). Let $a^{\nu,1} = (0, 1)$ and $a^{\nu,2} = (\frac{1}{\nu}, \frac{1}{\sqrt{\nu}})$.

Detail. Then $a^1 = (0, 1)$ and $a^2 = (0, 0)$. The lineality spaces of $\text{pos}W^\nu$ and $\text{pos}W$ are the null space, and therefore conditions 4.6(a) and 4.6(c) clearly hold. Since $a^{\nu,2} \neq 0$, condition (b) of [12, Theorem 2] is not satisfied. However, one still has $\limsup_\nu \text{pos}W^\nu \subset \text{pos}W$ according to the previous corollary.

The outer semicontinuity of the positive hulls can also be characterized in terms of the inner limits of their positive polar cones. Given a nonempty subset W of $\mathbb{R}^m$ the polar cone of W consists of linear functions that are positive on W, that is,

$$W^* := \{u \in \mathbb{R}^m : \langle u, t\rangle \geq 0, t \in W\}.$$

The *inner limit* of a collection of sets $\{C^\nu, \nu \in \mathbb{N}\}$ is denoted by $\liminf_\nu C^\nu$, which consists of those vectors v for which there exist $v_\nu \in C^\nu$ for every ν such that v is the limit of the sequence $\{v_\nu\}_\nu$.

PROPOSITION 4.8 (limits under polarity). *Let $C \subset \mathbb{R}^n$ be a closed convex cone. Then the following conditions are equivalent:*

(a) $\limsup_\nu \text{pos}W^\nu \subset C$*,*

(b) $\liminf_\nu (\text{pos}W^\nu)^* \supset C^*$*.*

Hence, under the assumptions of Theorem 4.1*,*

$$\liminf_\nu (W^\nu)^* \supset W^*.$$

Proof. Observe that the outer limit of posW^ν coincides with the outer limit of their closures. Now, apply [11, Corollary 11.35] to the closed convex cones $\text{cl}\,\text{posW}^\nu$

to obtain the first assertion. The second assertion is deduced from (a) and Theorem 4.1. □

Let us close up this section by some remarks on lower semicontinuity and upper semicontinuity of the positive hull mappings. It is clear that the inclusion

$$\liminf_{\nu} \mathrm{pos}W^{\nu} \supset \mathrm{pos}W,$$

which characterizes the lower semicontinuity of the positive hull, is true under a quite standard assumption that $\liminf_{\nu} W^{\nu} \supset W$. The upper semicontinuity means that, for any open set A containing $\mathrm{pos}W$, one can find an index ν_0 such that A contains all $\mathrm{pos}W^{\nu}$ for $\nu > \nu_0$. Since $\mathrm{pos}W^{\nu}$ are cones, the above condition implies that $\mathrm{pos}W$ contains all of the cones $\mathrm{pos}W^{\nu}$ whenever $\nu \geq \nu_0$. This is the reason why we focus our attention to outer semicontinuity only.

5. Application: Semi-infinite programs. Consider the following semi-infinite linear program,

$$\text{(siLP)} \qquad \begin{aligned} &\min \langle c, x\rangle \\ &\text{so that} \quad \langle a_t, x\rangle + \beta_t \leq 0,\ t \in T, \end{aligned}$$

where $a_t \in \mathbb{R}^n, \beta_t \in \mathbb{R}$, and T is supposed to be a compact metric space. The system of constraints $\langle a_t, x\rangle + \beta_t \leq 0,\ t \in T$, is denoted by σ and its solution set (the feasible set of (siLP)) is denoted by S. Set

$$W = \{a_t | t \in T\} \qquad \text{and} \qquad \hat{W} = \{(a_t, \beta_t) | t \in T\}.$$

The convex cones generated by W and $\hat{W}$ are, respectively, called the first-moment and the second-moment cone of the system σ (see [6]). We recall also that a sequence $\{x^{\nu}\}_{\nu}$ of vectors in $\mathbb{R}^n$ is said to be an asymptotic solution of the system σ if for every $t \in T$ one has $\liminf_{\nu\to\infty}\langle a_t, x^{\nu}\rangle + \beta_t \leq 0$. It is clear that the system σ is consistent (that is, S is nonempty) if and only if it has a bounded asymptotic solution. A system that has no asymptotic solutions is called strongly inconsistent. Of course, an inconsistent system may have asymptotic solutions as well.

Now we consider a collection of perturbed problems of the same form: For $\nu \in \mathbb{N}$,

$$\text{(siLP}^{\nu}\text{)} \qquad \begin{aligned} &\min \langle c^{\nu}, x\rangle \\ &\text{so that} \quad \langle a_t^{\nu}, x\rangle + \beta_t^{\nu} \leq 0,\ t \in T^{\nu}. \end{aligned}$$

As functions of t, a_t, a_t^{ν}, β_t, and β_t^{ν} are continuous. When the spaces $C(T)$ and $C(T^{\nu})$ of the continuous functions on T and T^{ν} are equipped with the max-norm topology, their (topological) duals are the spaces of measures $M(T)$ and $M(T^{\nu})$, respectively. The cone of positive measures is denoted by $M_+(T^{\nu})$, and $M_+^F(T^{\nu})$ is the cone of positive measures with finite support. The system σ^{ν} and the sets S^{ν}, W^{ν}, and $\hat{W}^{\nu}$ are defined accordingly as σ, S, W, and $\hat{W}$ above.

The first result that can be derived from Theorem 4.1 is on the stability of consistency and strong inconsistency of the system of constraints of (siLP).

PROPOSITION 5.1. *Assume that the collection* $\{\hat{W}; \hat{W}^{\nu}, \nu \in \mathbb{N}\}$ *satisfies the hypotheses of Theorem* 4.1. *Then the following assertions hold:*

(a) *If the system σ is consistent, then, for ν sufficiently large, the systems σ^ν are consistent.*

(b) *If the systems σ^ν are strongly inconsistent, then the system σ is strongly inconsistent.*

Proof. According to the consistency tests (Theorem 4.4 of [6]) the system σ is consistent if and only if the cone $\operatorname{clpos}\hat{W}$ does not contain the vector $e := (0, \ldots, 0, 1)$ of the space $\mathbb{R}^{n+1}$. Since the outer limit of $\operatorname{pos}\hat{W}^\nu$ coincides with the outer limit of $\operatorname{clpos}\hat{W}^\nu$, in view of Theorem 4.1, when ν is sufficiently large, the cone $\operatorname{clpos}\hat{W}^\nu$ does not contain that vector either, and hence the system σ^ν is consistent.

Now, if the systems σ^ν are strongly inconsistent, then again, in view of the consistency tests, the cones $\operatorname{pos}\hat{W}^\nu$ contain the vector e. By Theorem 4.1, the vector e belongs to the cone $\operatorname{pos}\hat{W}$, and hence the strong inconsistency of the system σ follows. □

It is known that the horizon cone of the feasible set S (when it is nonempty) is the solution set to the homogeneous system $\langle a_t, x\rangle \leq 0, t \in T$. It is exactly the negative polar cone of the set W. We derive the outer and inner continuity of the horizon cone of S as follows.

PROPOSITION 5.2. *The following assertions hold:*

(a) *If the collection $\{W; W^\nu, \nu \in \mathbb{N}\}$ satisfies the hypotheses of Theorem* 4.1 *and if the problems* (siLP$^\nu$) *have feasible solutions, then*

$$S_\infty \subseteq \liminf_\nu S^\nu_\infty.$$

(b) *If $W \subseteq \liminf_\nu W^\nu$ and the problem* (siLP) *has feasible solutions, then*

$$S_\infty \supseteq \limsup_\nu S^\nu_\infty.$$

Proof. If S is empty, then the inclusion of (a) holds trivially. If S is not empty, as we have already mentioned, the horizon cone of S is the negative polar cone of the set W. Then (a) follows from Theorem 4.1 and Proposition 4.8. The second assertion is obtained from Proposition 4.8 and the remark at the end of section 4 on the inner semicontinuity of the positive hull mappings. □

An immediate consequence of the above proposition is on the boundedness of the feasible set.

COROLLARY 5.3. *If nonempty, the feasible set S of* (siLP) *is bounded provided that the hypotheses of Theorem* 4.1 *are satisfied and that the feasible sets S^ν are nonempty and bounded. For ν sufficiently large, if nonempty, the sets S^ν are bounded provided that S is nonempty and bounded and that $W \subseteq \liminf_\nu W^\nu$.*

Proof. It is known that, being nonempty, the set S is bounded if and only if its horizon cone is trivial. The corollary now follows from Proposition 5.2. □

We notice that the hypotheses of Theorem 4.1 do not guarantee that problem (siLP) has feasible solutions even if (siLP$^\nu$) have. Next we turn to the stability the of existence of optimal solutions for (siLP).

PROPOSITION 5.4 (existence of solutions). *Assume the following:*

(a) (siLP) *and* (siLP$^\nu$) *satisfy the Slater condition;*

(b) *the collection $\{W; W^\nu, \nu \in \mathbb{N}\}$ satisfies the hypotheses of Theorem* 4.1 *and $c^\nu \to c$;*

(c) *problems* (siLP$^\nu$) *have finite optimal values.*

Then the optimal value of (siLP) *is finite.*

Proof. Direct calculation, either via the Lagrangian function or conjugate calculus (see also [3]), yields the dual problem of (siLP$^\nu$):

$$(siD^\nu) \qquad \begin{aligned} &\max \int_{T^\nu} \beta_t^\nu \, d\mu(t) \\ &\text{so that} \quad \mu \in M_+(T^\nu), \\ &\qquad c^\nu + \int_{T^\nu} a_t^\nu d\mu(t) = 0. \end{aligned}$$

The *finite support dual*, whose (dual) variables are restricted to be measures with finite support, is

$$(\text{siFD}^\nu) \qquad \begin{aligned} &\max \sum_{t\in \text{supp}(\mu)} \beta_t^\nu \mu(t) \\ &\mu \in M_+^F(T^\nu), \\ &c^\nu + \sum_{t\in \text{supp}(\mu)} a_t^\nu \mu(t) = 0, \end{aligned}$$

where $\text{supp}(\mu)$ designates the (finite) support of the measure μ. With $v(\cdot)$ denoting the optimal value, one has the following weak duality inequalities [3]:

$$v(\text{siLP}^\nu) \geq v(\text{siD}^\nu) \geq v(\text{siFD}^\nu).$$

By assumption $v(\text{siLP}^\nu)$ is finite, and hence the set of feasible solutions of (siFD$^\nu$) is nonempty [3, Theorem 5.27], which means that

$$0 \in c^\nu + \text{pos} W^\nu.$$

Passing to the limit and applying Theorem 4.1, one obtains

$$0 \in c + \limsup_\nu \text{pos} W^\nu \subset c + \text{pos} W.$$

Hence, (siFD) is feasible, and in turn this implies that v(siLP) is finite. □

Remark 5.5 (convergence of the solutions). The propositions that we have established in this section are direct applications of Theorem 4.1. Some more convergence properties which are less direct from the above-said theorem can also be said about (siLP). For instance, the convergence of the feasible solutions is stated as follows.

(i) If $\hat{W} \subseteq \liminf_\nu \hat{W}^\nu$, then

$$S \supseteq \limsup_\nu S^\nu;$$

(ii) if the family $\hat{W}^\nu$ is relaxed upper semicontinuous in the sense that, for every $\epsilon > 0$, there is some integer ν_0 such that all $\hat{W}^\nu$, $\nu \geq \nu_0$, are within an ϵ-neighborhood of $\hat{W}$ in the product space $\mathbb{R}^m \times \mathbb{R}$, and if S admits a strong Slater point, that is, a point $\overline{x} \in S$, with $\langle a_t, \overline{x}\rangle + \beta_t \leq -\delta$ for all $t \in T$ and some $\delta > 0$, then

$$S \subseteq \liminf_\nu S^\nu.$$

The proof of these assertions presents no difficulty, so we omit it. The convergence of the optimal solutions and the optimal value of (siLP) can also be obtained by using the methods of [4, 6], in which a rather complete analysis of convergence of (siLP) has been exposed in the case when the index sets T^ν are common for all ν.

6. Application: Stochastic programs. Here, we consider an extension of the *"linear" stochastic program recourse model* [2, 10] to one where the recourse problem takes on the form of a generalized linear program. This brings us into the realm of problems with nonlinear recourse, and, for this particular class, we derive rather explicit expressions for the induced constraints as well as a lower semicontinuity result for the optimal value function. Let $\Im$ stand for the random components of the problem that takes its values, denoted by ξ, in $\Xi \subset \mathbb{R}^d$, the (closed) support of the associated probability measure P. The *recourse cost function* is

$$Q(\xi, x) = \inf\left\{\sum_{j=1}^{J} q_j y_j \,|\, y_j \geq 0, \text{ with } \xi - Tx = \sum_{j=1}^{J} t^j y_j, t^j \in W, j = 1, \ldots, J, \text{ and } J \in \mathbb{N}\right\},$$

where, for $j = 1, \ldots, J$,

$$\begin{pmatrix} q_j \\ t^j \end{pmatrix} \in C \subset \mathbb{R}^{m+1},$$

with C convex; it could be the epigraph of (nonlinear) convex functions, for example. The *expected recourse function* is

$$EQ(x) = E\{Q(\Im, x)\} = \int_{\Xi} Q(\xi, x)\, P(d\xi),$$

with the stochastic program

$$\begin{aligned} &\min f_0(x) + EQ(x) \\ \text{so that} \quad &Ax = b, x \geq 0, \end{aligned}$$

with $f_0 : \mathbb{R}^n \to \mathbb{R}$. Let

$$K_1 = \left\{x \in \mathbb{R}^n_+ \,|\, Ax = b\right\} \quad \text{and} \quad K_2 = \text{dom}EQ.$$

The problem is said to have *relatively complete recourse* when $K_1 \subset K_2$. That's a desirable property, but, unfortunately, this is not universally the case. When $E\{\Im\}$ is finite [13], as we now assume, for the set of *induced constraints*, one has

$$K_2 = \text{dom}EQ = \bigcap_{\xi \in \Xi} \text{dom}Q(\xi, \cdot) = \bigcap_{\xi \in \Xi} \{x \,|\, \xi - Tx \in \text{pos}W\}.$$

Our emphasis here will be on the dependence of the set K_2 on perturbations affecting T, W, and Ξ, in particular when the set of feasible solutions of the stochastic program is not necessarily bounded. Let's denote these perturbed versions by T^ν, W^ν, and Ξ^ν.

LEMMA 6.1 (outer semicontinuity of the induced constraints). *When*

(a) $\liminf_\nu \Xi^\nu \supset \Xi$ *and* $T^\nu \to T$,

(b) *the collection* $\{W; W^\nu, \nu \in \mathbb{N}\}$ *satisfies the hypotheses of Theorem* 4.1,

then $\limsup_\nu K_2^\nu \subset K_2$.

Proof. Let $\{x^\nu \in K_2^\nu\}_{\nu=1}^\infty$ converge to x and $\xi \in \Xi$. In view of (a), one can find $\xi^\nu \in \Xi^\nu$ such that $\lim_\nu \xi^\nu = \xi$. Then $\lim_\nu (\xi^\nu - T^\nu x^\nu) = \xi - Tx$ that belongs to pos$W$ according to Theorem 4.1. It follows that $x \in K_2$. ∎

Generally, the inclusion $\limsup_\nu D^\nu \subset D$ does not imply $\limsup_\nu^\infty D^\nu \subset D^\infty$. A remarkable exception is the case of induced constraints.

LEMMA 6.2 (the horizon outer limit of the induced constraints). *Under the hypothesis of Lemma* 6.1*, one has*

$$\limsup_\nu^\infty K_2^\nu \subset K_2^\infty.$$

Moreover, when $K_1 \cap K_2 \neq \emptyset$*, also* $\limsup_\nu^\infty (K_1 \cap K_2^\nu) \subset (K_1 \cap K_2)^\infty$.

Proof. Observe first that a vector x belongs to the horizon cone K_2^∞ if and only if $-Tx$ belongs to clpos$W$. Let $z \in \limsup_\nu^\infty K_2^\nu$, say, $z = \lim_k \lambda_k z^{\nu_k}$, where $z^{\nu_k} \in K_2^{\nu_k}$ and $\lambda_k \downarrow 0$. For $\xi \in \Xi$, in view of Lemma 6.1(a), one can find $\xi^k \in \Xi^{\nu_k}$ such that $\xi^k \to \xi$. Thus, $\xi^k - T^{\nu_k} z^{\nu_k} \in \text{pos}W^{\nu_k}$ and $\lambda_k \xi^k - T^{\nu_k}(\lambda_k z^{\nu_k}) \in \text{pos}W^{\nu_k}$, as well. Passing to the limit when k goes to ∞ yields $-Tz \in \text{pos}W$ via Lemma 6.1. Thus, $z \in K_2^\infty$. The second assertion is obtained from the first one by relying on [11, Proposition 3.9]. □

This leads us to a lower-semicontinuity result for the optimal value. Here, we also allow for perturbations f_0^ν of the objective function f_0 as well as for perturbations P^ν of the probability measure P. The expected recourse functions $E^\nu Q$ of the perturbed problems are then defined by

$$EQ^\nu(x) = \int_{\Xi^\nu} \left[\inf \left\{ \sum_{j=1}^J q_j y_j | y_j \geq 0 \text{ such that } \xi - T^\nu x = \sum_{j=1}^J t^j y_j, t^j \in W^\nu, \right. \right.$$
$$\left. \left. j = 1, \ldots, J, \text{with } J \text{ finite} \right\} \right] P^\nu(d\xi).$$

Let v and v^ν denote the optimal values of the given stochastic program and of its perturbations, respectively.

PROPOSITION 6.3 (lower semicontinuity of the optimal value). *When*

(a) $\liminf_\nu \Xi^\nu \supset \Xi$ *and* $T^\nu \to T$,

(b) *the collection* $\{W; W^\nu, \nu \in \mathbb{N}\}$ *satisfies the hypotheses of Theorem* 4.1,

(c) *given* (b), $E^\nu Q$ *epiconverges to* EQ,[1]

(d) *the functions* $f^\nu = f_0^\nu + E^\nu Q$ *are quasi-convex,*

(e) *the functions* f_0^ν *converge continuously to* f_0,

(f) *the set of solutions of our (given) stochastic program is nonempty and bounded, then* $\liminf_\nu v^\nu \geq v$.

Proof. A standard argument about the inf-projection and the summation, or integration, of convex functions yields the convexity of EQ and $E^\nu Q$. Assumptions (c) and (e) imply that the functions f^ν epiconverge to f [11, Theorem 7.46(b)]. We now proceed by contradiction. Suppose that one can find x^ν such that $\liminf_\nu f^\nu(x^\nu) < v$. Let $x^0 \in \text{argmin}_{K_1 \cap K_2} f$. If the sequence $\{x^\nu, \nu \in \mathbb{N}\}$ is bounded, in view of Lemma 6.1, one may assume that it converges to some $x \in K_1 \cap K_2$. This means, by epiconvergence, that $\liminf_\nu f^\nu(x^\nu) \geq f(x) \geq v$, and that's a contradiction. If the sequence $\{x^\nu\}$ is unbounded, we may assume that $\{x^\nu/|x^\nu|\}$ converges to some nonzero vector $z \in \limsup_\nu^\infty (K_1 \cap K_2^\nu)$. By Lemma 6.2, $z \in (K_1 \cap K_2)^\infty$; i.e., given $\lambda > 0$, one can find $\lambda_\nu > 0$ converging to 0 such that $x^0 + \lambda_\nu(x^\nu - x^0)$ converges to $x^0 + \lambda z$. Since the f^ν are quasi-convex, one must have

$$f^\nu(x^0 + \lambda_\nu(x^\nu - x^0)) \leq \max\{f^\nu(x^0), f^\nu(x^\nu)\} \leq \max\{f^\nu(x^0), v\}.$$

[1] This is not a very demanding condition, and it actually occurs under rather minimal assumptions; cf. [14, sections 6 and 8] for a brief survey.

This implies that

$$f(x^0 + \lambda z) \leq \max\{f(x^0), v\} = v.$$

Thus, for all $\lambda \geq 0$, $x^0 + \lambda z$ minimizes f on $K_1 \cap K_2$, in contradiction with assumption (f). □

Example 6.4 (without quasi-convexity). If the functions f^ν are not quasi-convex, then the inequality $\liminf_\nu v^\nu \geq v$ is no longer valid.

Detail. To see this, let us define the objective function $f_0 : \mathbb{R} \to \mathbb{R}_+$ by

$$f(x) = \begin{cases} -x - 1 & \text{if } x \leq 0, \\ x - 1 & \text{if } 0 < x < 3, \\ 2 & \text{if } x \geq 3 \end{cases}$$

and choose the constraints so that the feasible set $K_1 \cap K_2 = \mathbb{R}_+$. (The set W consists of one element 1, T is the matrix (-1), and the random variable Ξ takes the values $0, 1, 2, \ldots$.) The perturbed problems are given with the same feasibility set, and the objective functions are

$$f^\nu(x) = \begin{cases} f(x) & \text{if } x \leq \nu, \\ -x + \nu + 1 & \text{if } \nu < x < \nu + 3, \\ -f(x) & \text{if } x \geq \nu + 3. \end{cases}$$

Then all of the hypotheses of the proposition are fulfilled except for the quasi convexity of the f^ν. The optimal value v^ν is -2 while the optimal value $v = -1$.

Example 6.5 (without boundedness of $\operatorname{argmin} f$). The conclusion of Proposition 6.3 is no longer true if the set of minimizers of f on $K_1 \cap K_2$ is unbounded as demonstrated by the following example.

Detail. Again, let $K_1 \cap K_2 = \mathbb{R}_+$. The objective function is

$$f(x) = \begin{cases} -x - 1 & \text{if } x \leq 0. \\ -1 & \text{if } x > 0, \end{cases} \text{ and } f^\nu(x) = \begin{cases} f(x) & \text{if } x \leq \nu, \\ -x + \nu - 1 & \text{if } \nu < x < \nu + 1, \\ x - \nu - 3 & \text{if } x \geq \nu + 1. \end{cases}$$

Then the set of minimizers of f is unbounded, and the optimal value of the perturbed problems is $v^\nu = -2$ while $v = -1$.

Acknowledgment. The authors are thankful to a referee for bringing their attention to references [4] and [6].

REFERENCES

[1] A. Auslender and M. Teboulle, *Asymptotic Cones and Functions in Optimization and Variational Inequalities*, Springer, New York, 2003.
[2] J. Birge and F. Louveaux, *Introduction to Stochastic Programming*, Springer, New York, 1997.
[3] F. Bonnans, *Optimization Continue*, Dunod, Paris, 2006.
[4] M. J. Canovas, M. A. Lopez, J. Parra, and M. I. Todorov, *Stability and well-posedness in linear semi-infinite programming*, SIAM J. Optim., 10 (1999), pp. 82–98.
[5] G. B. Dantzig, *Linear Programming and Extensions*, Princeton University Press, Princeton, NJ, 1963.
[6] M. Goberna and M. Lopez, *Linear Semi-Infinite Optimization*, Wiley, New York, 1998.

[7] F. Flores-Bazán and R. Lopez, *The linear complementarity problem under asymptotic analysis*, Math. Oper. Res., 30 (2006), pp. 73–90.
[8] A. Jofré and R. J.-B. Wets, *Variational Convergence of Bivariate Functions: Motivating Applications* I, manuscript, 2008.
[9] M. A. Mansour, *Sensibilité et stabilité des points d'équilibre et quasi-équilibre: Applications aux inégalités hemi-variationnelles et variationnelles*, Ph.D. thesis, Université Cadi Ayyad, Marrakech, 2002.
[10] A. Prekopa, *Stochastic Programming*, Akadémiai Kiado, Budapest, Hungary, 1995.
[11] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, 2nd ed., Springer, New York, 2004.
[12] D. Walkup and R. J.-B. Wets, *Continuity of some convex-cone valued mappings*, Proc. Amer. Math. Soc., 18 (1967), pp. 229–235.
[13] R. J.-B. Wets, *Stochastic programs with fixed recourse: The equivalent deterministic problem*, SIAM Rev., 16 (1974), pp. 309–339.
[14] R. J.-B. Wets, *Stochastic programming*, in Optimization, Handbooks Oper. Res. Management Sci. 1, G. L. Nemhauser, A. H. G. Rinnooy, and M. J. Todd, eds., North-Holland, Amsterdam, 1989, pp. 573–629.

SIAM J. OPTIM.
Vol. 19, No. 2, pp. 714–739

FINDING BEST APPROXIMATION PAIRS RELATIVE TO A CONVEX AND PROX-REGULAR SET IN A HILBERT SPACE*

D. RUSSELL LUKE†

Abstract. We study the convergence of an iterative projection/reflection algorithm originally proposed for solving what are known as phase retrieval problems in optics. There are two features that frustrate any analysis of iterative methods for solving the phase retrieval problem: nonconvexity and infeasibility. The algorithm that we developed, called relaxed averaged alternating reflections (RAAR), was designed primarily to address infeasibility, though our strategy has advantages for nonconvex problems as well. In the present work we investigate the asymptotic behavior of the RAAR algorithm for the general problem of finding points that achieve the minimum distance between two closed convex sets in a Hilbert space with empty intersection, and for the problem of finding points that achieve a local minimum distance between one closed convex set and a closed *prox-regular set*, also possibly nonintersecting. The nonconvex theory includes and expands prior results limited to convex sets with nonempty intersection. To place the RAAR algorithm in context, we develop parallel statements about the standard alternating projections algorithm and gradient descent. All of the various algorithms are unified as instances of iterated averaged alternating proximal reflectors applied to a sum of regularized maximal monotone mappings.

Key words. best approximation pair, convex set, prox-regular, inconsistent feasibility problems, projection, relaxed averaged alternating reflections, fixed point, resolvent, maximal monotone mappings

AMS subject classifications. 90C26, 49M27, 49M20, 49J53, 65K05

DOI. 10.1137/070681399

1. Introduction.

Projection algorithms are simple yet powerful iterative techniques for finding the intersections of sets. Perhaps the most prevalent example of a projection algorithm is the alternating projections onto convex sets (POCS) dating back to von Neumann [54]. This and algorithms like it have been applied in image processing [19], medical imaging and economics [14], and optimal control [27] to name a few. For a review and historical background, see [4]. The theory for these algorithms is limited mainly to convex setting and to *consistent feasibility problems*, that is, problems where the set intersection is nonempty; if the intersection is empty, then the problem is referred to as an *inconsistent* feasibility problem; examples are abound of practitioners using these methods for nonconvex and/or inconsistent problems. We have in recent years been particularly interested in projection algorithms in crystallography and astronomy [6, 39], and more recently in inverse scattering [34, 33, 15, 12, 11]. Until now, we have been forced to rely on convex heuristics to justify certain strategies [7, 38] for want of an adequate nonconvex theory. In the absence of a nonconvex theory, practitioners resort to ad hoc stopping criteria and other strategies to get their algorithms to work according to user defined criteria. Depending on the algorithms, iterates tend to either stagnate at an undesirable point, or "blow up." We are particularly interested in those algorithms that appear to be unstable. Using convex heuristics we were able to provide plausible explanations [8] and remedies [38] for algorithmic instabilities; however, a general theory was not pursued.

*Received by the editors January 30, 2007; accepted for publication (in revised form) March 16, 2008; published electronically July 2, 2008. This research was supported by NSF grant DMS-0712796.
http://www.siam.org/journals/siopt/19-2/68139.html
†Department of Mathematical Sciences, University of Delaware, Newark, DE 19716 (rluke@math.udel.edu).

Our goal in this paper is two-fold: first, to prove the convergence in the convex setting of an algorithm that we have proposed to solve inconsistent feasibility problems [38], and second, to modify the theory to accommodate nonconvexity. Our algorithm, called relaxed averaged alternating reflections (RAAR), can be viewed as a relaxation of a fixed point mapping used by Lions and Mercier to solve generalized equations involving the sum of maximal monotone mappings [37] and which is an extension of an implicit iterative algorithm by Douglas and Rachford [24] for solving linear partial differential equations.

In section 2 we analyze the RAAR algorithm in the convex setting. Our task here is to characterize the fixed point set of the mapping, as well as to verify the assumptions of classical theorems. Our main result for this section is Theorem 2.7 which establishes convergence of the RAAR algorithm with approximate evaluation of the fixed point operator and variable relaxation parameter. The novelty of our mapping is that it addresses the crucial instance of inconsistent feasibility problems. Inconsistency is a source of instability for more conventional strategies. To place our new algorithm in the context of better-known strategies, we show in Proposition 2.5 that RAAR, alternating projections, and gradient descent are all instances of iterated alternating averaged proximal reflectors—the Lions–Mercier algorithm—applied to the problem of minimizing the sum of two regularized maximal monotone mappings.

In section 3 we expand our convex results to accommodate nonconvexity. In addition to characterizing the fixed point set, we formulate local, nonconvex versions of convex theorems, in particular formulations where *prox-regularity* is central. Our main result in this section is Theorem 3.12 which establishes local convergence of nonconvex applications of the RAAR algorithm. While the nonconvex theory includes the convex case, we present both separately to highlight the places where nonconvexity requires extra care, and to make the nonconvex theory more transparent. The nature of nonconvexity requires focused attention to specific mappings; however, we generalize wherever possible. Failing that, we detail parallel statements about the more conventional alternating projection algorithm; this also allows comparison of our algorithm with gradient descent methods for solving nonlinear least squares problems.

Our analysis complements other results on the convergence of projection algorithms for consistent nonconvex problems [22, 36, 35]. In particular, we point out that the key assumption that we rely upon for convergence, namely a type of local nonexpansiveness of the fixed point mapping, does not appear to yield *rates* of convergence as are achieved in [36, 35] using notions of *regularity* of the intersection. On the other hand, regularity, in addition to assuming the intersection is nonempty, is a strong condition on the intersection that, in particular, is not satisfied for ill-posed inverse problems, our principal motivation.

To close this subsection, we would like to clarify the relationship of the present work to previous work on the phase retrieval problem in crystallography and astronomy that has been a major motivation for these investigations. The results developed in this work apply in principle to the finite-dimensional phase retrieval problem (that is, discrete bandlimited images), so long as certain regularity of the fixed point mapping, namely local firm nonexpansiveness, can be determined. Such an investigation is beyond the scope of this work. The infinite dimensional phase retrieval problem, on the other hand, as studied in [13] does *not* fall within the theory developed here because the sets generated by the magnitude constraints are not weakly closed [39, Property 4.1], hence they are not prox-regular.

1.1. Basic tools and results. We begin with the central tools and basic results that we will use in our analysis. Throughout this paper $\mathcal{H}$ is a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$. For $A, B \subset \mathcal{H}$ closed, the underlying problem is to

$$\text{find } x \in A \cap B. \tag{1.1}$$

Note that it could happen that $A \cap B = \varnothing$, in which case one might naturally formulate the problem as a nonlinear least squares problem

$$\underset{x}{\text{minimize}}\, J(x) = \tfrac{1}{2}\left(\tfrac{1}{2}\operatorname{dist}_A^2(x) + \tfrac{1}{2}\operatorname{dist}_B^2(x)\right), \tag{1.2}$$

where $\operatorname{dist}_C(x)$ is the distance of the point x to a set C:

$$\operatorname{dist}_C(x) := \inf_{c \in C} |x - c|.$$

If x_* is a locally optimal point, then $0 \in \partial J(x_*)$, where ∂ denotes the subdifferential [17, 51, 18, 30, 31, 32, 43]. Another characterization of a locally optimal point x_* is to associate it with a *best approximation* pair (a, b) satisfying $b \in P_B x_* \subset P_B a$ and $a \in P_A x_* \subset P_A b$, where P_C is the projection defined by

$$P_C x := \operatorname{argmin}_{c \in C} |x - c| = \{ y \in C \subset \mathcal{H} \mid |x - y| = \operatorname{dist}_C(x) \}.$$

If C is convex, then the projection is single-valued. If in addition C is closed and nonempty, then $P_C x$ is characterized by [23, Theorem 4.1]

$$P_C x \in C \quad \text{and} \quad \langle c - P_C x, x - P_C x \rangle \leq 0 \quad \text{for all } c \in C. \tag{1.3}$$

If C is not convex, then the projection, if it exists, is a set. If the projection exists and is single-valued near all points in C, then C is said to be *prox-regular* [49]. The relationship between the subdifferential of the squared distance function to a prox-regular set C and the projection is shown in [49, Proposition 3.1] to be

$$\partial\left(\operatorname{dist}_C^2(x)\right) = 2(x - P_C x) \tag{1.4}$$

for x in a neighborhood of C. If C is convex, then this relationship holds globally. In particular, for A and B prox-regular and x in a proximal neighborhood of both sets, we have

$$\partial J(u) = \tfrac{1}{2}\left((x - P_A x) + (x - P_B x)\right).$$

Example 1.1 (gradient descent and averaged projections). Consider the steepest descent iteration with step length λ_n where P_A and P_B are single-valued:

$$\begin{aligned} x_{n+1} &= x_n - \lambda_n \partial J(x_n) \\ &= (1 - \lambda_n) x_n + \lambda_n \tfrac{1}{2}\left(P_A x_n + P_B x_n\right). \end{aligned} \tag{1.5}$$

In other words, gradient descent for least squares minimization is a *relaxed averaged projection algorithm*. We will come back to this particular algorithm later.

Projection algorithms seek to find an element in $A \cap B$, or best approximation thereof, by iterated projections, possibly with some relaxation strategy, onto A and

B separately. The example above interprets the steepest descent algorithm as a relaxed averaged projection algorithm. Another elementary projection algorithm is the well-known alternating projections algorithm: Given $x_0 \in \mathcal{H}$ generate the sequence $\{x_n\}_{n\in\mathbb{N}}$ by

$$x_n = (P_A P_B)\, x_{n-1}. \tag{1.6}$$

Example 1.2 (averaged projections and alternating projections). A standard formulation in the product space [46] identifies the averaged projections (and hence steepest descent) given by (1.5) with alternating projections. To see this, consider the product space $\mathcal{H}\times\mathcal{H}$ with inner product $\langle (x_1,x_2),\ (y_1,y_2)\rangle := \frac{1}{2}\left((x_1,y_1)+(x_2,y_2)\right)$. Let $C = \{(x,y)\in\mathcal{H}\times\mathcal{H} \mid x = y\}$ and $D = \{(x,y)\in\mathcal{H}\times\mathcal{H} \mid x\in A, y\in B\}$; then

$$P_C P_D(x,x) = \left(\tfrac{1}{2}(P_A+P_B)x, \tfrac{1}{2}(P_A+P_B)x\right).$$

Our focus in this paper is on the convergence of projection algorithms, but the above example serves to emphasize that convergence results about such algorithms can be very broadly applied.

When $A\cap B = \varnothing$ we say that the feasibility problem (1.1) is inconsistent. The distinction between inconsistent and consistent feasibility problems has profound implications for the stability and convergence of numerical algorithms. It is convenient to define *difference set* $B-A$. The *gap* between the sets A and B is the point in $B-A$ closest to the origin. Specifically,

$$G := \overline{B-A}, \qquad g := P_G 0, \qquad E := A\cap(B-g), \quad \text{and} \quad F := (A+g)\cap B, \tag{1.7}$$

where $\overline{B-A}$ denotes the closure of $B-A$. Note that these definitions only make sense when A and B are convex. We will generalize these sets in section 3. Basic characterizations are given in [3, 8]. We note that if $A\cap B\neq\varnothing$, then $E=F=A\cap B$. Even in the case where $A\cap B=\varnothing$, the gap vector g is unique and always well defined. A useful characterization of the gap vector g is via the *normal cone mapping of* G: for G convex $-g\in N_G(g)$, where $N_G(g)$ is defined by

$$N_G\colon g\mapsto \begin{cases} \{y\in\mathcal{H} \mid \langle c-g, y\rangle \le 0 \quad \text{for all } c\in G\} & \text{if } g\in G,\\ \varnothing & \text{otherwise.}\end{cases} \tag{1.8}$$

Example 1.3 (projections and normal cone mappings for convex sets). If C is convex, then the normal cone mapping is the subdifferential of the (infinite) indicator function, ι_C, of the set C:

$$\iota_C(x) := \begin{cases} 0 & \text{for } x\in C,\\ \infty & \text{else,}\end{cases} \tag{1.9}$$

$$\partial\iota_C(x) \;=\; N_C(x), \tag{1.10}$$

$$(I+\rho N_C)^{-1}x \;=\; P_C x \quad \text{for all } \rho>0. \tag{1.11}$$

The mapping $(I+N_C)^{-1}$ is called the *resolvent* of the normal cone mapping, or equivalently in this case, the resolvent of $\partial\iota_C$. Specializing to $C=A\cap B$, the indicator function of the intersection is the sum of the indicator functions of the individual sets

$$\iota_{A+B} = \iota_A + \iota_B$$

and, for A and B convex, the resolvent $(I + (N_A + N_B))^{-1}$ is the projection onto $A \cap B$ supposing this is nonempty. In other words, an element of $A \cap B$ is a zero of $\partial \iota_{A\cap B}$. The parameter ρ in (1.11) is interpreted as a *step size* consistent with backward-stepping descent algorithms (see [25]).

Throughout this work we compare the asymptotic properties of alternating projections to more recent projection strategies. A common framework in the convex setting that provides an elegant synthesis of these algorithms is through operator splitting strategies for solving

$$\underset{x\in\mathcal{H}}{\text{minimize}}\ f_1(x) + f_2(x), \tag{1.12}$$

where f_1 and f_2 are proper, lower semi-continuous (l.s.c.) convex functions from $\mathcal{H}$ to $\mathbb{R}\cup\{\infty\}$. The subdifferentials ∂f_j are then *maximal monotone* [44, Proposition 12.b.]; that is, $\operatorname{gph} \partial f_j$ cannot be enlarged in $\mathcal{H}\times\mathcal{H}$ without destroying the monotonicity of ∂f_j defined by

$$\langle v_2 - v_1, x_2 - x_1\rangle \geq 0 \quad \text{whenever} \quad v_1 \in \partial f_j(x_1),\ v_2 \in \partial f_j(x_2).$$

We then seek points that satisfy the inclusion for the sum of two maximal monotone mappings:

$$0 \in \partial f_1(x) + \partial f_2(x). \tag{1.13}$$

Iterative techniques for solving (1.13) are built on combinations of forward- and backward-stepping mappings of the forms $(I - \lambda\partial f_j)$ and $(I + \lambda\partial f_j)^{-1}$, respectively. For proper, l.s.c. convex functions f_j Moreau [44] showed the correspondence between the resolvent $(I + \lambda\partial f_j)^{-1}$ and the argmin of the regularized mapping f_j centered on x. In particular, define the Moreau envelope, $e_{\lambda f}$, and the proximal mapping, $\operatorname{prox}_{\lambda,f}$, of a function f by

$$e_{\lambda,f}x := \inf_w \left\{ f(w) + \frac{1}{2\lambda}|w - x|^2 \right\} \quad \text{and}$$

$$\operatorname{prox}_{\lambda,f} x := \operatorname{argmin}_w \left\{ f(w) + \frac{1}{2\lambda}|w - x|^2 \right\}.$$

Then by [44, Proposition 6.a] we have

$$\operatorname{prox}_{1,f_j} x = J_{\partial f_j}x := (x + \partial f_j(x))^{-1}, \tag{1.14}$$

where $J_{\partial f_j}$ is the resolvent of ∂f_j. The Moreau envelope at zero, $e_{\lambda,f}0$, is perhaps better known as Tikhonov regularization [53, 52].

Maximal monotonicity of ∂f_j is equivalent to *firm nonexpansiveness* of the resolvent $J_{\partial f_j}$, whose domain is all of $\mathcal{H}$ [42]. A mapping $T : \operatorname{dom} T = \mathbb{X} \to \mathbb{X}$ is *nonexpansive* on the closed convex subset $\mathbb{X} \subset \mathcal{H}$ if

$$|Tx - Ty| \leq |x - y| \quad \text{for all} \ \ x, y \in \mathbb{X}; \tag{1.15}$$

we say that T is firmly nonexpansive on $\mathbb{X}$ when

$$|Tx - Ty|^2 \leq \langle x - y, Tx - Ty\rangle \quad \text{for all} \ \ x, y \in \mathbb{X}. \tag{1.16}$$

Firmly nonexpansive mappings also satisfy the following convenient relation:

$$|Tx - Ty|^2 + |(I-T)x - (I-T)y|^2 \leq |x-y|^2 \quad \text{for all} \quad x, y \in \mathbb{X}. \tag{1.17}$$

For more background, see [28, Theorem 12.1] and [23, Theorem 5.5].

Example 1.4 (projections onto and reflections across convex sets). Let C be a nonempty closed convex set in $\mathcal{H}$. The projection onto C is firmly nonexpansive on $\mathcal{H}$ [23, Theorem 5.5], and the corresponding reflection, defined by $R_C := 2P_C - I$, is nonexpansive.

The following central result we build upon concerns the convergence of iterated nonexpansive mappings allowing for *approximate evaluation* of *dynamically relaxed* mappings with variable *step sizes*. Our formulation follows [20], which is a generalization of an analogous result in [25]. Both [25] and [20] synthesize previous work of Rockafellar [50] and Gol'stein and Tret'yakov [29], and are also related to work of Martinet [40, 41] and Brezis and Lions [10] concerning resolvents of maximally monotone mappings. The theorem is formulated for a common relaxation of the fixed point mapping T. For any arbitrary nonexpansive mapping T, the standard relaxation of the iteration $x_{n+1} = Tx_n$ is to a *Krasnoselski–Mann* iteration [9] given by

$$x_{n+1} = U(T, \lambda_n)x_n := \lambda_n T x_n + (1-\lambda_n)x_n, \quad 0 < \lambda_n < 2. \tag{1.18}$$

By Example 1.1, gradient descent for the squared distance objective (1.2) with step length λ_n is equivalent to a Krasnoselski–Mann relaxation of the averaged projection mapping $T := \frac{1}{2}(P_A + P_B)$. In general, the Krasnoselski–Mann relaxation does not change the set of *fixed points* of T denoted $\mathsf{Fix}\ T$.

LEMMA 1.5. *Let $T = (I+\rho S)^{-1}$ $(\rho > 0)$ be firmly nonexpansive with* $\operatorname{dom} T = \mathcal{H}$. *Then $\mathsf{Fix}\ T = \varnothing$ if and only if there is no solution to $0 \in Sx$.*

Proof. T with $\operatorname{dom} T = \mathcal{H}$ is *firmly nonexpansive* if and only if it is the resolvent of a maximally monotone mapping $F : \mathcal{H} \to 2^{\mathcal{H}}$ [42]. Direct calculation then shows that $\mathsf{Fix}\ T = \varnothing$ is equivalent to $\{x \in \mathcal{H} \mid Fx = 0\} = \varnothing$. ∎

THEOREM 1.6 (inexact evaluation of firmly nonexpansive mappings). *Let T be a firmly nonexpansive mapping on $\mathcal{H}$ with* $\operatorname{dom} T = \mathcal{H}$. *Given any $x_0 \in \mathcal{H}$, let the sequence $\{x_n\}_{n\in\mathbb{N}}$ be generated by*

$$x_{n+1} = (1-\lambda_n)x_n + \lambda_n\left(Tx_n + \epsilon_n\right) \quad \textit{for all } n \geq 0, \tag{1.19}$$

where $\{\lambda_n\}_{n\in\mathbb{N}} \subset]0, 2[$ and $\{\epsilon_n\}_{n\in\mathbb{N}} \subset \mathcal{H}$ are sequences with

$$\sum_{n=0}^{\infty} |\epsilon_n| < \infty, \quad \lambda_- = \inf_{n\geq 0} \lambda_n > 0, \quad \lambda_+ = \sup_{n\geq 0} \lambda_n < 2. \tag{1.20}$$

Then if T possesses a fixed point, x_n converges weakly to a fixed point of T. Convergence is strong if any one of the following holds:

- $\underline{\lim} \operatorname{dist}_{\mathsf{Fix}\ T}(x_n) = 0$.
- $\operatorname{int} \mathsf{Fix}\ T \neq \varnothing$.
- *T is* demicompact *at 0; that is, for every bounded sequence $\{x_n\}_{n\in\mathbb{N}}$ with $Tx_n - x_n$ converging strongly to y, the set of strong cluster points of $\{x_n\}_{n\in\mathbb{N}}$ is nonempty.*

If T is firmly nonexpansive with $\operatorname{dom} T = \mathcal{H}$ *and* $\mathsf{Fix}\ T = \varnothing$, *then $\{x_n\}_{n\in\mathbb{N}}$ is unbounded.*

Proof. All but the last statement is the content of Theorem 5.5 in [20]. To show that x_n is unbounded if T does not have a fixed point for T firmly nonexpansive with

dom $T = \mathcal{H}$, we note that by Lemma 1.5 $\mathsf{Fix}\, T = \varnothing$ if and only if there is no solution to $0 \in Fx$, where T is the resolvent of the maximally monotone mapping F. The result now follows from [25, Theorem 3]. □

For the remainder of this paper we will be concerned with applying the above results to particular instances of the mapping T for convex and nonconvex settings. Our principal task, therefore, is to characterize $\mathsf{Fix}\,(T)$ and to modify the above theory to accommodate nonconvexity. To account for realistic limitations in computing accuracy, we consider fixed point iterations where T is only approximately evaluated. With this in mind, and in the context of (1.12), we compare the following approximate algorithms.

ALGORITHM 1.7 (approximate alternating proximal mappings). *Choose* $x_0 \in \mathcal{H}$. *For* $n \in \mathbb{N}$ *set*

$$x_{n+1} = (1-\lambda_n)x_n + \lambda_n \left(\operatorname{prox}_{1,f_1}\left(\operatorname{prox}_{1,f_2} x_n\right) + \epsilon_n\right). \tag{1.21}$$

ALGORITHM 1.8 (approximate averaged alternating proximal reflections). *Choose* $x_0 \in \mathcal{H}$. *For* $n \in \mathbb{N}$ *set*

$$x_{n+1} = (1-\lambda_n)x_n + \frac{\lambda_n}{2}\left(R_{f_1}\left(R_{f_2}x_n + \epsilon_n\right) + \rho_n + x_n\right), \tag{1.22}$$

where $R_{f_j}x := 2\operatorname{prox}_{1,f_j} x - x$.

The parameter λ_n is the Krasnoselski–Mann relaxation parameter, and the terms ρ_n and ϵ_n account for the error made in the calculation of each of the resolvents separately.

The exact version of Algorithm 1.8 was proposed by Lions and Mercier [37] who adapted the Douglas–Rachford [24] algorithm to solving $0 \in F + G$ for general maximal monotone mappings F and G. Convergence results for the application of this algorithm hinge on the following assumption.

Assumption 1.9. There exist $x \in \mathcal{H}$, $a \in \partial f_1(x)$, and $b \in \partial f_2(x)$ such that $a + b = 0$.

The key result of Lions and Mercier adapted to our setting is that, if Assumption 1.9 holds, then the sequence of iterates $\{x_n\}_{n\in\mathbb{N}}$ generated by Algorithm 1.8 with $\epsilon_n = \rho_n = 0$ for all n converges weakly to $\overline{x} \in \mathcal{H}$ as $n \to \infty$ such that $x_* = J_{\partial f_2}\overline{x}$ solves (1.12) [37, Theorem 1].

Example 1.10 (specialization of (1.12) to convex feasibility). Let $f_1 = \iota_A$ and $f_2 = \iota_B$ in (1.12), where A and B are convex. Then, following Example 1.3 we have

$$\operatorname{prox}_{1,f_1}\operatorname{prox}_{1,f_2} = P_A P_B, \tag{1.23}$$

$$\tfrac{1}{2}(R_{f_1}R_{f_2} + I) = \tfrac{1}{2}(R_A R_B + I). \tag{1.24}$$

Specialization of Algorithm 1.7 to this setting yields the classical alternating projection algorithm. Convergence of the exact algorithm was obtained in [16, Theorem 4] under the assumption that either (a) one of A or B is compact, or (b) one of A or B is finite dimensional and the distance between the sets is attained. In other words, $A \cap B$ can be empty. *Rates* of convergence, however, appear to require a certain regularity of the intersection [35].

Specialization of Algorithm 1.8 yields the averaged alternating reflection (AAR) algorithm studied in [8]. It follows immediately from Example 1.4 that $\frac{1}{2}(R_A R_B + I)$ is firmly nonexpansive (see also [8, Proposition 3.1]). Assumption (1.9) reduces to $A \cap B \neq \varnothing$. If $A \cap B = \varnothing$, then by Theorem 1.6 we have $|\frac{1}{2}(R_A R_B + I)x_n| \to \infty$

as $n \to \infty$. Nevertheless, as long as there exist nearest points in B to A, then the sequences $\{P_B x_n\}_{n\in\mathbb{N}}$ and $\{P_A P_B x_n\}_{n\in\mathbb{N}}$ are bounded with weak cluster points belonging to the sets F and E defined by (1.7) [8, Theorem 3.13]. Indeed, regardless of whether or not $A \cap B = \varnothing$, the set $\mathsf{Fix}\,(T_{AAR} + g)$ is closed and convex and [8, Theorem 3.5]

$$F + N_G(g) \subset \mathsf{Fix}\,(T_{AAR} + g) \subset g + F + N_G(g). \tag{1.25}$$

In other words, if $A \cap B = \varnothing$, then T_{AAR} does not have fixed points, but rather has fixed *directions* or *velocities*. Examples 3.7 and 3.8 of [8] show that the upper and lower bounds on this fixed point set are tight, consistent with the case $A \cap B \neq \varnothing$. The salient point here is that convergence of this algorithm is contingent on the consistency of the feasibility problem.

Generalizations of Lions and Mercier's results to approximate evaluation of the resolvents of maximally monotone mappings have been investigated in [25, 20, 21]. The following theorem, adapted from [21], is a specialization of Theorem 1.6 to Algorithms 1.7 and 1.8.

COROLLARY 1.11 (specialization to Algorithms 1.7 and 1.8). *Let f_1 and f_2 be proper, l.s.c. convex functions from $\mathcal{H}$ to $\mathbb{R} \cup \{\infty\}$, let $\{\rho_n\}_{n\in\mathbb{N}}$ and $\{\epsilon_n\}_{n\in\mathbb{N}}$ be sequences in $\mathcal{H}$, and let $\{\lambda_n\}_{n\in\mathbb{N}}$ be a sequence in $]0,2[$.*

(i) *Let $E := \mathsf{Fix}\,\mathrm{prox}_{1,f_1}\mathrm{prox}_{1,f_2} \neq \varnothing$ and $\{\epsilon_n\}_{n\in\mathbb{N}}$ and $\{\lambda_n\}_{n\in\mathbb{N}}$ satisfy* (1.20). *Then every sequence $\{x_n\}_{n\in\mathbb{N}}$ of Algorithm* 1.7 *converges weakly to a point in E. If $\mathrm{int}\, E \neq \varnothing$, then convergence is strong.*

(ii) *If Assumption* 1.9 *holds and $\{\lambda_n\}_{n\in\mathbb{N}} \subset]0,2[$ with $\sum_{n\in\mathbb{N}} \lambda_n(2-\lambda_n) = \infty$ and $\sum_{n\in\mathbb{N}} \lambda_n\left(\|\rho_n\| + \|\epsilon_n\|\right) < \infty$, then the sequence $\{x_n\}_{n\in\mathbb{N}}$ generated by Algorithm* 1.8 *converges weakly to $\overline{x} \in \mathcal{H}$ as $n \to \infty$ such that $x_* = J_{\partial f_2}\overline{x}$ solves* (1.12). *If Assumption* 1.9 *does not hold, then the sequence $\{x_n\}_{n\in\mathbb{N}}$ generated by Algorithm* 1.8 *is unbounded.*

Proof. (i) is an immediate specialization of Theorem 1.6. For (ii), all but the last statement is [21, Corollary 5.2] with $\gamma = 1$. The last statement of (ii) follows from Theorem 1.6 since $\frac{1}{2}(R_{f_1}R_{f_2} + I)$ is firmly nonexpansive. ∎

2. Convex analysis. For this section we will assume that the sets A and B are closed and convex. Denote

$$T_{AP} := P_A P_B \quad \text{and} \quad T_{AAR} := \tfrac{1}{2}(R_A R_B + I) \tag{2.1}$$

discussed in Example 1.10. As discussed in Example 1.10, the existence of fixed points of T_{AP} is independent of whether or not the feasibility problem is consistent. Indeed, it is easy to see that $\mathsf{Fix}\, T_{AP} = E$ for E defined by (1.7). This is not the case for T_{AAR}. We will argue in the nonconvex setting that the fact that T_{AAR} has no fixed points if $A \cap B = \varnothing$ has tremendous algorithmic potential since it means that averaged alternating reflections will not get stuck in a local minimum. Other algorithms for solving feasibility problems do not suffer from such instabilities with inconsistent problems (alternating projections for instance), but for nonconvex problems, this stability is at the cost of getting caught in local minima. It is this resilience of the AAR algorithm in nonconvex applications that first attracted our attention and, we believe, warrants a closer look. In the next section we compare the behavior of these algorithms in the convex setting.

2.1. Relaxations/regularizations. In this subsection we consider relaxations of T_{AAR} whose associated mappings have fixed points independent of whether or not

$A \cap B = \varnothing$. The common relaxation that we have already discussed is of the form

$$U(T, \lambda) := \lambda T + (1 - \lambda)I, \quad 0 < \lambda < 2, \tag{2.2}$$

for the generic mapping T. If the mapping T is firmly nonexpansive (for instance T_{AP} or T_{AAR}), then this property is preserved under the relaxation $U(T, \beta)$ for $\beta \in]0, 1[$. Krasnoselski–Mann iterations have been extensively studied in Hilbert spaces and more general normed spaces [9] so there is ample theory to draw from for the study of the relaxation $U(T, \beta)$.

An advantage and disadvantage of this relaxation is that the fixed points of $U(T, \lambda)$ are the same as those of T. In particular, since T_{AAR} has a fixed point if and only if $A \cap B \neq \varnothing$, it follows immediately that the same holds for $U(T_{AAR}, \lambda)$: for inconsistent problems neither mapping has a fixed point. To remedy this we consider the following alternative relaxation:

$$V(T, \beta) := \beta T + (1 - \beta)P_B, \quad 0 < \beta < 1. \tag{2.3}$$

Like the Krasnoselski–Mann relaxation, for A and B convex and T firmly nonexpansive, $V(T, \beta)$ is also firmly nonexpansive since it is the convex combination of firmly nonexpansive mappings. Hence if $\mathsf{Fix}\, V(T_{AAR}, \beta)$ is nonempty, then the associated approximate fixed point iteration converges to the fixed point set according to Theorem 1.6. One of the principal advantages of this relaxation is that, as we show in Lemma 2.1, $\mathsf{Fix}\, V(T_{AAR}, \beta)$ is independent of whether or not the associated problem (1.1) is feasible. Moreover, the relaxation parameter β can be used to exert some control on the iterates (see subsection 2.2).

In characterizing the fixed points we note that the relaxation $V(T_{AAR}, \beta)$ is fundamentally different than the standard relaxation $U(T_{AAR}, \lambda)$ which has little qualitative effect on the set of fixed points. The two are independent and may be used together without any redundancy of effect. There can, however, be diminishing returns to the addition of parameters to algorithms of this sort. For our application we have found no significant advantage to employing relaxations of the form (2.2). Nevertheless, by Example 1.1, for cases where the relaxation is related to a step length in a gradient descent algorithm, the optimization of λ_n in (2.2) can clearly lead to improved performance. We therefore retain this relaxation and, for the sake of generality, consider nested relaxations of the form

$$\begin{gathered} U\left(V\left(U(T_{AAR}, \lambda_1), \beta\right), \lambda_2\right) = \lambda_2 V\left(U(T_{AAR}, \lambda_1), \beta\right) + (1 - \lambda_2)I, \\ \text{where } \lambda_2 \in]0, 2[\,. \end{gathered} \tag{2.4}$$

The next theorem is a generalization of [38, Theorem 2.2] where we determined the fixed points of the mapping $V(T_{AAR}, \beta)$ alone. The following analysis of the nested relaxations demonstrates the relative importance of the relaxation strategies. This is discussed in greater detail following the proof of the next observation.

LEMMA 2.1 (characterization of fixed points). *Let* $\beta \in]0, 1[$ *and* $\lambda_1, \lambda_2 \in]0, 2[$. *Then*

$$\mathsf{Fix}\, U\left(V\left(U(T_{AAR}, \lambda_1), \beta\right), \lambda_2\right) = F - \frac{\beta\lambda_1}{1 - \beta} g, \tag{2.5a}$$

where F *and* g *are defined by* (1.7). *Moreover,* $\mathsf{Fix}\, U\left(V\left(U(T_{AAR}, \lambda_1), \beta\right), \lambda_2\right)$ *is closed*

and convex and, for every $x \in \mathsf{Fix}\, U\left(V\left(U(T_{AAR}, \lambda_1), \beta\right), \lambda_2\right)$, *we have the following:*

$$x = P_B x - \frac{\beta\lambda_1}{1-\beta} g, \tag{2.5b}$$

$$P_B x - P_A R_B x = g, \tag{2.5c}$$

$$P_B x \in F, \quad \text{and} \quad P_A P_B x \in E. \tag{2.5d}$$

In the special case where $\beta = 1$, *we have*

$$F + N_G(g) \subset \mathsf{Fix}\,(U(T_{AAR}, \lambda) + \lambda g) \subset g + F + N_G(g). \tag{2.5e}$$

By comparison,

$$\mathsf{Fix}\, U(V(U(T_{AP}, \lambda_1), \beta), \lambda_2) = F - \beta g. \tag{2.6}$$

Proof. For all $\beta \in [0, 1[$, since $\mathsf{Fix}\,(\lambda T + (1-\lambda)I) = \mathsf{Fix}\, T$ for *any* mapping T, the fixed point set is invariant with respect to the outer relaxation $(\lambda_2 V(U(T, \lambda_1)) + (1 - \lambda_2)I)$, so without loss of generality we ignore this relaxation.

Equation (2.6) follows immediately from $\mathsf{Fix}\, T_{AP} = E$.

What remains, then, is to show (a) that $F - \frac{\beta\lambda_1}{(1-\beta)} g \subset \mathsf{Fix}\, V\left(U(T_{AAR}, \lambda_1), \beta\right)$ and, conversely, (b) that $\mathsf{Fix}\, V\left(U(T_{AAR}, \lambda_1), \beta\right) \subset F - \frac{\beta\lambda_1}{(1-\beta)} g$.

We first establish the inclusion $F - \frac{\beta\lambda_1}{1-\beta} g \subset \mathsf{Fix}\, V\left(U(T_{AAR}, \lambda_1), \beta\right)$. Pick $f \in F$, let $x = f - \frac{\beta\lambda_1}{1-\beta} g$, and define $e := f - g$. Now, since $f \in F$ and $g \in P_G 0$, $e \in E$, $-\gamma g \in N_B(f)$, and $\gamma g \in N_A(e)$ for all $\gamma > 0$. Hence $P_B x = f$ and $P_A(e + \gamma g) = e$; thus

$$R_B x = 2P_B x - x = f + \frac{\beta\lambda_1}{1-\beta} g,$$

and

$$P_A R_B x = P_A\left(f + \frac{\beta\lambda_1}{1-\beta} g\right) = P_A\left(e + \frac{1 + \beta(\lambda_1 - 1)}{1-\beta} g\right) = e = f - g.$$

Hence $P_B x - P_A R_B x = g$. This together with the observation that

$$x - T_{AAR} x = P_B x - P_A R_B x \quad \text{for all } x \in \mathcal{H} \tag{2.7}$$

implies $x - \beta U(T_{AAR}, \lambda_1) x - (1-\beta) P_B x = \beta\lambda_1 (x - T_{AAR} x) + (1-\beta)(x - P_B x) = \beta\lambda_1 g + (1-\beta)(x - f) = 0$. Thus, as previously claimed, $F - \frac{\beta\lambda_1}{1-\beta} g \subset \mathsf{Fix}\,(\beta U(T_{AAR}, \lambda_1) + (1-\beta) P_B)$.

We show next that $\mathsf{Fix}\,(\beta U(T_{AAR}, \lambda_1) + (1-\beta) P_B) \subset F - \frac{\beta\lambda_1}{1-\beta} g$. To see this, pick any $x \in \mathsf{Fix}\,(\beta U(T_{AAR}, \lambda_1) + (1-\beta) P_B)$. Let $f = P_B x$ and $y = x - f$. Recall that

$$P_A(2f - x) = P_A(2P_B x - x) = P_A R_B x. \tag{2.8}$$

This, together with the identity (2.7), yields

$$P_A(2f - x) = f + T_{AAR} x - x. \tag{2.9}$$

For our choice of x we have $\beta U(T_{AAR},\lambda_1)x+(1-\beta)P_Bx=\beta\lambda_1 T_{AAR}x+\beta x-\beta\lambda_1 x+(1-\beta)P_Bx=x$, which yields

$$T_{AAR}x-x=\frac{1-\beta}{\beta\lambda_1}(x-P_Bx). \tag{2.10}$$

Then (2.9) and (2.10) give

$$P_A(2f-x)=f+\frac{1-\beta}{\beta\lambda_1}(x-f)=f+\frac{1-\beta}{\beta\lambda_1}y. \tag{2.11}$$

Now, for any $a\in A$, since A is nonempty, closed, and convex, we have

$$\langle a-P_A(2f-u),(2f-u)-P_A(2f-u)\rangle\leq 0, \tag{2.12}$$

and hence

$$\begin{aligned}0&\geq\left\langle a-\left(f+\frac{1-\beta}{\beta\lambda_1}y\right),\ (2f-x)-\left(f+\frac{1-\beta}{\beta\lambda_1}y\right)\right\rangle\\&=\left\langle a-\left(f+\frac{1-\beta}{\beta\lambda_1}y\right),\ -y-\frac{1-\beta}{\beta\lambda_1}y\right\rangle\\&=\frac{\beta(\lambda_1-1)+1}{\beta\lambda_1}\langle -a+f,\ y\rangle+\frac{(1-\beta)(\beta(\lambda_1-1)+1)}{(\beta\lambda_1)^2}|y|^2. \qquad (2.13)\end{aligned}$$

Here we have used (2.12), (2.11), and the fact that $y=x-f$. On the other hand, for any $b\in B$, since B is a nonempty, closed convex set and $f=P_Bx$, we have

$$\langle b-P_Bx,x-f\rangle\leq 0, \tag{2.14}$$

which yields

$$\langle b-f,\ y\rangle=\langle b-f,\ x-f\rangle\leq 0. \tag{2.15}$$

Note that, for $\beta\in\,]0,1[$ and $\lambda_1\in\,]0,2[$, the numerator $\beta(\lambda_1-1)+1>0$; thus (2.13) and (2.15) yield

$$\langle b-a,\ y\rangle\leq-\frac{1-\beta}{\beta\lambda_1}|y|^2\leq 0. \tag{2.16}$$

Now take a sequence $\{a_n\}_{n\in\mathbb{N}}$ in A and a sequence $\{b_n\}_{n\in\mathbb{N}}$ in B such that $g_n=b_n-a_n\to g$. Then

$$\langle g_n,\ y\rangle\leq-\frac{1-\beta}{\beta\lambda_1}|y|^2\leq 0\quad\text{for all } n\in\mathbb{N}. \tag{2.17}$$

Taking the limit and using the Cauchy–Schwarz inequality yields

$$|y|\leq\frac{\beta\lambda_1}{1-\beta}|g|. \tag{2.18}$$

Conversely, $x-(\beta U(T_{AAR},\lambda_1)x+(1-\beta)P_Bx)=\beta\lambda_1\left(f-P_A(2f-x)\right)+(1-\beta)y=0$ gives

$$|y|=\frac{\beta\lambda_1}{1-\beta}\Big|f-P_A(2f-x)\Big|\geq\frac{\beta\lambda_1}{1-\beta}|g|. \tag{2.19}$$

Hence $|y| = \frac{\beta\lambda_1}{1-\beta}|g|$ and, taking the limit in (2.17), $y = -\frac{\beta\lambda_1}{1-\beta}g$, which confirms the identity (2.5b). From (2.8) and (2.11) with $y = -\frac{\beta\lambda_1}{1-\beta}g$ it follows that $f - P_A R_B x = g$, which proves (2.5c) and, by definition, implies that $P_B x = f \in F$ and $P_A P_B x \in E$. This yields identity (2.5d) and proves (2.5a). The closedness and convexity of the fixed point set then follows from the fact that F is closed and convex. (Generally, the fixed point set of any nonexpansive map defined everywhere in a Hilbert space is closed convex; see [28, Lemma 3.4].)

For the special case where $\beta = 1$, a straightforward calculation shows that $\mathsf{Fix}\,(T_{AAR} + g) = \mathsf{Fix}\,(U(T_{AAR}, \lambda) + \lambda g)$. Since, by [8, Theorem 3.5], we have

$$F + N_G(g) \subset \mathsf{Fix}\,(T_{AAR} + g) \subset g + F + N_G(g), \tag{2.20}$$

the result follows immediately, which completes the proof. □

Remark 2.2. Lemma 2.1 shows that the inner relaxation parameter λ_1 has only a marginal effect on the set of fixed points of T_{AAR} compared to the β relaxation, which, provided $g \neq 0$, is unbounded as $\beta \to 1$; it has no effect on the set of fixed points of T_{AP}. The outer relaxation parameter λ_2 has no effect on either mapping. In stark contrast to these, the relaxation parameter β in the relaxation $V(T, \beta)$ has a profound impact on the set of fixed points of T_{AAR} and marginal impact on the fixed points of T_{AP}. Indeed, from (2.20) and (2.5a) it is clear that, for all $0 < \beta < 1$, $\mathsf{Fix}\; V(T_{AAR}, \beta) \subset \mathsf{Fix}\,(T_{AAR} + g)$; thus, by definition, $x_\beta - T_{AAR} x_\beta = g$, where $x_\beta \in \mathsf{Fix}\; V(T_{AAR}, \beta)$. More interestingly, however, the fixed point set becomes vastly larger at $\beta = 1$. Similarly, at $\beta = 0$ the fixed point set becomes all of B.

Having characterized the fixed points of $V(T_{AAR}, \beta)$, we turn our attention to *inexact* Relaxed Averaged Alternating Reflections (RAAR) iterations.

ALGORITHM 2.3 (inexact RAAR algorithm). *Choose $x_0 \in \mathcal{H}$ and the sequence $\{\beta_n\}_{n\in\mathbb{N}} \subset]0, 1[$. For $n \in \mathbb{N}$ set*

$$x_{n+1} = \frac{\beta_n}{2}\left(R_A\left(R_B x_n + \epsilon_n\right) + \rho_n + x_n\right) + (1 - \beta_n)\left(P_B x_n + \frac{\epsilon_n}{2}\right). \tag{2.21}$$

The analogous algorithm to this for inexact alternating projections is the following.

ALGORITHM 2.4 (inexact alternating projection algorithm). *Choose $x_0 \in \mathcal{H}$ and the sequence $\{\eta_n\}_{n\in\mathbb{N}} \subset]0, 1[$. For $n \in \mathbb{N}$ set*

$$x_{n+1} = (1 - \eta_n)x_n + \eta_n\left(P_A\left(P_B x_n + \epsilon_n\right) + \rho_n\right). \tag{2.22}$$

For fixed relaxation parameter β, additional insight into the relaxation $V(T_{AAR}, \beta)$ and the Krasnoselski–Mann-relaxed alternating projection algorithm is gained by considering regularizations of iterated proximal mappings applied to (1.12).

PROPOSITION 2.5 (unification of algorithms).

(i) *Algorithm* 1.8 *applied to* (1.12) *with*

$$f_1(x) = \frac{\beta}{2(1-\beta)}\,\mathrm{dist}^{\,2}_A(x) \quad \text{and} \quad f_2(x) = \iota_B(x) \tag{2.23}$$

and $\lambda_n = 1$ for all n is equivalent to Algorithm 2.3 *with $\beta_n = \beta$ for all n.*

(ii) *Algorithm* 1.8 *applied to* (1.12) *with*

$$f_1(x) = \tfrac{1}{2}\,\mathrm{dist}^{\,2}_A(x) \quad \text{and} \quad f_2(x) = \tfrac{1}{2}\,\mathrm{dist}^{\,2}_B(x) \tag{2.24}$$

and relaxation parameter λ_n is equivalent to Algorithm 2.4 *with $\eta_n = \lambda_n/2$.*

(iii) *Algorithm* 1.8 *applied to* (1.12) *on the product space with* f_1 *and* f_2 *defined by*

$$f_1(x,y) = \tfrac{1}{2}\operatorname{dist}^2_C(x,y) \quad \text{and} \quad f_2(x,y) = \tfrac{1}{2}\operatorname{dist}^2_D(x,y) \tag{2.25}$$

for $C = \{(x,y) \in \mathcal{H} \times \mathcal{H} \mid x = y\}$, $D = \{(x,y) \in \mathcal{H} \times \mathcal{H} \mid x \in A, y \in B\}$, $\epsilon_n = \rho_n = 0$ *for all* n *and relaxation parameter* λ_n *is equivalent to gradient descent with step length* $\lambda_n/2$ *applied to the nonlinear least squares problem* (1.2).

Proof. (i) Let

$$f_1(x) = \frac{\beta}{2(1-\beta)}\operatorname{dist}^2_A(x) \quad \text{and} \quad f_2(x) = \iota_B(x);$$

then $\operatorname{prox}_{1,f_2}(x) = P_B x$ and a short calculation yields $\operatorname{prox}_{1,f_1}(x) = x + \beta(P_A x - x)$. The result then follows upon substituting these expressions into (1.22) with $\lambda_n = 1$ for all n and ρ_n of Algorithm 1.8 replaced by ρ_n of (2.21) scaled by β.

(ii) A similar calculation shows that when $f_1(x) = \frac{1-\beta}{2\beta}\operatorname{dist}^2_A(x)$ and $f_2(x) = \frac{\beta}{2(1-\beta)}\operatorname{dist}^2_B(x)$, the recursion (1.22) is equivalent to

$$\begin{aligned} x_{n+1} &= (1-\lambda_n)x_n \\ &\quad + \lambda_n\left((1-\beta)P_A y_n + \beta(2\beta-1)P_B x_n + 2(\beta-\beta^2)x_n + (\beta - \tfrac{1}{2})\epsilon_n + \tfrac{1}{2}\rho_n\right), \end{aligned} \tag{2.26}$$

where $y_n = (1-2\beta)x_n + 2\beta P_B x_n + \epsilon_n$. In particular, when $\beta = 1/2$ we have $x_{n+1} = (1 - \frac{\lambda_n}{2})x_n + \frac{\lambda_n}{2}(P_A(P_B x_n + \epsilon_n) + \rho_n)$, the Krasnoselski–Mann relaxation of approximate alternating projections given by (2.22) with relaxation parameter $\eta_n = \lambda_n/2$ for all n.

(iii) We use the product space formulation as in Example 1.2. By (ii) of this theorem, Algorithm 1.8, with relaxation parameter λ_n applied to (1.12), where f_1 and f_2 are defined by (2.25), is equivalent to alternating projections on the product space—Algorithm 2.4 with $x_{n+1} = (1-\lambda_n/2)x_n + \lambda_n/2 P_C P_D x_n$. But by Example 1.2 alternating projections is equivalent to Krasnoselski–Mann-relaxed averaged projections on the product space with relaxation parameter $\lambda_n/2$. To complete the proof we note that, by Example 1.1, Krasnoselski–Mann-relaxed averaged projections on the product space with relaxation parameter $\lambda_n/2$ is equivalent to gradient descent with step length $\lambda_n/2$ applied to the nonlinear least squares problem (1.2). ∎

In other words, $V(T_{AAR}, \beta)$ is not a relaxation of T_{AAR} but rather the exact instance of Algorithm 1.8 applied to (1.12) with f_1 and f_2 defined by (2.23). Similarly, alternating projections is also an instance of Algorithm 1.8, which, in turn, yields the equivalence of this algorithm to gradient descent.

Proposition 2.5 yields a proof of the next theorem by direct application of Corollary 1.11 with $\lambda_n = 1$ for all n.

THEOREM 2.6 (the inexact RAAR algorithm with fixed β). *Let* $\{\rho_n\}_{n\in\mathbb{N}}$ *and* $\{\epsilon_n\}_{n\in\mathbb{N}}$ *be sequences in* $\mathcal{H}$ *such that* $\sum_{n\in\mathbb{N}} \|\rho_n\| + \|\epsilon_n\| < \infty$, *and fix* $\beta \in]0,1[$, $x_0 \in \mathcal{H}$, *and* $\lambda_n = 1$ *for all* n. *If* F *defined by* (1.7) *is nonempty, then the sequence* $\{x_n\}_{n\in\mathbb{N}}$ *generated by Algorithm* 2.3 *with* $\beta_n = \beta$ *for all* n *converges weakly to* $\overline{x} \in \mathcal{H}$ *as* $n \to \infty$ *such that* $x_* = P_B\overline{x}$ *solves*

$$\operatorname*{minimize}_{x\in\mathcal{H}} \frac{\beta}{2(1-\beta)}\operatorname{dist}^2_A(x) + \iota_B(x). \tag{2.27}$$

If $F = \varnothing$, then the sequence $\{x_n\}_{n\in\mathbb{N}}$ generated by (2.21) is unbounded.

Proof. The result follows from Corollary 1.11(ii) once the equivalence of the condition $F \neq \varnothing$ to Assumption (1.9) is established. To see this, note that by Proposition 2.5 the recursion (2.21) is equivalent to (1.22) with $\lambda_n = 1$ for all n applied to (2.27). Moreover, for $f_1 = \frac{\beta}{2(1-\beta)} \operatorname{dist}^2_A(x)$ and $f_2 = \iota_B$, we have $\partial f_1(x) = \frac{\beta}{1-\beta}(x - P_A x)$ $\partial f_2(x) = N_B(x)$ (see (1.10) and (1.4)), so the existence of points $x \in F$ implies that $\frac{\beta}{1-\beta}(x - P_A x) = \frac{\beta}{1-\beta} g \in -N_B(x)$; hence Assumption 1.9 holds. Conversely, the existence of points $x \in \mathcal{H}$ and $a \in \partial f_1(x)$ and $b \in \partial f_2(x)$ such that $a + b = 0$ implies that, for such x, $\frac{\beta}{1-\beta}(x - P_A x) \in -N_B(x)$; hence $x \in F$, which completes the proof. ∎

While Theorem 2.6 takes advantage of regularizations to reinterpret the relaxation (2.3), it does not easily allow us to verify the effect of variable β. To account for variable β_n we take a different approach.

THEOREM 2.7 (the inexact RAAR algorithm with variable β). *Fix $\beta \in]0,1[$ and $x_0 \in \mathcal{H}$. Let $\{\beta_n\}_{n\in\mathbb{N}}$ be a sequence in $]0,1[$, and let $\{x_n\}_{n\in\mathbb{N}} \in \mathcal{H}$ be generated by Algorithm 2.3 with corresponding errors $\{\epsilon_n\}_{n\in\mathbb{N}}, \{\rho_n\}_{n\in\mathbb{N}} \subset \mathcal{H}$. Define*

$$\nu_n = 2|\beta_n - \beta|\,|(P_A - I)R_B x_n|\,. \tag{2.28}$$

If $F \neq \varnothing$ and

$$\sum_{n\in\mathbb{N}} |\epsilon_n| + |\rho_n| + \nu_n < +\infty, \tag{2.29}$$

then $\{x_n\}_{n\in\mathbb{N}}$ converges weakly to a point $x_* \in F - \beta g/(1-\beta)$.

Proof. The mapping $V(T_{AAR}, \beta) = \beta T_{AAR} + (1-\beta)P_B$. Then $V(T_{AAR}, \beta)$ is firmly nonexpansive as a convex combination of the two firmly nonexpansive mappings T_{AAR} and P_B. Accordingly, the mapping $R = 2V(T_{AAR}, \beta) - I$ is nonexpansive since $V(T_{AAR}, \beta)$ is firmly nonexpansive if and only if $2V(T_{AAR}, \beta) - I$ is nonexpansive [28, Theorem 12.1]. Moreover, it follows from Lemma 2.1 that $\mathsf{Fix}\, R = \mathsf{Fix}\, V(T_{AAR}, \beta) = F - \beta g/(1-\beta) \neq \varnothing$. Setting $r_n = 2x_{n+1} - x_n$, an elementary calculation shows that

$$\begin{aligned} |r_n - Rx_n| \leq \beta_n |R_A\left(R_B x_n + \epsilon_n\right) - R_A R_B x_n| + \beta_n|\rho_n| \\ + (1-\beta_n)|\epsilon_n| + 2|\beta_n - \beta|\,|(P_A - I)R_B x_n|\,. \end{aligned} \tag{2.30}$$

Now, since R_A is nonexpansive

$$|R_A\left(R_B x_n + \epsilon_n\right) - R_A R_B x_n| \leq |\epsilon_n|$$

and from (2.28)

$$|r_n - Rx_n| \leq |\rho_n| + |\epsilon_n| + \nu_n. \tag{2.31}$$

The recursion (2.21) can thus be rewritten as

$$x_0 \in \mathcal{H} \quad \text{and} \quad x_{n+1} = \tfrac{1}{2}x_n + \tfrac{1}{2}r_n \quad \text{for all } n \in \mathbb{N}, \tag{2.32}$$

where, from (2.29) and (2.31),

$$\sum_{n\in\mathbb{N}} |r_n - Rx_n| < +\infty. \tag{2.33}$$

It then follows from [20, Theorem 5.5(i)] that $\{x_n\}_{n\in\mathbb{N}}$ converges weakly to a fixed point of R, which proves the result. □

Analogous results for the important case of the RAAR algorithm with $\beta = 1$, that is, the AAR algorithm, have been treated extensively in [8]. Surprisingly, the proof techniques for these two cases are distinct, and it appears that a unification is not readily available.

A more detailed picture of the behavior of iterates of the exact RAAR algorithm can be obtained in the following restricted setting.

COROLLARY 2.8 (exact RAAR algorithm in Euclidean space). *Let $\mathcal{H}$ be a Euclidean space. Fix $\beta \in]0,1[$ and let*

$$x_0 \in \mathbb{R}^n \quad \text{and} \quad x_{n+1} = V(T_{AAR},\beta)x_n \quad \text{for all } n \in \mathbb{N}.$$

Suppose that $F \neq \varnothing$. Then $\{x_n\}_{n\in\mathbb{N}}$ converges to some point $x \in F - \beta g/(1-\beta)$ and, furthermore,

(i) $P_B x_n - P_A P_B x_n \to g$;
(ii) $P_B x_n \to P_B x$ *and* $P_A P_B x_n \to P_A P_B x$;
(iii) $P_B x - P_A P_B x = g$; *hence* $P_B x \in F$ *and* $P_A P_B x \in E$.

Proof. The convergence of $\{x_n\}_{n\in\mathbb{N}}$ follows from Theorem 2.7 (with $\epsilon_n = \rho_n = \nu_n := 0$); denote the limit by x. From (1.7) we can write $x \in F - \beta g/(1-\beta)$ as $x = f - \beta g/(1-\beta)$, where $f = P_B x \in F$ (see also [8, Proposition 2.4.(ii)]). Since the mappings P_A, P_B, R_A, R_B are continuous, (ii) follows. Next, using (2.5c), we have

$$P_B x_n - P_A R_B x_n \to P_B x - P_A R_B x = g. \tag{2.34}$$

Hence

$$|g| \leq |P_B x_n - P_A P_B x_n| \leq |P_B x_n - P_A R_B x_n| \to |g|, \tag{2.35}$$

and thus $|P_B x_n - P_A P_B x_n| \to |g|$. Now (i) follows from [8, Proposition 2.5]. Taking the limit in (i) yields (iii). □

We would like to note in closing this subsection that the duality theory for (1.12) with f_1 and f_2 given by (2.23) has been detailed in [2, section 2]. The connection between algorithms (1.8) and (1.7) and (2.3) allows for an attractive synthesis in the convex setting. However, at this time the nonconvex theory is much less developed than the convex theory. A notable exception is the recent work of Moudafi [45], who studies the convergence of the prox-gradient method in a prox-regular setting. Nevertheless, the view of the parameter β as a weight in a regularized objective does not, in our opinion, lead to a natural justification for dynamic β_n as does the interpretation of this parameter as a relaxation. This is discussed in greater detail in the next subsection.

2.2. Controlling the iterates. The implementation of the RAAR algorithm that we studied in [38] was motivated by the following observation that indicates that the relaxation parameter β might be used to *steer* the iterates.

PROPOSITION 2.9. *Let $x \in \mathcal{H}$ and suppose that $F \neq \varnothing$.*

(i) $\operatorname{dist}(x, \operatorname{Fix} V(T_{AP},\beta)) = \operatorname{dist}(x, F + \beta g)$ *for all* $\beta \in (0,1]$.
(ii) *If $A \cap B \neq \varnothing$, then* $\operatorname{dist}(x, \operatorname{Fix} V(T_{AAR},\beta)) = \operatorname{dist}(x, A \cap B)$ *for all* $\beta \in]0,1[$; *otherwise,* $\lim_{\beta\nearrow 1} \operatorname{dist}(x, \operatorname{Fix} V(T_{AAR},\beta)) = +\infty$.

Proof. The proof of (i) follows immediately from (2.6). To see (ii), note that if $A \cap B \neq \varnothing$, then $g = 0$ and $\operatorname{Fix} V(T_{AAR},\beta) = A \cap B$, which proves the first part of the statement. Now assume $A \cap B = \varnothing$ and fix $f_0 \in F$. Then $g \neq 0$ and F is contained in

the hyperplane $\{x \in \mathcal{H} \mid \langle x - f_0,\ g\rangle = 0\}$ [3, Lemma 2.2(v)]. Hence, it follows from Lemma 2.1 that

(2.36)
$$\mathsf{Fix}\ V(T_{AAR},\beta) = F - \frac{\beta}{1-\beta}g \subset \left\{x \in \mathcal{H} \;\middle|\; \left\langle x + \frac{\beta}{1-\beta}g - f_0,\ g\right\rangle = 0\right\} = H_\beta.$$

Accordingly,

$$\text{dist}\ \big(x, \mathsf{Fix}\ V(T_{AAR},\beta)\big) \geq \text{dist}\ (x, H_\beta) = \frac{\left|\left\langle x + \frac{\beta}{1-\beta}g - f_0,\ g\right\rangle\right|}{|g|} \geq \frac{\beta}{1-\beta}|g| - \frac{|\langle x - f_0,\ g\rangle|}{|g|}, \tag{2.37}$$

which proves the second assertion of part (ii). □

By Proposition 2.9, for any estimate x_n "close" to $\mathsf{Fix}\ V(T_{AAR},\beta_n)$, there is a β_{n+1} such that x_n is comparatively distant from $\mathsf{Fix}\ V(T_{AAR},\beta_{n+1})$. It will become clear in the next section that it is the proximity to the set F, rather than $\mathsf{Fix}\ V(T_{AAR},\beta_n)$, that is critical to the quality of an iterate x_n. We therefore use the relaxation parameter β_n to control the *step size* of an iterate toward the set F. By comparison, the relaxation parameter β has very little effect on the iterates x_n of the alternating projection algorithm. The next proposition shows that by varying β the step size can be regulated in the direction of the gap vector g.

PROPOSITION 2.10. *Let $x \in \mathcal{H}$ satisfy $|x - x_{\beta_1}| < \delta$, where $x_{\beta_1} \in \mathsf{Fix}\ V(T_{AAR},\beta_1)$, $\delta > 0$ and $\beta_1 \in\ ]0,1[$. Then, for all $\beta_2 \in\ ]0,1[$, we have*

$$\left|V(T_{AAR},\beta_2)x - \left(f_{\beta_1} - \frac{\beta_2}{1-\beta_1}g\right)\right| < \delta, \tag{2.38}$$

where $f_{\beta_1} = P_B x_{\beta_1} \in F$.

Proof. This proof was proved in [38, Proposition 2.3]. □

This ability to control the step lengths with the relaxation parameter stands out next to other relaxed projection algorithms. For this reason descent algorithms are often preferred since there is ample theory for determining optimal step sizes.

3. Nonconvex analysis.

3.1. Prox-regular sets. In this section A is still convex, but we allow the set B to be nonconvex. Such a situation is encountered in the numerical solution to the phase retrieval problem in inverse scattering [39, 6, 38], and is therefore of great practical interest. Indeed, our results form the basis for proving local convergence of some phase retrieval algorithms for inconsistent (noisy) problems which, to our knowledge, would be the first such results. The central notion for getting a handle on this situation is *prox-regularity* as developed by Poliquin and Rockafellar [48, 47]. Prox-regular sets were, to our knowledge, first introduced by Federer [26] though he called them sets of *positive reach*, and are characterized as those sets C for which the projection is locally single-valued and continuous from the strong topology in the domain to the weak topology in the range [49, Theorem 3.1]. The main difficulty for our analysis is that prox-regularity is a local property relative to elements of the set B while the fixed points of the mapping $V(T,\beta)$ lie somewhere in the normal cone to B

at the local best approximation points of this set. A localized normal cone mapping is obtained through the *truncated normal cone mapping*

$$N_C^r(x) := \begin{cases} N_C(x) \cap \operatorname{int} \mathbb{B}(0,r), & x \in C, \\ \varnothing, & x \notin C, \end{cases} \tag{3.1}$$

where $\mathbb{B}(0,r)$ is the closed ball of radius r centered on the origin. The principal result we draw from can be found in [49, Corollary 2.2 and Proposition 3.1].

LEMMA 3.1 (properties of prox-regular sets). *Let $C \subset \mathcal{H}$ be prox-regular at $\overline{x}$. Then for some $r > 0$ and a neighborhood of $\overline{x}$, denoted $\mathcal{N}(\overline{x})$, the truncated normal cone mapping N_C^r is hypomonotone on $\mathcal{N}(\overline{x})$; that is, there is a $\sigma > 0$ such that*

$$\langle y_1 - y_2,\ x_1 - x_2 \rangle \geq -\sigma |x_1 - x_2|^2 \quad \text{whenever } y_i \in N_C^r(x_i) \text{ and } x_i \in \mathcal{N}(\overline{x}).$$

As suggested by Proposition 2.9 we can control to some extent the location of the fixed points of $V(T,\beta)$ by adjusting the parameter β. In particular, note that for $\beta = 0$ we have $V(T,0) = P_B$; hence we can adjust β so that the fixed points remain in prox-neighborhoods of the best approximation points in B.

The next result is a prox-regular analogue of (1.3).

LEMMA 3.2. *For C prox-regular at $\overline{x}$ there exist $\epsilon > 0$ and $\sigma > 0$ such that whenever $x \in C$ and $v \in N_C(x)$ with $|x - \overline{x}| < \epsilon$ and $|v| < \epsilon$ one has*

$$\langle x' - x, v \rangle \leq \sigma |x' - x|^2 \quad \text{for all } x' \in C \quad \text{with } |x' - x| < \epsilon. \tag{3.2}$$

Proof. Since C is prox-regular at $\overline{x}$, by Lemma 3.1 the truncated normal cone mapping $N_C^r(x)$ is hypomonotone on a neighborhood $\mathcal{N}(\overline{x})$, that is, there are $\sigma > 0$ and $\epsilon > 0$ such that

$$\langle x' - x, v \rangle \leq \sigma |x' - x|^2,$$

whenever $v \in N_C(x)$, and $0 \in N_C(x')$ with $|v| < \epsilon$ and $|x' - x| < \epsilon$. ∎

A stronger version (with a different proof) of the above proposition can be found in [49, Proposition 1.2].

For this prox-regular setting we must define local versions of the sets G, E, and F defined in (1.7).

DEFINITION 3.3 (local best approximation points). *For A convex and B nonconvex, a point $f \in B$ is a* local best approximation point *if there exists a neighborhood $\mathcal{N}(f)$ on which $|f - P_A f| \leq |b - P_A b|$ for all $b \in B \cap \mathcal{N}(f)$. For such a point, we define*

$$G_{\mathcal{N}(f)} := \overline{(B \cap \mathcal{N}(f)) - A} \quad \text{and} \quad \text{for } g := f - P_A f \in P_{G_{\mathcal{N}(f)}} 0,$$

$$E_{\mathcal{N}(f)}(g) := A \cap ((B \cap \mathcal{N}(f)) - g), \quad F_{\mathcal{N}(f)}(g) := (A + g) \cap (B \cap \mathcal{N}(f)). \tag{3.3a}$$

If $|f - P_A f| \leq |b - P_A b|$ for all $b \in B$, then $f \in B$ is a global best approximation point*. Whether or not such a point exists, the following sets are well defined:*

$$G := \overline{B - A} \quad \text{and} \quad \text{for } g \in P_G 0,$$

$$E(g) := A \cap (B - g), \quad F(g) := (A + g) \cap B. \tag{3.3b}$$

From Definition 3.3 it is immediate that any global best approximation point is also a local best approximation point. Note that $P_{G_{\mathcal{N}(f)}} 0$ and $P_G 0$ are now possibly *sets*

of gap vectors since, for A convex and B prox-regular, $G = \overline{B - A}$ is not in general convex. For any $g_1, g_2 \in P_{G_{\mathcal{N}(f)}}0$, although it may happen that $g_1 \neq g_2$, it is still the case that $|g_1| = |g_2|$; hence the (local) gap between the sets A and B in the nonconvex setting is still well defined.

We will assume the following throughout the rest of this work.

Assumption 3.4 (prox-regularity of G). The set G is prox-regular at all $g \in P_G 0$, and, for every local best approximation point $f \in B$, the set $G_{\mathcal{N}(f)}$ is prox-regular at the corresponding point $g_f := f - P_A f \in P_{G_{\mathcal{N}(f)}}0$.

Example 3.5. Consider the example in $\mathbb{R}^2$ where $A = \{(0, x_2) \text{ for } x_2 \in [-2, \epsilon]\}$ for $\epsilon \geq 0$ and

$$B = \left\{ (x_1, x_2) \in \mathbb{R}^2 \;\middle|\; \begin{array}{cc} x_1 = \pm\sqrt{1 - x_2^2} & \text{for } x_2 \geq 0, \\ x_1 = -1 & \text{for } x_2 \in [0, -1], \\ x_1 = -\sqrt{1 - (x_2 + 1)^2} & \text{for } x_2 \in [-2, -1] \end{array} \right\}.$$

The corresponding set G is not prox-regular everywhere. In particular, if $\epsilon = 0$ it is not prox-regular at the point $(-\sqrt{3}/2, 1/2)$ since the projection onto G is multivalued along the line segment $(-\sqrt{3}/2, 1/2) + \tau(1, 0)$ for all $\tau \in [0, \sqrt{3}/2]$. For this example, however, this is the only point in G where prox-regularity fails. Since A and B intersect at the point $(0, -2)$, the global gap vector is $(0, 0)$, and $F = E = \{(0, -2)\}$. Each of the vectors in the set $\left\{(x_1, x_2) \in \mathbb{R}^2 \,\middle|\, x_1 = \pm\sqrt{1 - x_2^2} \quad \text{for } x_2 \geq 0, \right\}$, on the other hand, is a local gap vector relative to some neighborhood of points in B. At the point $f = (0, 1)$, for example, all of these vectors are local gap vectors of the set $G_{\mathcal{N}(f)}$, where $\mathcal{N}(f)$ is a disk of radius $\sqrt{2}$. The corresponding local best approximation points in B are $F_{\mathcal{N}(f)}(g) = \{g\}$, while the local best approximation points in A are $E_{\mathcal{N}(f)}(g) = \{(0, 0)\}$ for all $g \in \left\{(x_1, x_2) \in \mathbb{R}^2 \,\middle|\, x_1^2 + x_2^2 = 1 \text{ for } x_2 \geq 0\right\}$. We call this collection of best approximation points a *fan*, characterized by the nonuniqueness of the gap vector. We shall disallow such regions in what follows. Indeed, if in the definition of A we set $0 < \epsilon \ll 1$, then there is no such fan region and the unique best approximation point in B corresponding to $(0, \epsilon) \in A$ is $f = (0, 1)$ with corresponding unique gap vector $g = (0, 1 - \epsilon)$.

The next fact is an adjustment of [8, Proposition 2.5] for B prox-regular.

PROPOSITION 3.6. *Let A be closed convex and B prox-regular subsets of $\mathcal{H}$. Suppose that $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ are sequences in A and $B \cap \mathcal{N}(f)$, respectively, with f a local best approximation point, $\mathcal{N}(f)$ a suitable neighborhood and $b_n - a_n \to g = f - P_A f \in P_{G_{\mathcal{N}(f)}}0$. Then the following hold.*

(i) *$b_n - P_A b_n \to g$ while $P_B a_n - a_n \to \widetilde{G}_{\mathcal{N}(f)} \subset P_{G_{\mathcal{N}(f)}}0$.*

(ii) *The cluster points of $\{a_n\}_{n\in\mathbb{N}}$ and $\{P_A b_n\}_{n\in\mathbb{N}}$ belong to $E_{\mathcal{N}(f)}(g)$. The cluster points of $\{b_n\}_{n\in\mathbb{N}}$ belong to $F_{\mathcal{N}(f)}(g)$. Consequently, the cluster points of the sequences*

$$\left\{(a_n, b_n)\right\}_{n\in\mathbb{N}},\ \left\{(P_A b_n, b_n)\right\}_{n\in\mathbb{N}}$$

are local best approximation pairs relative to (A, B).

(iii) *If g is the unique gap vector on $\mathcal{N}(f)$, that is, if $P_{G_{\mathcal{N}(f)}}0 = g$, then*

$$b_n - a_n \to g \quad \Longleftrightarrow \quad |b_n - a_n| \to |g|.$$

Proof. Since

$$\begin{aligned}|b_n - a_n| &\geq \max\left\{|b_n - P_A b_n|, |P_B a_n - a_n|\right\}\\ &\geq \min\left\{|b_n - P_A b_n|, |P_B a_n - a_n|\right\}\\ &\geq |g|,\end{aligned}$$

we conclude that $\left\{|b_n - P_A b_n|\right\}_{n\in\mathbb{N}}$ and $\left\{|P_B a_n - a_n|\right\}_{n\in\mathbb{N}}$ both converge to $|\widetilde{g}|$ for any $\widetilde{g} \in P_{G_{\mathcal{N}(f)}}0$. Since A is convex, $P_A b_n$ is single-valued and continuous; hence $b_n - P_A b_n \to g$. Since B is prox-regular, P_B is possibly set-valued and $\lim_{n\to\infty} P_B a_n - a_n = \widetilde{G}_{\mathcal{N}(f)}$, a subset of $P_{G_{\mathcal{N}(f)}}0$. Hence (i) holds. Let $a \in A$ be a cluster point of $\{a_n\}_{n\in\mathbb{N}}$, say $a_{n_k} \to a$. Then $b_{n_k} \to g + a \in B \cap \mathcal{N}(f) \cap (g + A) = F_{\mathcal{N}(f)}(g)$, and hence $a \in A \cap (B \cap \mathcal{N}(f) - g) = E_{\mathcal{N}(f)}(g)$. The arguments for $\{b_n\}_{n\in\mathbb{N}}$ and $\{P_A b_n\}_{n\in\mathbb{N}}$ are similar. Finally, (iii) follows from (i) and the fact that g is the unique gap vector. ∎

Remark 3.7. Convergence of $|b_n - a_n| \to |g|$ does not, in general, imply that $b_n - a_n \to g$. To see this, consider B and A in Example 3.5 with $\epsilon = 0$. Construct the sequences $a_n := (0, -1/n)$ and $b_n := (-1, -1/n)$ for all n and let $g = (0, 1)$. Now, $a_n \to (0, 0)$, $b_n \to (-1, 0)$, and $b_n - a_n \to (-1, 0) = \widetilde{g} \neq g$, even though both belong to $P_{G_{\mathcal{N}(f)}}0$ and $|b_n - a_n| \to |g|$ when $f = (0, 1)$ and $\mathcal{N}(f)$ is a disk of radius $\sqrt{2}$. Note also that $P_B a_n \to (-1, 0)$ while $P_B a = \left\{(x_1, x_2) \in \mathbb{R}^2 \;\middle|\; x_1^2 + x_2^2 = 1 \text{ and } x_2 \geq 0\right\}$.

3.2. Relaxed averaged alternating reflections: Prox-regular sets. The difference between T_{AP} and T_{AAR} is in the control exerted on the fixed points of the respective mappings by the relaxation strategy $V(T, \beta)$. As shown in (2.6) in the case of T_{AP}, the relaxation $\beta \in (0, 1]$ simply shifts the set of fixed points from best approximation points in A to their corresponding points in B. In the nonconvex setting this shift property is restricted to local best approximation points. Hence, the relaxation parameter does not change in any significant way the set of fixed points and, in particular, it does not change the correspondence between the set of local best approximation points and the fixed points of $V(T_{AP}, \beta)$. For T_{AAR} with the relaxation parameter in (2.21) the nonconvex situation is quite different. Indeed, as we show in Lemma 3.8, the relaxation parameter β can be chosen to *limit* the set of local best approximation points that correspond to fixed points of $V(T_{AAR}, \beta)$, thus eliminating *bad* local best approximation points.

Before proceeding with the prox-regular versions of Theorem 2.7 and Corollary 2.8, we need to define what we mean by $V(T, \beta)x$ in the case when $P_B x$ is multivalued. We shall define this as

$$V(T_{AP}, \beta)x := \{v = \beta P_A b + (1 - \beta)b \quad | \quad b \in P_B x\}, \tag{3.4a}$$

$$V(T_{AAR}, \beta)x := \left\{v = \frac{\beta}{2}\left(R_A(2b - x) + x\right) + (1 - \beta)b \quad | \quad b \in P_B x\right\}. \tag{3.4b}$$

LEMMA 3.8 (characterization of fixed points). *For A closed and convex and B prox-regular, suppose that Assumption* 3.4 *holds, and define $V(T_{AAR}, \beta)$ by* (3.4b) *for $\beta \in]0, 1[$ fixed. Then*

$$\begin{aligned}&\mathsf{Fix}\, V(T_{AAR}, \beta)\\ &\quad\subset \left\{f - \frac{\beta}{1 - \beta}(f - P_A f) \mid f \in B \text{ is a local best approximation point}\right\}\end{aligned} \tag{3.5a}$$

and the two sets are equal for all $\beta \leq 1/2$. Moreover, for every $x \in \mathsf{Fix}\, V(T_{AAR}, \beta)$, there is a local best approximation point f and corresponding gap vector g_f with

$$x \in P_B x - \frac{\beta}{1-\beta} g_f, \tag{3.5b}$$

$$g_f \in P_B x - P_A R_B x. \tag{3.5c}$$

By comparison, for all $\beta \in]0,1[$ fixed,

$$\mathsf{Fix}\, V(T_{AP}, \beta) = \{ f - \beta(f - P_A f) \mid f \in B \text{ is a local best approximation point} \}. \tag{3.6}$$

Proof. The proof is almost identical to that of Lemma 2.1. We skip most of the details and point only to the main differences.

To prove (3.5a)–(3.5c) we must take account of two issues: first, that P_B might not be single-valued at all $x \in \mathsf{Fix}\, V(T_{AAR}, \beta)$ and second, that the relation (2.14) does not hold for B prox-regular. The possible multivaluedness of $P_B x$ is handled by choosing $f \in P_B x$ for a given $x \in \mathsf{Fix}\, V(T_{AAR}, \beta)$ and setting $y = x - f$. The corresponding gap vector is uniquely determined by $g_f = f - P_A(f)$. This changes (2.8) and (2.9) to *inclusions*

$$P_A(2f - x) \in P_A R_B x \quad \text{and} \quad P_A R_B x - P_B x = T_{AAR} x - x \tag{3.7}$$

by (2.7). The second equation is actually an expression of set equality. When T_{AAR} is restricted to the selection $f \in P_B x$, which we write as $T_{AAR}|_f$, this yields

$$P_A(2f - x) - f = T_{AAR}|_f x - x. \tag{3.8}$$

For $x \in \mathsf{Fix}\,(V(T_{AAR}, \beta))$, (3.7) and (3.8) give

$$(1-\beta)(x - f) = \beta(T_{AAR}|_f - x) = \beta(P_A(2f - x) - f); \tag{3.9}$$

hence, with $y = x - f$,

$$f + \frac{1-\beta}{\beta} y = P_A(2f - x). \tag{3.10}$$

This is the same result as (2.11) for the selection $f \in P_B x$. As with (2.13), using (2.12) and (3.10) we have, for any $a \in A$ nonempty, closed, and convex,

$$\frac{1}{\beta} \langle -a + f,\ y \rangle + \frac{1-\beta}{\beta^2} |y|^2 \leq 0. \tag{3.11}$$

On the other hand, since B is nonempty prox-regular and $f \in P_B x$, by Lemma 3.2 we have

$$\langle b - f, x - f \rangle \leq \sigma |f - b|^2, \tag{3.12}$$

where $x - f \in N_B(f)$, and $0 \in N_B(b)$ for b close enough to f. This yields (compare to (2.15))

$$\langle b - f,\ y \rangle = \langle b - f,\ x - f \rangle \leq \sigma |b - f|^2. \tag{3.13}$$

Now, (3.11) and (3.13) yield

$$\langle b-a,\ y\rangle \le \langle b-f,\ y\rangle - \frac{1-\beta}{\beta}|y|^2 \le \sigma\,|f-b|^2 - \frac{1-\beta}{\beta}|y|^2. \tag{3.14}$$

The right-hand side is nonpositive for all b close enough to f. The rest of the proof follows the proof of Lemma 2.1 with the caveat that the sequence $b_n \to f$ be chosen close enough to f that

$$\sigma\,|f-b_n|^2 - \frac{1-\beta}{\beta}|y|^2 \le 0 \quad \text{for all } n.$$

The identities (3.5b)–(3.5c) follow immediately since $f \in P_B x$ is a local best approximation point.

To prove that the set inequality in (3.5a) is not, in general, tight we show that, given a local best approximation point $f \in B$ and corresponding gap vector g_f,

$$f - \frac{\beta}{1-\beta}g_f \in \mathsf{Fix}\, V(T_{AAR},\beta) \quad \text{if and only if} \quad f \in P_B\left(f - \frac{\beta}{1-\beta}g_f\right). \tag{3.15}$$

The "easy" implication is that the left-hand side of (3.15) implies the right-hand side (expand $0 \in (I - V(T_{AAR},\beta))\,(f - \frac{\beta}{1-\beta}g_f)$ in terms of P_B and P_A and "solve" for $P_B(f - \frac{\beta}{1-\beta}g_f)$). The other implication follows exactly as in Lemma 2.1 with the generalization to inclusions since the projection onto B need not be single-valued.

Finally, to show that set equality holds in (3.5a) for all $\beta \le 1/2$ note that, for any local best approximation point $f \in B$ with B prox-regular, $f \in P_B(f - \frac{\beta}{1-\beta}g_f)$ for all $\beta \in [0,1/2]$, where g_f is the corresponding gap vector. The result then follows from (3.15).

For T_{AP}, since the parameter β simply shifts the fixed point within the gap of a local best approximation pair as β ranges from 0 to 1, the fixed points of $V(T_{AP},\beta)$ coincide precisely with the local best approximation points of A and B, whence (3.6). ∎

COROLLARY 3.9. *$\mathsf{Fix}\, V(T_{AAR},\beta) = \varnothing$ for all $\beta > 0$ if and only if B does not contain a local best approximation point.*

Proof. This follows immediately from (3.5). ∎

Remark 3.10. Note that, while in the case of T_{AAR} all local best approximation points correspond to fixed points of $V(T_{AAR},\beta)$ for $\beta \in [0,1/2]$, this is not the case for $\beta > 1/2$. Indeed, by (3.15) of Lemma 3.8, the fixed points of $V(T_{AAR},\beta)$ consist only of local best approximation points for which $f - \frac{\beta}{1-\beta}g_f$ is in a proximal neighborhood of B. This certainly will not hold for all $\beta \in [1/2,1[$. One might envision an algorithmic strategy for *filtering out* certain local best approximation points by choosing β large enough. Of course, how such a filtering might work in practice depends entirely on local proximal properties of the set B. The point is that, by simply increasing β, one can avoid local minima. This is a potentially powerful algorithmic tool for global projection algorithms for nonconvex problems.

We finish this section with nonconvex versions of Theorem 2.7 and Corollary 2.8. The convex results exploited the firm nonexpansiveness of the fixed point mapping, or equivalently maximal monotonicity. We show that, for the nonconvex problem, if this property holds locally, then local versions of the results of section 2.1 follow. This is not an empty assumption as Example 3.5 illustrates. Indeed, for the sets defined there with $\epsilon > 0$, an elementary calculation of $V(T_{AAR},1/2)x$ for points x on convex neighborhoods of $(0,\epsilon) \in \mathsf{Fix}\, V(T_{AAR},1/2)$ (not even very small neighborhoods)

shows that $2V(T_{AAR}, 1/2) - I$ is nonexpansive, and hence $V(T_{AAR}, 1/2)$ is locally firmly nonexpansive.

One consequence of such an assumption is the following.

PROPOSITION 3.11. *For either $T = T_{AAR}$ or $T = T_{AP}$, if $V(T, \beta)$ is firmly nonexpansive on a neighborhood $\mathcal{N}(x_0)$ of $x_0 \in \mathsf{Fix}\, V(T, \beta)$, then $g_0 = P_B x_0 - P_A P_B x_0$ is the unique gap vector on $\mathcal{N}(x_0)$; that is, for all $x \in \mathsf{Fix}\, V(T, \beta) \cap \mathcal{N}(x_0)$ one has $P_B x - P_A P_B x = P_B x_0 - P_A P_B x_0$. Moreover, P_B is single-valued on $\mathcal{N}(x_0)$.*

Proof. We prove the statement for $T = T_{AAR}$ as the proof for $T = T_{AP}$ is almost identical. Let x_1 be any fixed point on $\mathcal{N}(x_0)$ with corresponding gap vectors $g_1 \in P_B x_1 - P_A P_B x_1$, and let $b_j \in P_B x_j$, for $j = 0, 1$. Then by Lemma 3.8 $x_j = b_j - \frac{1-\beta}{\beta} g_j$, $j = 0, 1$. Since P_A is nonexpansive we have

$$|g_1 - g_0|^2 = |b_1 - b_0|^2 + |P_A b_1 - P_A b_0|^2 - 2\langle b_1 - b_0,\ P_A b_1 - P_A b_0\rangle \tag{3.16}$$
$$\leq |b_1 - b_0|^2 - |P_A b_1 - P_A b_0|^2$$

and

$$|P_A b_1 - P_A b_0| \leq |b_1 - b_0|. \tag{3.17}$$

If $|P_A b_1 - P_A b_0| = |b_1 - b_0|$, then by (3.16) $g_1 = g_0$. If, on the other hand, $|P_A b_1 - P_A b_0| < |b_1 - b_0|$, then $|y_1 - x_0| < |x_1 - x_0|$, where $y_1 := b_1 - \left(\frac{1-\beta}{\beta} + \epsilon\right) g_1$ for some $\epsilon > 0$ small enough. A straightforward calculation shows that

$$Ry_1 = b_1 - \left(\frac{1-\beta}{\beta} + (2\beta - 1)\epsilon\right) g_1 \quad \text{and} \quad Rx_0 = x_0,$$

where $R := 2V(T_{AAR}, \beta) - I$. So for $\beta < 1$, $|y_1 - x_0| < |Ry_1 - Rx_0|$, which contradicts the assumption that $V(T_{AAR}, \beta)$ is firmly nonexpansive. It must hold, then, that $|P_A b_1 - P_A b_0| = |b_1 - b_0|$; hence $g_1 = g_0$.

To see that the projection P_B is single-valued on $\mathcal{N}(x_0)$, consider any $x \in \mathcal{N}(x_0)$ and the corresponding vectors $y_1 = b_1 - x$ and $y_2 = b_2 - x$ with $b_j \in P_B x$ $(j = 1, 2)$. The same argument as above applies here with A replaced by $\{x\}$ to show that $y_1 = y_2$ since $V(T_{AAR}, \beta)$ is firmly nonexpansive; hence $b_1 = b_2$. Since x was arbitrarily chosen, this completes the proof. ∎

The inexact RAAR algorithm, Algorithm 2.3, is modified in the obvious way to inclusions for B prox-regular. For variable relaxation parameters, we then have the following generalization of Theorem 2.7.

THEOREM 3.12 (inexact prox-regular RAAR algorithm, variable β). *For A closed and convex and B prox-regular, suppose that Assumption 3.4 holds. Let $\beta \in]0, 1[$ be small enough that $\mathsf{Fix}\, V(T_{AAR}, \beta) \neq \varnothing$. Suppose that $V(T_{AAR}, \beta)$ is firmly nonexpansive on a convex neighborhood $\mathcal{N}(\overline{x})$ of $\overline{x} \in \mathsf{Fix}\, V(T_{AAR}, \beta)$ with $\operatorname{dom} V(T_{AAR}, \beta) = \mathcal{H}$. Choose $x_0 \in \mathcal{N}(\overline{x})$, let $\{\beta_n\}_{n\in\mathbb{N}}$ be a sequence in $]0, 1[$, and let $\{x_n\}_{n\in\mathbb{N}} \subset \mathcal{H}$ be generated by Algorithm 2.3 with corresponding errors $\{\epsilon_n\}_{n\in\mathbb{N}}, \{\rho_n\}_{n\in\mathbb{N}} \subset \mathcal{H}$. Define*

$$\nu_n = 2|\beta_n - \beta|\,|(P_A - I)R_B x_n| \tag{3.18}$$

and suppose that

$$\sum_{n\in\mathbb{N}} |\epsilon_n| + |\rho_n| + \nu_n = M \tag{3.19}$$

for $M \in \mathbb{R}$ small enough that $\{x_n\}_{n\in\mathbb{N}} \subset \mathcal{N}(\overline{x})$. Then $\{x_n\}_{n\in\mathbb{N}}$ converges weakly to a point in

$$\operatorname{Fix} V(T_{AAR},\beta) \cap \mathcal{N}(\overline{x}) \subset \left\{ f - \frac{\beta}{1-\beta}(f - P_A f) \;\middle|\; f \in F_{P_B\mathcal{N}(\overline{x})} \right\},$$

where $F_{P_B\mathcal{N}(\overline{x})}$ denotes the set of best approximation points in B corresponding to the projected neighborhood $P_B\mathcal{N}(\overline{x})$. Convergence is strong if any one of the following holds:

- $\underline{\lim} \operatorname{dist}_{\operatorname{Fix} V(T_{AAR},\beta)}(x_n) = 0$;
- $\operatorname{int} \operatorname{Fix} V(T_{AAR},\beta) \neq \varnothing$;
- $V(T_{AAR},\beta)$ *is* demicompact at 0.

Proof. Let $R = 2V(T_{AAR},\beta) - I$ and note that $\operatorname{Fix} R = \operatorname{Fix} V(T_{AAR},\beta)$, which, by assumption, is nonempty. Moreover, it follows from Lemma 3.8 that

$$\operatorname{Fix} R \cap \mathcal{N}(\overline{x}) = \operatorname{Fix} V(T_{AAR},\beta) \cap \mathcal{N}(\overline{x}) \subset \left\{ f - \frac{\beta}{1-\beta}(f - P_A f) \;\middle|\; f \in F_{P_B\mathcal{N}(\overline{x})} \right\}.$$

Following the proof of Theorem 2.7, let $r_n = 2x_{n+1} - x_n$ and rewrite the recursion (2.21) as

$$x_0 \in \mathcal{H} \quad \text{and} \quad x_{n+1} = \tfrac{1}{2}x_n + \tfrac{1}{2}r_n \quad \text{for all } n \in \mathbb{N}, \tag{3.20}$$

where, from (3.19) and

$$|r_n - Rx_n| \leq |\rho_n| + |\epsilon_n| + \nu_n, \tag{3.21}$$

it holds that

$$\sum_{n\in\mathbb{N}} |r_n - Rx_n| < M \tag{3.22}$$

for M small enough that $\{x_n\}_{n\in\mathbb{N}} \subset \mathcal{N}(\overline{x})$. Since $V(T_{AAR},\beta)$ is firmly nonexpansive on this neighborhood with $\operatorname{dom} V(T_{AAR},\beta) = \mathcal{H}$, it follows that R is nonexpansive on the same neighborhood with $\operatorname{dom} R = \mathcal{H}$. The result then follows from [20, Theorem 5.5]. ∎

A more detailed picture of the behavior of iterates of the exact RAAR algorithm can be obtained in the following restricted setting.

COROLLARY 3.13 (exact prox-regular RAAR algorithm in Euclidean space). *Let $\mathcal{H}$ be a Euclidean space. For the assumptions of Theorem* 3.12, *fix $\beta \in]0,1[$ small enough that $\operatorname{Fix} V(T_{AAR},\beta) \neq \varnothing$, and let*

$$x_0 \in \mathcal{N}(\overline{x}) \quad \textit{and} \quad x_{n+1} = V(T_{AAR},\beta)x_n \quad \textit{for all } n \in \mathbb{N},$$

where $\overline{x} \in \operatorname{Fix} V(T_{AAR},\beta)$ with corresponding local best approximation point $f = P_B\overline{x}$ and gap vector $g_f = f - P_A f$. Then $\{x_n\}_{n\in\mathbb{N}}$ converges to a point $x \in F_{\mathcal{N}(f)}(g_f) - \beta g_f/(1-\beta)$ and

(i) *$P_B x_n - P_A P_B x_n \to g_f$;*
(ii) *$P_B x_n \to P_B x$ and $P_A P_B x_n \to P_A P_B x$;*
(iii) *$P_B x - P_A P_B x = g_f$; hence $P_B x \in F_{\mathcal{N}(f)}(g_f)$ and $P_A P_B x \in E_{\mathcal{N}(f)}(g_f)$.*

Proof. The convergence of $\{x_n\}_{n\in\mathbb{N}}$ follows from Theorem 3.12 (with $\mu_n := \nu_n := 0$); denote the limit by x. From (3.3a) we can write $x \in F_{\mathcal{N}(f)}(g_f) - \beta g_f/(1-\beta)$ as $x = f - \beta g_f/(1-\beta)$, where $f = P_B x \in F_{\mathcal{N}(f)}(g_f)$ (see also [8, Proposition 2.4.(ii)]).

Since the mappings P_A, P_B, R_A, R_B are continuous, (ii) follows. Next, using (3.5c), we have

$$P_B x_n - P_A R_B x_n \to P_B x - P_A R_B x = g_f. \tag{3.23}$$

Hence

$$|g_f| \leq |P_B x_n - P_A P_B x_n| \leq |P_B x_n - P_A R_B x_n| \to |g_f|, \tag{3.24}$$

and thus $|P_B x_n - P_A P_B x_n| \to |g_f|$. Now (i) follows from Propositions 3.6 and 3.11. Taking the limit in (i) yields (iii). □

4. Conclusion and open problems. In this work we have laid some groundwork for a comprehensive theory of the asymptotic behavior of projection algorithms in prox-regular settings, with particular focus on the RAAR algorithm. The RAAR algorithm has many attractive features, namely that it is robust for consistent and inconsistent problems, the relaxation parameter can be interpreted as a *step length* and thus can be optimized, and, moreover, the relaxation parameter can be used to avoid "bad" local minima. In the convex setting the RAAR algorithm, together with the classical alternating projections and averaged projections algorithms, can be viewed as instances of the classical Lions–Mercier/Douglas–Rachford algorithm applied to the problem of minimizing the sum of two maximal monotone mappings; hence the analysis of the RAAR algorithm can be broadly applied. We conjecture that these correspondences carry over to the nonconvex setting; however, the details of this correspondence are beyond the scope of the present study.

The analytical tools that we use derive from analogues in convex theory. As one would expect from nonconvex problems, our most general results are local in nature. Our hope is that this analysis can serve as a guide to the analysis of similar algorithms. For the purpose of proving the convergence of the RAAR algorithm for the phase retrieval problem in crystallography, it remains to be shown that the fixed point mapping $V(T_{AAR}, \beta)$ is firmly nonexpansive on a neighborhood of a fixed point. This question would be quickly resolved by sufficient conditions under which firm nonexpansiveness holds in nonconvex settings. Less restrictive notions of firm nonexpansiveness, namely *quasi*-1/2-*averaged* mappings as studied in [5, 20], could also be quite fruitful here. Another key assumption for our results was the prox-regularity of the set $G_{\mathcal{N}(f)} = \overline{B \cap \mathcal{N}(f) - A}$ at local best approximation points f. We conjecture that this assumption as well as the assumption of local firm nonexpansiveness of the corresponding reflection can be removed on neighborhoods of local best approximation points. The assumption that one of the sets is convex, not just prox-regular, was useful for our proofs, but is probably not necessary in general. With the exception of what we called fan regions, local best approximation points will by definition be in the proximal neighborhood of the other set. We therefore conjecture that these results can be extended to two prox-regular sets. Finally, we note that our use of hypomonotonicity defined in Lemma 3.1 might be relaxed to approximate monotonicity of the closed set C at a point $\overline{x} \in C$ defined by

$$\langle y_1 - y_2,\ x_1 - x_2 \rangle \geq -\sigma |x_1 - x_2| \quad \text{whenever } y_i \in N_C^r(x_i) \text{ and } x_i \in \mathcal{N}(\overline{x})$$

for all $\sigma > 0$. By [35, Corollary 4.11] this is equivalent to the super-regularity of C [35, Definition 4.4], a weaker condition than prox-regularity. This generalization would immediately extend our results to sets with other types of regularity such as subsmooth sets [1].

Acknowledgments. The author would like to thank the anonymous reviewers for their thoughtful comments, particularly with regard to the interpretation of RAAR as a regularization rather than relaxation.

REFERENCES

[1] D. Aussel, A. Daniilidis, and L. Thibault, *Subsmooth sets: Functional characterizations and related concepts*, Trans. Amer. Math. Soc., 357 (2004), pp. 1275–1301.

[2] H. H. Bauschke and J. M. Borwein, *On the convergence of von Neumann's alternating projection algorithm for two sets*, Set-Valued Anal., 1 (1993), pp. 185–212.

[3] H. H. Bauschke and J. M. Borwein, *Dykstra's alternating projection algorithm for two sets*, J. Approx. Theory, 79 (1994), pp. 418–443.

[4] H. H. Bauschke and J. M. Borwein, *On projection algorithms for solving convex feasibility problems*, SIAM Rev., 38 (1996), pp. 367–426.

[5] H. H. Bauschke and P. L. Combettes, *A weak-to-strong convergence principle for Fejér-monotone methods in Hilbert space*, Math. Oper. Res., 26 (2001), pp. 248–264.

[6] H. H. Bauschke, P. L. Combettes, and D. R. Luke, *Phase retrieval, error reduction algorithm and Fienup variants: A view from convex feasibility*, J. Opt. Soc. Amer. A., 19 (2002), pp. 1334–1345.

[7] H. H. Bauschke, P. L. Combettes, and D. R. Luke, *A hybrid projection reflection method for phase retrieval*, J. Opt. Soc. Amer. A., 20 (2003), pp. 1025–1034.

[8] H. H. Bauschke, P. L. Combettes, and D. R. Luke, *Finding best approximation pairs relative to two closed convex sets in Hilbert spaces*, J. Approx. Theory, 127 (2004), pp. 178–192.

[9] J. M. Borwein, S. Reich, and I. Shafrir, *Krasnoselski-Mann iterations in normed spaces*, Canad. Math. Bull., 35 (1992), pp. 21–28.

[10] H. Brézis and P.-L. Lions, *Produits infinis de resolvantes*, Israel J. Math., 29 (1978), pp. 329–345.

[11] M. Brignone, J. Coyle, and M. Piana, *The use of the linear sampling method for obtaining super-resolution effects in Born approximation*, J. Comput Appl. Math., 203 (2007), pp. 145–158.

[12] M. Brignone and M. Piana, *The use of constraints for solving inverse scattering problems: Physical optics and the linear sampling method*, Inverse Problems, 21 (2005), pp. 207–222.

[13] J. V. Burke and D. R. Luke, *Variational analysis applied to the problem of optical phase retrieval*, SIAM J. Control Optim., 42 (2003), pp. 576–595.

[14] Y. Censor and S. A. Zenios, *Parallel Optimization: Theory Algorithms and Applications*, Oxford University Press, Oxford, UK, 1997.

[15] R. Chapko and R. Kress, *A hybrid method for inverse boundary value problems in potential theory*, J. Inverse Ill-Posed Probl., 13 (2005), pp. 27–40.

[16] W. Cheney and A. A. Goldstein, *Proximity maps for convex sets*, Proc. Amer. Math. Soc., 10 (1959), pp. 448–450.

[17] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Classics Appl. Math. 5, SIAM, Philadelphia, 1990.

[18] F. H. Clarke, R. J. Stern, Y. S. Ledyaev, and P. R. Wolenski, *Nonsmooth Analysis and Control Theory*, Grad. Texts in Math. 178, Springer-Verlag, New York, 1998.

[19] P. L. Combettes, *The convex feasibility problem in image recovery*, in Advances in Imaging and Electron Physics, Vol. 95, P. W. Hawkes, ed., Academic Press, New York, 1996, pp. 155–270.

[20] P. L. Combettes, *Quasi-Fejérian analysis of some optimization algorithms*, in Inherently Parallel Algorithms for Feasibility and Optimization, D. Butnariu, Y. Censor, and S. Reich, eds., North-Holland/Elsevier, Amsterdam, 2001, pp. 115–152.

[21] P. L. Combettes, *Solving monotone inclusions via compositions of nonexpansive averaged operators*, Optimization, 53 (2004), pp. 475–504.

[22] P. L. Combettes and H. J. Trussell, *Method of successive projections for finding a common point of sets in metric spaces*, J. Optim. Theory Appl., 67 (1990), pp. 487–507.

[23] F. Deutsch, *Best Approximation in Inner Product Spaces*, CMS Books Math./Ouvrages Math. SMC 7, Springer-Verlag, New York, 2001.

[24] J. Douglas and H. H. Rachford, *On the numerical solution of heat conduction problems in two or three space variables*, Trans. Amer. Math. Soc., 82 (1956), pp. 421–439.

[25] J. Eckstein and D. P. Bertsekas, *On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators*, Math. Programming, 55 (1992), pp. 293–318.

[26] H. Federer, *Curvature measures*, Trans. Amer. Math. Soc., 93 (1959), pp. 418–491.
[27] A. E. Frazho, K. M. Grigoriadis, and S. M. Kherat, *Alternating projection methods for mixed H^2 and H^∞ Nehari problems*, IEEE Trans. Automat. Control, 40 (1995), pp. 2127–2131.
[28] K. Goebel and W. A. Kirk, *Topics in Metric Fixed Point Theory*, Cambridge University Press, Cambridge, UK, 1990.
[29] E. G. Gol'shtein and N. Tret'yakov, *Modified Lagrangians in convex programming and their generalizations*, Math. Programming Stud., 10 (1979), pp. 86–97.
[30] A. D. Ioffe, *Approximate subdifferentials and applications*: II, Mathematika, 33 (1986), pp. 111–128.
[31] A. D. Ioffe, *Approximate subdifferentials and applications*: III: *The metric theory*, Mathematika, 36 (1989), pp. 1–38.
[32] A. D. Ioffe, *Proximal analysis and approximate subdifferentials*, J. London Math Soc. (2), 41 (1990), pp. 175–192.
[33] R. Kress and W. Rundell, *Nonlinear integral equations and the iterative solution for an inverse boundary value problem*, Inverse Problems, 21 (2005), pp. 1207–1223.
[34] R. Kress and P. Serranho, *A hybrid method for two-dimensional crack reconstruction*, Inverse Problems, 21 (2005), pp. 773–784.
[35] A. S. Lewis, D. R. Luke, and J. Malick, *Local Linear Convergence of Alternating and Averaged Projections*, preprint, 2007.
[36] A. S. Lewis and J. Malick, *Alternating projections on manifolds*, Math. Oper. Res., 33 (2008), pp. 216–234.
[37] P. L. Lions and B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM J. Numer. Anal., 16 (1979), pp. 964–979.
[38] D. R. Luke, *Relaxed averaged alternating reflections for diffraction imaging*, Inverse Problems, 21 (2005), pp. 37–50.
[39] D. R. Luke, J. V. Burke, and R. G. Lyon, *Optical wavefront reconstruction: Theory and numerical methods*, SIAM Rev., 44 (2002), pp. 169–224.
[40] B. Martinet, *Régularisation d'inéquations variationnelles par approximations successives*, Rev. Française Informat. Recherche Opérationnelle Sér. R-3, 4 (1970), pp. 154–158.
[41] B. Martinet, *Détermination approchée d'un point fixe d'une application pseudo-contractante. Cas de l'application prox*, C. R. Acad. Sci. Paris Sér. A, 274 (1972), pp. A163–A165.
[42] G. J. Minty, *Monotone (nonlinear) operators in Hilbert space*, Duke. Math. J., 29 (1962), pp. 341–346.
[43] B. S. Mordukhovich and Y. Shao, *Nonsmooth sequential analysis in Asplund spaces*, Trans. Amer. Math. Soc., 348 (1996), pp. 1235–1280.
[44] J. J. Moreau, *Proximité et dualité dans un espace Hilbertian*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.
[45] A. Moudafi, *An algorithmic approach to prox-regular variational inequalities*, Appl. Math. Comput., 155 (2004), pp. 845–852.
[46] G. Pierra, *Éclatement de contraintes en parallèle pour la minimisation d'une forme quadratique*, in Proceedings of the 7th IFIP Conference on Optimization Techniques, Lecture Notes in Comput. Sci. 41, Springer-Verlag, London, 1975, pp. 200–218.
[47] R. A. Poliquin and R. T. Rockafellar, *Generalized Hessian properties of regularized nonsmooth functions*, SIAM J. Optim., 6 (1996), pp. 1121–1137.
[48] R. A. Poliquin and R. T. Rockafellar, *Prox-regular functions in variational analysis*, Trans. Amer. Math. Soc., 348 (1996), pp. 1805–1838.
[49] R. A. Poliquin, R. T. Rockafellar, and L. Thibault, *Local differentiability of distance functions*, Trans. Amer. Math. Soc., 352 (2000), pp. 5231–5249.
[50] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.
[51] R. T. Rockafellar and R. J. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[52] A. N. Tihonov, *On the regularization of ill-posed problems*, Dokl. Akad. Nauk SSSR, 153 (1963), pp. 49–52 (in Russian).
[53] A. N. Tihonov, *On the solution of ill-posed problems and the method of regularization*, Dokl. Akad. Nauk SSSR, 151 (1963), pp. 501–504 (in Russian).
[54] J. von Neumann, *On rings of operators, reduction theory*, Ann. of Math. (2), 50 (1949), pp. 401–485.

SIAM J. OPTIM.
Vol. 19, No. 2, pp. 740–755

PITCHFORK BIFURCATIONS IN NONLINEAR LEAST SQUARES PROBLEMS–APPLICATIONS TO TOA GEOLOCATION*

LOUIS A. ROMERO† AND JEFF MASON‡

Abstract. We prove a general result concerning nonlinear least squares problems in the neighborhood of a singular Hessian. Assuming the Hessian has a one-dimensional null space, we show that in the neighborhood of such a point, there will be three solutions with small residuals. Counter to our intuition, assuming certain transversality conditions are satisfied, the solution with the largest residual is actually the most accurate solution. We illustrate this general theory by applying it to a time of arrival (TOA) geolocation problem.

Key words. bifurcations, nonlinear least squares, geolocation, TOA, TDOA, global positioning system

AMS subject classifications. 62L12, 65H10, 65F15

DOI. 10.1137/070697598

1. Introduction.

The results in this paper were motivated by phenomena observed while solving overdetermined systems of equations arising in time of arrival (TOA) geolocation as described in section 6 and [10, 7, 5]. TOA geolocation is used in a variety of settings including global positioning systems (GPS). As part of a program to make the relevant algorithms as robust as possible, we sought configurations of satellites where the algorithms were severely tested, in particular, where the Hessian (3.4) was singular or nearly singular. In the language of GPS, we were seeking out situations where the geometrical dilution of precision (GDOP) was very high. Here we observed a phenomena that seems counterintuitive, and that we have not seen reported elsewhere in the literature.

To describe this phenomena, we consider a system of equations

$$\mathbf{f}(\mathbf{z}, \mathbf{d}) = \mathbf{0}, \tag{1.1}$$

where $\mathbf{z}$ is a vector that we are trying to determine, $\mathbf{d}$ is a vector of data, and $\mathbf{f}$ is a nonlinear function of $\mathbf{z}$. Assuming we have more equations than unknowns, this gives us an overdetermined system of nonlinear equations that can be solved using least squares [4]. That is, we try to minimize

$$P(\mathbf{z}) = \frac{1}{2}\mathbf{f}^T(\mathbf{z})\mathbf{f}(\mathbf{z}). \tag{1.2}$$

Here it is assumed that the weighting has been absorbed into our equations.

In the absence of noise, we have data $\mathbf{d}_0$ such that there is a scripted solution $\mathbf{z}_0$ that exactly solves the overdetermined system

$$\mathbf{f}(\mathbf{z}_0, \mathbf{d}_0) = \mathbf{0}. \tag{1.3}$$

*Received by the editors July 17, 2007; accepted for publication (in revised form) February 28, 2008; published electronically July 3, 2008. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the US Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

http://www.siam.org/journals/siopt/19-2/69759.html

†Computational Mathematics and Algorithms Department, Sandia National Laboratories, MS 1320, P.O. Box 5800, Albuquerque, NM 87123-1320 (lromero@sandia.gov).

‡Radar and Signal Analysis Department, Sandia National Laboratories, MS 0519, P.O. Box 5800, Albuquerque, NM 87123-1320 (jjmason@sandia.gov).

We now add noise to our system, but for the sake of analysis, we do this in a deterministic way. That is, we suppose that the data is given by

$$\mathbf{d} = \mathbf{d}_0 + \epsilon \mathbf{n}, \tag{1.4}$$

where ϵ gives the level of the noise, and $\mathbf{n}$ is a unit vector that is random, but held fixed during our analysis. We are now ready to state the counterintuitive results that we have encountered.

When we put the satellites in a configuration where the Hessian evaluated at $\mathbf{z}_0$ is singular (with a one-dimensional null space), and then add noise to the data, we observe that typically there are three critical points of our least squares problem that are all close to $\mathbf{z}_0$. Two of these, which we will call $\mathbf{z}_+$ and $\mathbf{z}_-$, differ from $\mathbf{z}_0$ by an amount that is proportional to the square root of the noise.

$$\mathbf{z}_\pm = \mathbf{z}_0 + \nu_\pm \boldsymbol{\phi} + O(\epsilon), \quad \nu_\pm = \pm constant \times \sqrt{\epsilon}. \tag{1.5}$$

The third answer, which we call $\mathbf{z}_M$, differs from the scripted answer by an amount that is proportional to the noise.

$$\mathbf{z}_M = \mathbf{z}_0 + \epsilon \mathbf{q}_0 + \nu_M \boldsymbol{\phi} + O(\epsilon^2), \quad \nu_M = constant \times \epsilon. \tag{1.6}$$

In these equations $\boldsymbol{\phi}$ is the null vector of the Hessian in the absence of noise. For small noise levels, ϵ will be much smaller than $\sqrt{\epsilon}$; hence $\mathbf{z}_M$ is much more accurate than the solutions $\mathbf{z}_\pm$, and is clearly the answer that we would like to report. However, of the three candidates for the solution, we will show that $\mathbf{z}_M$ always gives the largest residual P (as defined in (1.2)).

$$P(\mathbf{z}_M, \mathbf{d}) > P(\mathbf{z}_\pm, \mathbf{d}). \tag{1.7}$$

This clearly violates our intuition since we feel as though the best solution should be the one with the minimum residual, which it is not.

At first sight it may seem impossible to know which solution to report (since in practice we do not know $\mathbf{z}_0$). However, it can be shown that $\mathbf{z}_M$ is much less sensitive to changes in the data. This sensitivity can be computed without knowing the scripted solution, and hence this offers a viable way of determining which solution to report. Alternatively, (1.5) and (1.6) show that if we average all three of these solutions we get a solution that differs from the scripted solution by $O(\epsilon)$. This suggests that if we see three solutions close to each other, the average may be an accurate estimate of the solution we seek.

In this paper we use techniques from bifurcation theory to show that this is in fact a general phenomenon that occurs when solving nonlinear least squares problems. In particular, we use what is known as the Liapanov–Schmidt reduction [2, 6, 3, 9] to effectively turn this into a one-dimensional problem. This reduction allows us to solve a one-dimensional equation for the determination of the constant ν in front of the eigenvector $\boldsymbol{\phi}$ in (1.5) and (1.6). We will see that to leading order, ν satisfies an equation

$$c_0 \epsilon^2 + 2a_1 \epsilon \nu + 4a_3 \nu^3 = 0, \tag{1.8}$$

where

$$a_3 > 0, \tag{1.9}$$

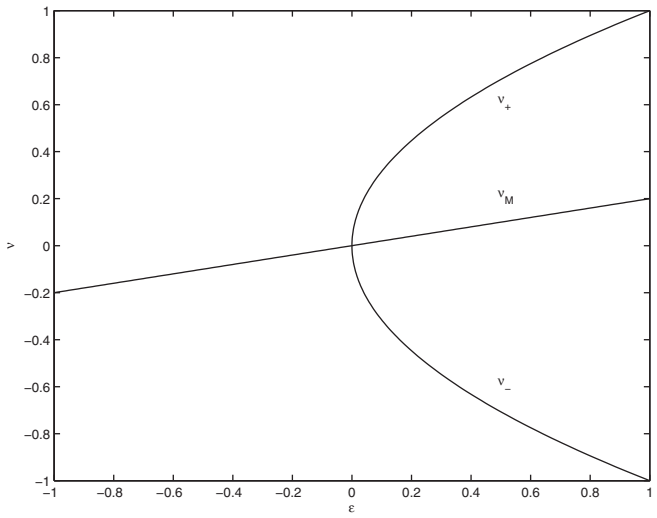


FIG. 1.1. *This is a schematic illustrating the pitchfork bifurcation described by* (1.8). *Note that this is not a symmetry breaking bifurcation. That is, it is not symmetrical except for asymptotically small values of* ϵ.

and the objective function can be written as

$$P(\nu, \epsilon) = a_0\epsilon^2 + a_1\epsilon\nu^2 + a_3\nu^4 + \text{Higher order terms.} \tag{1.10}$$

Note that (1.8) is missing the linear term in ϵ and both the linear and quadratic terms in ν. This is a general property of nonlinear least squares problems in the vicinity of a singularity. An analysis of (1.8) shows that it exhibits a pitchfork bifurcation, having three real solutions for one sign of ϵ, and one real solution for the opposite sign. The form in (1.10) for the residual P is noteworthy in that two of the constants appearing in it are the same as the constants in (1.8). This, along with the fact that $a_3 > 0$, will allow us to prove that the solutions $\mathbf{z}_\pm$ have lower residuals than $\mathbf{z}_M$ for small values of ϵ.

Figure 1.1 illustrates the pitchfork bifurcation described by (1.8). It should be noted that this is not a symmetry-breaking bifurcation. That is, the bifurcation is only symmetrical for asymptotically small values of ϵ.

A relevant question is: How likely are we to see such degenerate configurations? In order for such a configuration to exist, we will see that we must have a vector $\boldsymbol{\phi}$ such that $\mathbf{J}\boldsymbol{\phi} = 0$, where $\mathbf{J}$ is the Jacobian of $\mathbf{f}$. If we have $N+1$ equations in N unknowns, $\mathbf{J}$ will be an $(N+1)\times N$ matrix, and typically we cannot find a nontrivial vector $\boldsymbol{\phi}$ that satisfies $\mathbf{J}\boldsymbol{\phi} = 0$. However, such a solution is generic if we adjust two parameters. For example, in a GPS problem we have four unknowns: the three-dimensional vector giving the location of the receiver, and the clock bias. If we have five satellites this gives us five equations in four unknowns. If we fix the location of the receiver, then it is not likely that at some point in time the five satellites will pass through such a configuration. However, it is likely that at some point in time there is a point on a curve (such as the equator) that will see the satellites as being degenerate. On the other hand, if we have $N + 2$ equations in N unknowns, we will need to adjust three parameters.

We now describe the outline of this paper. In section 2 we give a review of the Liapanov–Schmidt reduction in bifurcation theory. In section 3 we apply this theory

to the case of nonlinear least squares problems, in section 4 we discuss the residuals of the solutions in the neighborhood of the singular point, and in section 5 we discuss a criterion for choosing the best solution. In section 6 we give examples of this phenomena occurring in geolocation problems, and in section 7 we give conclusions.

2. Review of the Liapanov–Schmidt reduction. We will carry out the bifurcation analysis using a standard technique in bifurcation theory known as the Liapanov–Schmidt reduction. Though this procedure can be found in many different sources [2, 6, 3, 9], we have chosen to review it here for the convenience of the reader, and to introduce the notation that we will be using. We limit our discussion of this procedure to systems that have a simple zero eigenvalue. This technique allows us to analyze an N-dimensional system near a bifurcation point by analyzing a one-dimensional equation that solves for the component ν of the solution in the direction of the critical eigenvector.

Suppose we have N nonlinear equations in N unknowns

$$\mathbf{G}(\mathbf{z}, \epsilon) = \mathbf{0}, \tag{2.1}$$

where ϵ is a parameter, $\mathbf{G}$ is an N-dimensional vector, and we are solving for the N-dimensional vector $\mathbf{z}$. We suppose that for $\epsilon = 0$ we have a solution $\mathbf{z}_0$, and that the Jacobian of this system is singular at $(\mathbf{z}, \epsilon) = (\mathbf{z}_0, 0)$, with a simple real eigenvector $\boldsymbol{\phi}$, and adjoint eigenvector $\boldsymbol{\psi}$. This is equivalent to making the following assumptions:

Assumptions 1 (assumptions for Liapanov–Schmidt reduction). *We assume that for* $(\mathbf{z}, \epsilon) = (\mathbf{z}_0, 0)$ *we have*

$$\mathbf{G}(\mathbf{z}_0, 0) = \mathbf{0}. \tag{2.2}$$

The Jacobian $\frac{\partial \mathbf{G}}{\partial \mathbf{z}}$ *has a simple zero eigenvector* $\boldsymbol{\phi}$

$$\frac{\partial \mathbf{G}}{\partial \mathbf{z}}(\mathbf{z}_0, 0)\boldsymbol{\phi} = \mathbf{0} \tag{2.3}$$

and a left eigenvector

$$\boldsymbol{\psi}^T \frac{\partial \mathbf{G}}{\partial \mathbf{z}}(\mathbf{z}_0, 0) = \mathbf{0}^T. \tag{2.4}$$

The simplicity requirement requires that

$$\boldsymbol{\psi}^T \boldsymbol{\phi} \neq 0. \tag{2.5}$$

For the problems we will be concerned with, the Jacobian will be symmetric, and we will have $\boldsymbol{\psi} = \boldsymbol{\phi}$. However, we will not make this assumption in this brief review.

Since our Jacobian is singular at $(\mathbf{z}_0, 0)$, we have a violation of the conditions of the implicit function theorem [1], and we are not guaranteed that we can uniquely solve for $\mathbf{z}(\epsilon)$ in the neighborhood of such a point. However, since we have a simple zero eigenvalue, it can be shown that we can solve for $N - 1$ of the components of $\mathbf{z}$ in terms of the component ν of $\mathbf{z}$ that is in the direction of the critical eigenvector $\boldsymbol{\phi}$. Once we have expressed the other components in terms of ν, we can reduce the N-dimensional problem to a one-dimensional problem.

To be precise we will write our solution as

$$\mathbf{z} = \mathbf{z}_0 + \mathbf{v} + \nu\boldsymbol{\phi} \tag{2.6}$$

and require that

$$\boldsymbol{\psi}^T \mathbf{v} = 0. \tag{2.7}$$

These equations express the fact that ν is the component of $\mathbf{z} - \mathbf{z}_0$ in the direction of $\boldsymbol{\phi}$, and that $\mathbf{v}$ gives the components in the other directions.

We would like to solve for $\mathbf{v}(\nu, \epsilon)$ in the neighborhood of $(\mathbf{z}_0, 0)$. To do that we will require that all of the components of the equation $\mathbf{G}(\mathbf{z}, \epsilon) = \mathbf{0}$ are satisfied, except for the component in the direction $\boldsymbol{\phi}$. We can do this by using the equation

$$\mathbf{G}(\mathbf{z}_0 + \mathbf{v} + \nu\boldsymbol{\phi}, \epsilon) = \kappa\boldsymbol{\phi}. \tag{2.8}$$

Here κ is a variable that we will solve for. By introducing κ, (2.8) does not require that $\mathbf{G}(\mathbf{z}, \epsilon) = 0$ in the direction of $\boldsymbol{\phi}$. Equations (2.7) and (2.8) give us $N + 1$ equations in the $N + 1$ unknowns $\mathbf{v}$ and κ, with ν and ϵ as parameters. These equations are satisfied by $\mathbf{v} = \kappa = 0$, when $\epsilon = \nu = 0$. Furthermore, when we linearize about this point, it can be shown that the Jacobian of this extended system is nonsingular with respect to the variables $\mathbf{v}$ and κ (we omit the details here). This implies that we can uniquely solve for $\mathbf{v}(\nu, \epsilon)$ and $\kappa(\nu, \epsilon)$.

LEMMA 2.1. *Under the assumptions* (1), (2.7) *and* (2.8) *have a unique solution* $\mathbf{v}(\nu, \epsilon)$, $\kappa(\nu, \epsilon)$ *for* (ν, ϵ) *in the neighborhood of* $(0, 0)$, *and* $\mathbf{v}$ *in the neighborhood of* $\mathbf{0}$.

For any values of ν and ϵ in the neighborhood of $(0, 0)$, we will have

$$\mathbf{q}^T \mathbf{G}(\mathbf{z}_0 + \mathbf{v}(\nu, \epsilon) + \nu\boldsymbol{\phi}, \epsilon) = 0; \text{ for } \mathbf{q}^T\boldsymbol{\phi} = 0. \tag{2.9}$$

This implies that $\mathbf{z} = \mathbf{z}_0 + \mathbf{v}(\nu, \epsilon) + \nu\boldsymbol{\phi}$ will satisfy the equation $\mathbf{G}(\mathbf{z}, \epsilon) = \mathbf{0}$ in all directions except for one, the direction of $\boldsymbol{\phi}$. In order to ensure that $\mathbf{z} = \mathbf{z}_0 + \mathbf{v}(\nu, \epsilon) + \nu\boldsymbol{\phi}$ is in fact a solution to $\mathbf{G}(\mathbf{z}, \epsilon) = 0$, we need only require that the equation is also satisfied in the direction $\boldsymbol{\phi}$. To do this we require that

$$g(\nu, \epsilon) = \boldsymbol{\psi}^T \mathbf{G}(\mathbf{z}_0 + \mathbf{v}(\nu, \epsilon) + \nu\boldsymbol{\phi}, \epsilon) = 0. \tag{2.10}$$

This last equation gives us a one-dimensional equation $g(\nu, \epsilon) = 0$ to solve. This is the end result of the Liapanov–Schmidt reduction. Our main goal is to determine the topology of the solution space in the neighborhood of $(\mathbf{z}_0, 0)$. That is, we would like to know how many solutions there are as a function of ϵ, and what the leading order behavior of these solutions is. Through our reduction, this is equivalent to determining the topology of the solutions to $g(\nu, \epsilon) = 0$ near $(\nu, \epsilon) = (0, 0)$. This can be determined if we know the leading order terms in the Taylor series expansion of $g(\nu, \epsilon)$, which can be found by taking derivatives of g. That is, the process we carry out in the next section for the case where $\mathbf{G}$ comes from a nonlinear least squares problem.

A general property of this Liapanov–Schmidt reduction is that $\frac{\partial \mathbf{v}}{\partial \nu}(0, 0) = 0$. This follows from taking the derivative of (2.7) and (2.8) with respect to ν. If we do this we get

$$\frac{\partial \mathbf{G}}{\partial \mathbf{z}} \left(\frac{\partial \mathbf{v}}{\partial \nu} + \boldsymbol{\phi} \right) = \frac{\partial \kappa}{\partial \nu} \boldsymbol{\phi}, \tag{2.11}$$

$$\boldsymbol{\psi}^T \frac{\partial \mathbf{v}}{\partial \nu} = 0. \tag{2.12}$$

If we evaluate these equations at $\nu = \epsilon = 0$, and use the fact that at that point $\frac{\partial \mathbf{G}}{\partial \mathbf{z}}\boldsymbol{\phi} = 0$, we get the equation

$$\frac{\partial \mathbf{G}}{\partial \mathbf{z}}\frac{\partial \mathbf{v}}{\partial \nu} = \frac{\partial \kappa}{\partial \nu}\boldsymbol{\phi}. \tag{2.13}$$

If we multiply both sides of this equation on the left by $\boldsymbol{\psi}^T$ and use (2.4) and (2.5), we see that $\frac{\partial \kappa}{\partial \nu} = 0$. It follows that $\frac{\partial \mathbf{G}}{\partial \mathbf{z}}\frac{\partial \mathbf{v}}{\partial \nu} = 0$ and hence $\frac{\partial \mathbf{v}}{\partial \nu}$ is proportional to $\boldsymbol{\phi}$, but (2.12) shows that the constant of proportionality must in fact be zero.

LEMMA 2.2. *Under the assumptions* (1)*, we have*

$$\frac{\partial \mathbf{v}}{\partial \nu} = 0 \quad \textit{for } \nu = \epsilon = 0. \tag{2.14}$$

Collecting all of this we get the following:

THEOREM 2.3 (Liapanov–Schmidt). *Under the assumptions* (1)*, as* $\epsilon \to 0$*, any solution* $\mathbf{z}(\epsilon)$ *(in the neighborhood of* $\mathbf{z}_0$*) to* $\mathbf{G}(\mathbf{z}, \epsilon) = 0$ *can be written as*

$$\mathbf{z}(\epsilon) = \mathbf{z}_0 + \mathbf{v}(\epsilon, 0) + \nu(\epsilon)\boldsymbol{\phi} + \ldots, \tag{2.15}$$

where $\nu(\epsilon)$ *is a solution to* $g(\nu, \epsilon) = 0$ *as defined in* (2.10)*. Similarly, any solution to* $g(\nu, \epsilon) = 0$ *gives a solution to* $\mathbf{G}(\mathbf{z}, \epsilon) = 0$*, whose leading order behavior is as in* (2.15)*.*

3. Bifurcation analysis. In this section we apply the Liapanov–Schmidt procedure to the specific case where the function $\mathbf{G}(\mathbf{z}, \epsilon)$ (as in the last section) comes from solving an overdetermined system of equations $\mathbf{f}(\mathbf{z}, \epsilon) = 0$ in a least squares sense. The analysis is applied to the case where the Jacobian of our system has a single null vector. In Lemma (3.2) we show that in this case the function $g(\nu, \epsilon)$ (as in (2.10)) is missing the order ϵ, ν, and ν^2 terms in the Taylor series about $(\nu, \epsilon) = (0, 0)$. This gives us the form as in (1.8). We will see that this implies that for small values of ϵ we have three solutions as in (1.5) and (1.6).

We assume we have a suitably differentiable function $\mathbf{f}(\mathbf{z}, \epsilon)$ where $\mathbf{z}$ is an N-dimensional vector, and $\mathbf{f}(\mathbf{z}, \epsilon)$ is an M-dimensional vector, with $M > N$. Here the parameter ϵ is meant to specify the level of noise in our system. We will solve this system of equations using nonlinear least squares. That is, we minimize the objective function

$$P(\mathbf{z}, \epsilon) = \frac{1}{2}\mathbf{f}^T(\mathbf{z}, \epsilon).\mathbf{f}(\mathbf{z}, \epsilon). \tag{3.1}$$

The equations for $\mathbf{z}$ come from setting the gradient of this to zero. This gives us the equations

$$\mathbf{G}(\mathbf{z}, \epsilon) = \mathbf{J}^T(\mathbf{z}, \epsilon)\mathbf{f}(\mathbf{z}, \epsilon) = \mathbf{0}, \tag{3.2}$$

where

$$\mathbf{J}(\mathbf{z}, \epsilon) = \frac{\partial \mathbf{f}(\mathbf{z}, \epsilon)}{\partial \mathbf{z}} \tag{3.3}$$

is the Jacobian of $\mathbf{f}$.

In general, the Hessian matrix is given by

$$\mathbf{H}(\mathbf{z}, \epsilon) = \frac{\partial \mathbf{J}^T}{\partial \mathbf{z}}\mathbf{f} + \mathbf{J}^T\mathbf{J}. \tag{3.4}$$

Here $\frac{\partial \mathbf{J}^T}{\partial \mathbf{z}}$ is a third rank tensor. Since we will always be using the Hessian when $\mathbf{f} = \mathbf{0}$, we do not need to further specify what we mean by this quantity.

We will assume that when there is no noise, there is a solution $\mathbf{z}_0$ that satisfies the equations exactly. That is, we have

$$\mathbf{f}(\mathbf{z}_0, 0) = \mathbf{0}. \tag{3.5}$$

Furthermore, we assume that we are at a point where the Hessian $\mathbf{H}$ is singular. Since $\mathbf{f} = \mathbf{0}$, this means that the Hessian is given by $\mathbf{H} = \mathbf{J}^T\mathbf{J}$. The only way this can be singular is if there is a vector $\boldsymbol{\phi}$ such that

$$\mathbf{J}(\mathbf{z}_0, 0)\boldsymbol{\phi} = \mathbf{0}. \tag{3.6}$$

We make the assumption that there is only one vector $\boldsymbol{\phi}$ in the null space of $\mathbf{J}$.

We now carry out the Liapanov–Schmidt reduction for equations of the form (3.2). In particular, we will show that the function $g(\nu, \epsilon)$ in the Liapanov–Schmidt reduction has no linear terms, and that it is missing the term ν^2. We will write the function $g(\nu, \epsilon)$ as

$$g(\nu, \epsilon) = \boldsymbol{\phi}^T \hat{\mathbf{J}}^T(\nu, \epsilon)\hat{\mathbf{f}}(\nu, \epsilon), \tag{3.7}$$

where

$$\hat{\mathbf{f}}(\nu, \epsilon) = \mathbf{f}(z_0 + \mathbf{v}(\nu, \epsilon) + \nu\boldsymbol{\phi}, \epsilon) \quad \text{and} \tag{3.8}$$

$$\hat{\mathbf{J}}(\nu, \epsilon) = \mathbf{J}(z_0 + \mathbf{v}(\nu, \epsilon) + \nu\boldsymbol{\phi}, \epsilon). \tag{3.9}$$

In our case (2.8) can be written as

$$\hat{\mathbf{J}}^T(\nu, \epsilon)\hat{\mathbf{f}}(\nu, \epsilon) = \kappa(\nu, \epsilon)\boldsymbol{\phi}. \tag{3.10}$$

The calculation of the low-order derivatives of $g(\nu, \epsilon)$ is simple due to the following lemma.

LEMMA 3.1. *At $\nu = \epsilon = 0$, we have $\hat{\mathbf{f}} = \mathbf{0}$, $\boldsymbol{\phi}^T\hat{\mathbf{J}}^T = \mathbf{0}^T$, and $\frac{\partial \hat{\mathbf{f}}}{\partial \nu} = \mathbf{0}$.*

Proof. The fact that $\hat{\mathbf{f}} = \mathbf{0}$ results from the fact that $\mathbf{v}(0, 0) = \mathbf{0}$, and $\mathbf{f}(\mathbf{z}_0, 0) = \mathbf{0}$. The fact that $\boldsymbol{\phi}^T\hat{\mathbf{J}}^T = \mathbf{0}^T$ follows from the fact that $\mathbf{J}\boldsymbol{\phi} = \mathbf{0}$ for $(\mathbf{z}, \epsilon) = (\mathbf{z}_0, 0)$. The last equality in the theorem arises from differentiating $\hat{\mathbf{f}}$ with respect to ν. If we do this, we get

$$\frac{\partial \hat{\mathbf{f}}}{\partial \nu} = \hat{\mathbf{J}}\left(\frac{\partial \mathbf{v}}{\partial \nu} + \boldsymbol{\phi}\right). \tag{3.11}$$

This equation is a direct consequence of the chain rule, and the fact that $\mathbf{J}$ is the gradient of $\mathbf{f}$. However, (2.14) shows that we have $\frac{\partial \mathbf{v}}{\partial \nu} = \mathbf{0}$ at $\nu = \epsilon = 0$. Using the fact that $\hat{\mathbf{J}}\boldsymbol{\phi} = \mathbf{0}$ at $\nu = \epsilon = 0$, we conclude that $\frac{\partial \hat{\mathbf{f}}}{\partial \nu}(0, 0) = \mathbf{0}$. ∎

The basic structure of the solutions near $\epsilon = 0$ results from the following lemma.

LEMMA 3.2. *At $\epsilon = \nu = 0$ we have $\frac{\partial g}{\partial \nu} = \frac{\partial g}{\partial \epsilon} = \frac{\partial^2 g}{\partial \nu^2} = 0$.*

Proof. To show that we are missing the linear terms in ν, we differentiate (3.7) with respect to ν to get

$$\frac{\partial g}{\partial \nu} = \boldsymbol{\phi}^T\hat{\mathbf{J}}^T\frac{d\hat{\mathbf{f}}}{d\nu} + \boldsymbol{\phi}^T\frac{d\hat{\mathbf{J}}^T}{d\nu}\hat{\mathbf{f}}. \tag{3.12}$$

When we evaluate this at $\nu = \epsilon = 0$, this clearly vanishes, since $\boldsymbol{\phi}^T \hat{\mathbf{J}}^T(0) = \mathbf{0}^T$, and $\hat{\mathbf{f}}(0,0) = \mathbf{0}$. Identical arguments hold for the vanishing of $\frac{\partial}{\partial \epsilon} g(0,0)$. To show that $\frac{\partial^2 g}{\partial \nu^2} = 0$ at $\nu = \epsilon = 0$, we differentiate (3.12) with respect to ν to get

$$\frac{\partial^2 g}{\partial \nu^2} = \boldsymbol{\phi}^T \left(\hat{\mathbf{J}}^T \frac{\partial^2 \hat{\mathbf{f}}}{\partial \nu^2} + 2 \frac{\partial \hat{\mathbf{J}}}{\partial \nu}^T \frac{\partial \hat{\mathbf{f}}}{\partial \nu} + \frac{\partial^2 \hat{\mathbf{J}}}{\partial \nu^2}^T \hat{\mathbf{f}} \right). \tag{3.13}$$

The first term in this sum vanishes at $\nu = \epsilon = 0$ since $\boldsymbol{\phi}^T \hat{\mathbf{J}} = \mathbf{0}^T$. The second term vanishes since $\frac{\partial \hat{\mathbf{f}}}{\partial \nu} = \mathbf{0}$ (see Lemma 3.1), and the third term vanishes since $\hat{\mathbf{f}} = \mathbf{0}$. □

Lemma 3.2 shows that near $\nu = \epsilon = 0$, the equation $g(\nu, \epsilon) = 0$ is given by

$$g(\nu, \epsilon) = c_0 \epsilon^2 + c_1 \epsilon \nu + c_3 \nu^3 + \cdots \tag{3.14}$$

We now determine these constants. If we differentiate (3.7) twice with respect to ϵ, we get

$$\frac{\partial^2 g}{\partial \epsilon^2} = \boldsymbol{\phi}^T \left(\hat{\mathbf{J}}^T \frac{\partial^2 \hat{\mathbf{f}}}{\partial \epsilon^2} + 2 \frac{\partial \hat{\mathbf{J}}}{\partial \epsilon}^T \frac{\partial \hat{\mathbf{f}}}{\partial \epsilon} + \frac{\partial^2 \hat{\mathbf{J}}}{\partial \epsilon^2} \hat{\mathbf{f}} \right). \tag{3.15}$$

If we evaluate this at $\epsilon = \nu = 0$, and use Lemma 3.1 we get

$$c_0 = \boldsymbol{\phi}^T \frac{\partial \hat{\mathbf{J}}^T}{\partial \epsilon} \frac{\partial \hat{\mathbf{f}}}{\partial \epsilon}, \tag{3.16}$$

where the derivatives are all evaluated at $\epsilon = \nu = 0$. A similar calculation shows that

$$c_1 = \boldsymbol{\phi}^T \frac{\partial \hat{\mathbf{J}}}{\partial \nu}^T \frac{\partial \hat{\mathbf{f}}}{\partial \epsilon} \tag{3.17}$$

and

$$c_3 = \frac{1}{2} \boldsymbol{\phi}^T \frac{\partial \hat{\mathbf{J}}}{\partial \nu}^T \frac{\partial^2 \hat{\mathbf{f}}}{\partial \nu^2}. \tag{3.18}$$

If we ignore the higher order terms in (3.14), and set $g(\nu, \epsilon) = 0$, we get a cubic equation for ν.

$$c_0 \epsilon^2 + c_1 \epsilon \nu + c_3 \nu^3 = 0. \tag{3.19}$$

THEOREM 3.3. *Assuming that none of c_0, c_1, and c_3 vanish, as $\epsilon \to 0$ the three roots of* (3.19) *are given by*

$$\nu_M = -\frac{c_0}{c_1} \epsilon + \cdots \tag{3.20}$$

$$\nu_\pm = \pm \sqrt{\frac{-\epsilon c_1}{c_3}} + \cdots \tag{3.21}$$

Proof. If $\nu(\epsilon)$ is a root of (3.19), then when we substitute it into that equation, we must have two of the terms be of the same order as $\epsilon \to 0$, and these terms must be

larger than the remaining term. This leaves us with three possibilities for the leading order behavior, either $c_0\epsilon^2 + c_1\epsilon\nu = 0$, $c_1\epsilon\nu + c_3\nu^3 = 0$, or $c_0\epsilon^2 + c_3\nu^3 = 0$. It is easily verified that the first two of these limits results in solutions where the ignored term is in fact smaller than the terms we kept. For example, if we assume that $c_0\epsilon^2 + c_1\epsilon\nu = 0$, then we will have $\nu = O(\epsilon)$, and the terms we kept are both order ϵ, but the term we did not keep is of order ϵ^3, which is in fact higher order than the terms we kept. Similar arguments hold for the equation $c_1\epsilon\nu + c_3\nu^3 = 0$. However, if we assume that $c_0\epsilon^2 + c_3\nu^3 = 0$, then $\nu = O(\epsilon^{2/3})$, but in such a solution the terms we kept are order ϵ^2, but the term we ignored is on the order of $\epsilon^{5/3}$, so this is not a valid solution. ∎

DEFINITION 3.4. *We will refer to* $\mathbf{z}_M(\epsilon)$ *as the solution associated with the root* ν_M *in Theorem* 3.3*, and* $\mathbf{z}_\pm(\epsilon)$ *as the roots associated with* $\nu_\pm$ *in Theorem* 3.3.

4. The objective function near the bifurcation point. In order to show that the residual is larger for $\mathbf{z}_M$ than for $\mathbf{z}_\pm$, we need to expand the objective function for $\epsilon \to 0$. In order to do this we introduce the function

$$\hat{P}(\nu, \epsilon) = \frac{1}{2}\hat{\mathbf{f}}^T(\nu, \epsilon)\hat{\mathbf{f}}(\nu, \epsilon). \tag{4.1}$$

If we have a solution $\nu(\epsilon)$ of (3.7), then the residual of the solution $\mathbf{z}(\epsilon)$ associated with this function will be $\hat{P}(\nu(\epsilon), \epsilon)$. Using the fact that $\hat{\mathbf{f}} = \frac{\partial \hat{\mathbf{f}}}{\partial \nu} = \mathbf{0}$, repeated differentiation shows that

$$\frac{\partial \hat{P}}{\partial \nu} = \frac{\partial \hat{P}}{\partial \epsilon} = \frac{\partial^2 \hat{P}}{\partial \nu^2} = \frac{\partial^2 \hat{P}}{\partial \nu \partial \epsilon} = \frac{\partial^3 \hat{P}}{\partial \nu^3} = 0 \quad \text{for } \nu = \epsilon = 0. \tag{4.2}$$

It follows that to leading order we have

$$\hat{P}(\nu, \epsilon) = a_0\epsilon^2 + a_1\epsilon\nu^2 + a_3\nu^4 + \cdots. \tag{4.3}$$

The constants a_0, a_1, and a_3 can be computed by computing the appropriate partial derivatives.

$$a_0 = \frac{1}{2}\frac{\partial \hat{\mathbf{f}}}{\partial \epsilon}^T \frac{\partial \hat{\mathbf{f}}}{\partial \epsilon} \quad \text{for } \nu = \epsilon = 0, \tag{4.4}$$

$$a_1 = \frac{1}{2}\frac{\partial^2 \hat{\mathbf{f}}}{\partial \nu^2}^T \frac{\partial \hat{\mathbf{f}}}{\partial \epsilon} \quad \text{for } \nu = \epsilon = 0, \tag{4.5}$$

$$a_3 = \frac{1}{8}\frac{\partial^2 \hat{\mathbf{f}}^T}{\partial \nu^2}\frac{\partial^2 \hat{\mathbf{f}}}{\partial \nu^2} \quad \text{for } \nu = \epsilon = 0. \tag{4.6}$$

In order to compare the residuals for $\mathbf{z}_M$ and $\mathbf{z}_\pm$ we need to express the coefficients c_1 and c_3 in (3.14) in terms of a_1 and a_3. The following lemma will help us do this.

LEMMA 4.1. *At* $\nu = \epsilon = 0$ *we have the identities*

$$\hat{\mathbf{J}}^T \frac{\partial \hat{\mathbf{f}}}{\partial \epsilon} = \mathbf{0} \tag{4.7}$$

$$\hat{\mathbf{J}}^T \frac{\partial^2 \hat{\mathbf{f}}}{\partial \nu^2} = \mathbf{0} \tag{4.8}$$

Proof. The first of these results is obtained from taking the derivative of (3.10) with respect to ϵ. If we do this and use the fact that $\hat{\mathbf{f}} = \mathbf{0}$ at $\nu = \epsilon = 0$, we get

$$\hat{\mathbf{J}}^T \frac{\partial \hat{\mathbf{f}}}{\partial \epsilon} = \frac{\partial \kappa}{\partial \epsilon}\boldsymbol{\phi}. \tag{4.9}$$

If we multiply both sides of this equation by $\boldsymbol{\phi}^T$, we see that the left-hand side vanishes, and hence we get $\frac{\partial \kappa}{\partial \epsilon} = 0$. This implies (4.7). We can derive (4.8) using a similar argument. In particular, by taking the second derivative of (3.10) with respect to ν, using the fact that $\hat{\mathbf{f}} = \frac{\partial \hat{\mathbf{f}}}{\partial \nu} = \mathbf{0}$ (see Lemma 3.1) for $\nu = \epsilon = 0$, and multiplying through by $\boldsymbol{\phi}^T$ to see that $\frac{\partial^2 \kappa}{\partial \nu^2} = 0$, and (4.8) follows from this. ∎

LEMMA 4.2. *We have the identity*

$$\frac{\partial^2 \hat{\mathbf{f}}}{\partial \nu^2} = \hat{\mathbf{J}} \frac{\partial^2 \mathbf{v}}{\partial \nu^2} + \frac{\partial \hat{\mathbf{J}}}{\partial \nu} \boldsymbol{\phi} \qquad \text{for } \nu = \epsilon = 0. \tag{4.10}$$

Proof. If we take the second derivative of (3.8) with respect to ν, we get

$$\frac{\partial^2 \hat{\mathbf{f}}}{\partial \nu^2} = \frac{\partial \hat{\mathbf{J}}}{\partial \nu} \left(\frac{\partial \mathbf{v}}{\partial \nu} + \boldsymbol{\phi} \right) + \hat{\mathbf{J}} \frac{\partial^2 \mathbf{v}}{\partial \nu^2}. \tag{4.11}$$

If we evaluate this at $\nu = \epsilon = 0$, and use the fact that $\frac{\partial \mathbf{v}}{\partial \nu}$ vanishes at that point (Lemma 2.2), we prove the lemma. ∎

LEMMA 4.3. *With a_1 and a_3 defined as in* (4.5) *and* (4.6)*, and c_1 and c_3 defined as in* (3.17) *and* (3.18)*, we have*

$$a_1 = \frac{1}{2} c_1, \tag{4.12}$$

$$a_3 = \frac{1}{4} c_3. \tag{4.13}$$

Proof. Using the expression for $\frac{\partial^2 \hat{\mathbf{f}}}{\partial \nu^2}$ from (4.10), and the fact that $\hat{\mathbf{J}}^T \frac{\partial^2 \hat{\mathbf{f}}}{\partial \nu^2} = \mathbf{0}$ (from (4.8)) we can write (4.6) as

$$a_3 = \frac{1}{8} \left(\hat{\mathbf{J}} \frac{\partial^2 \mathbf{v}}{\partial \nu^2} + \frac{\partial \hat{\mathbf{J}}}{\partial \nu} \boldsymbol{\phi} \right)^T \frac{\partial^2 \hat{\mathbf{f}}}{\partial \nu^2} = \frac{1}{8} \boldsymbol{\phi}^T \frac{\partial \hat{\mathbf{J}}}{\partial \nu}^T \frac{\partial^2 \hat{\mathbf{f}}}{\partial \nu^2} = \frac{1}{4} c_3. \tag{4.14}$$

Here the last equality has used (3.18). Similarly, if we use the expression for $\frac{\partial^2 \hat{\mathbf{f}}}{\partial \nu^2}$ from (4.10) and the fact that $\hat{\mathbf{J}}^T \frac{\partial \hat{\mathbf{f}}}{\partial \epsilon} = \mathbf{0}$ from (4.7), we can write (4.5) as

$$a_1 = \frac{1}{2} \left(\hat{\mathbf{J}} \frac{\partial^2 \mathbf{v}}{\partial \nu^2} + \frac{\partial \hat{\mathbf{J}}}{\partial \nu} \right)^T \frac{\partial \hat{\mathbf{f}}}{\partial \epsilon} = \frac{1}{2} \frac{\partial \hat{\mathbf{J}}^T}{\partial \nu} \frac{\partial \hat{\mathbf{f}}}{\partial \epsilon} = \frac{1}{2} c_1 \tag{4.15}$$

The last equality follows from (3.17). ∎

We can now collect our results to prove the following theorem.

THEOREM 4.4. *To leading order, we have*

$$\hat{P}(\nu_M(\epsilon), \epsilon) = a_0 \epsilon^2, \tag{4.16}$$

$$\hat{P}(\nu_\pm(\epsilon), \epsilon) = a_0 \epsilon^2 + \nu^2 \left(a_1 \epsilon + a_3 \nu^2 \right) = a_0 \epsilon^2 - \nu_\pm^4 a_3, \tag{4.17}$$

and since $a_3 > 0$,

$$\hat{P}(\nu_M(\epsilon), \epsilon) > \hat{P}(\nu_\pm(\epsilon), \epsilon) \quad \text{as } \epsilon \to 0. \tag{4.18}$$

Proof. We compute the objective function using (4.3) with $\nu(\epsilon)$ substituted for ν. Using (3.20), (4.16) follows from the fact that the terms ν_M^4 and $\epsilon\nu_M^2$ are higher order than $a_0\epsilon^2$. Equation (3.21) and the fact that $c_1 = 2a_1$, and $c_3 = 4a_3$, implies that $a_1\epsilon\nu_\pm + 2a_3\nu_\pm^3 = 0$, and hence $a_1\epsilon\nu_\pm + a_3\nu_\pm^3 = -a_3\nu_\pm^3$, which gives us the last equation on the right of (4.17). When we compare (4.16) and (4.17), and use the fact that $a_3 > 0$ (which follows from (4.6)), the inequality in the lemma follows. ∎

In this last theorem we showed that for small values of ϵ we have $P(\nu_\pm, \epsilon) \approx a_0\epsilon^2 - \nu_\pm^4 a_3$. It is not immediately clear that this is a positive quantity (which it must be). This can be shown to be positive by noting that as in the proof of Theorem 4.4 we have $\nu_\pm^2 = -\frac{\epsilon a_1}{2a_3}$, and hence $P(\nu_\pm, \epsilon) \approx \epsilon^2(a_0 - \frac{a_1^2}{4a_3}) = \epsilon^2 a_3(a_0a_3 - \frac{1}{4}a_1^2) > 0$. The last inequality follows from using the expressions for a_0, a_1, and a_3 in (4.4), (4.5), and (4.6) along with the Cauchy–Schwarz inequality.

5. A criterion for choosing the best solution. In this section we address how we can detect which of the three solutions $\mathbf{z}_\pm$ and $\mathbf{z}_M$ in the neighborhood of the singularity is actually the best solution. In our analysis, we have assumed that in the absence of noise, the data is given by $\mathbf{d}_0$, and that we have added noise $\epsilon\mathbf{n}$ to this vector. We have considered the solutions to be functions of ϵ, but we could more generally consider them to be functions of $\mathbf{d} - \mathbf{d}_0$. In this case the gradient of the solutions in the direction of $\mathbf{n}$ will just be the derivatives of the solutions with respect to ϵ. The derivatives of $\mathbf{z}_\pm$ with respect to ϵ will be on the order of $1/\sqrt{\epsilon}$, while the derivative of $\mathbf{z}_M$ will be order one.

On the other hand, if we change the noise in a direction perpendicular to $\mathbf{n}$, then we will keep ϵ fixed, but change the constants in (3.14). The gradient of the solutions in directions perpendicular to $\mathbf{n}$ will not be large for any of the three solutions. It follows that if we compute the gradient of our solutions with respect to the data, and find the direction where this gradient is largest, then the solutions $\mathbf{z}_\pm$ will give a large gradient, whereas the solution $\mathbf{z}_M$ will not.

To quantify this reasoning we can write our least squares equations as

$$\mathbf{G}(\mathbf{z}, \mathbf{d}) = \mathbf{0}. \tag{5.1}$$

When we change $\mathbf{d}$ by an amount $\delta\mathbf{d}$, the change in the solution $\mathbf{z}$ satisfies

$$\frac{\partial \mathbf{G}}{\partial \mathbf{z}}\delta\mathbf{z} + \frac{\partial \mathbf{G}}{\partial \mathbf{d}}\delta\mathbf{d} = \mathbf{0}. \tag{5.2}$$

This shows us that

$$\delta\mathbf{z} = R\delta\mathbf{d} \tag{5.3}$$

where

$$R = -\left(\frac{\partial \mathbf{G}}{\partial \mathbf{z}}\right)^{-1}\frac{\partial \mathbf{G}}{\partial \mathbf{d}}. \tag{5.4}$$

We can write

$$\frac{(\delta\mathbf{z})^T(\delta\mathbf{z})}{(\delta\mathbf{d})^T(\delta\mathbf{d})} = \frac{(\delta\mathbf{d})^T R^T R(\delta\mathbf{d})}{(\delta\mathbf{d})^T(\delta\mathbf{d})}. \tag{5.5}$$

The direction $\delta\mathbf{d}$ that causes the greatest change in the solution is given by the eigenvector $\mathbf{q}$ associated with the largest eigenvalue of $R^T R$. This suggests that we define the sensitivity as

$$\textit{sensitivity} = \text{largest eigenvalue of } R^T R \tag{5.6}$$

It follows that we can determine the sensitivity of the solutions by looking at the largest eigenvalue of $R^T R$. Using this measure, the solutions $\mathbf{z}_\pm$ will have a large sensitivity, and the solution $\mathbf{z}_M$ will have an order one sensitivity.

6. Example from TOA geolocation. We will now describe the application where the bifurcation was observed. The location $\mathbf{x}$ of a radio frequency (RF) emitter can be determined by measuring time of arrival t_k of the RF signal at a geographically dispersed constellation of receivers with positions $\mathbf{s}_k$ using the relation

$$| \mathbf{x} - \mathbf{s}_k | = \tau_k - \tau \quad \text{for } k = 1, N, \tag{6.1}$$

where τ_k is the arrival time times the speed of light; i.e., $\tau_k = ct_k$ and τ is the similarly scaled signal transmission time, which is also a solution variable. There are four unknowns in this problem: the 3×1 vector $\mathbf{x}$ and τ. We will consider the case of five receivers giving five equations, which make an overdetermined system.

Equation (6.1) is also used for the navigation problem where an RF receiver wishes to locate its self using signals transmitted from a constellation of overhead transmitters, such as GPS. Here the transmit time of any section of the received signal can be determined from the modulation impressed on the signal. The GPS satellites carry accurate clocks so that we can assume the transmit time to be known exactly. The receiver, however, measures the signal arrival time with a clock that has an unknown bias. For this application τ_k in (6.1) is the receiver's estimate of the range to satellite k using its biased clock and τ is the clock bias, which must be solved for along with the receiver's position. Again we will consider the case of five satellites giving five equations in four unknowns, $\mathbf{x}$ and τ.

It is easier to treat these equations algebraically if we square both sides to get the equations

$$| \mathbf{x} - \mathbf{s}_k |^2 = (\tau_k - \tau)^2 \quad \text{for } k = 1, N. \tag{6.2}$$

This gives us N equations for the four unknowns $\mathbf{x}$ and τ. We will write this as a four-dimensional vector of unknowns

$$\mathbf{z}^T = (\mathbf{x}^T, \tau). \tag{6.3}$$

and keeping with the notation used throughout this paper, the overdetermined system of (6.2) is written as $\mathbf{f}(\mathbf{z}) = 0$. The Jacobian $\mathbf{J}(\mathbf{z})$ is obtained by linearizing $\mathbf{f}(\mathbf{z})$ about the vector $\mathbf{z}$.

$$\mathbf{J}\delta\mathbf{z} = \delta\mathbf{f} \tag{6.4}$$

If we use the notation $\delta\mathbf{z}^T = (\delta\mathbf{x}^T, \delta\tau)$, we have

$$2\left(\mathbf{x} - \mathbf{s}_k\right)^T \delta\mathbf{x} + 2\left(\tau_k - \tau\right)\delta\tau = kth \text{ component of } \mathbf{J}\delta\mathbf{z}. \tag{6.5}$$

If the system is singular, this implies that there is a vector $\boldsymbol{\phi}$ such that $\mathbf{J}\boldsymbol{\phi} = 0$. If we write

$$\boldsymbol{\phi} = \begin{pmatrix} \mathbf{p} \\ \alpha \end{pmatrix}, \tag{6.6}$$

this implies that we have

$$\left(\mathbf{s}_k - \mathbf{x}\right)^T \mathbf{p} = (\tau_k - \tau)\alpha. \tag{6.7}$$

This can be written as

$$\mathbf{e}_k^T \mathbf{p} = \alpha \quad \text{for } k = 1, N, \tag{6.8}$$

where

$$\mathbf{e}_k = \frac{\mathbf{s}_k - \mathbf{x}}{\tau_k - \tau} = \frac{\mathbf{s}_k - \mathbf{x}}{\mid \mathbf{x} - \mathbf{s}_k \mid}. \tag{6.9}$$

Geometrically, (6.8) implies that the satellites all lie on a cone whose vertex is at the receiver $\mathbf{x}$.

In GPS the sensitivity to location errors arising from errors in the data is typically defined using the GDOP. The GDOP is defined as $GDOP = \sqrt{tr(\mathbf{H}^{-1})}$. Clearly the GDOP becomes large when the Hessian $\mathbf{H} = \mathbf{J}^T\mathbf{J}$ becomes nearly singular. Hence, in the language of GPS, we are considering situations where the satellite configurations have extremely large GDOP.

In [8] we show that the least squares solution to these equations can be obtained by solving a ninth order eigenvalue problem. We also show that there is something quite degenerate about the situation in GPS problems where all of the satellites are equidistant from the center of the earth. This is a very common situation in GPS problems, but it is not the only situation of interest. In our singular cases, it turns out that when the satellites are at equal radius, the coefficient c_3 in (3.14) comes out to be extremely close to zero. This makes such situations even more degenerate than the case we have discussed in this paper. For this reason we chose examples from situations where not all of the satellites are at the same radius. In particular, we will have four satellites at the middle earth orbit (MEO) radius, and one of them at geosynchronous earth orbit (GEO) radius. This corresponds to a wide area augmentation system (WAAS) augmented GPS contellation of satellites. We use coordinates made dimensionless by the radius of the earth. In this case, four of our satellites have a radius of 4.16, and the fifth has a radius of $2^{2/3}$ times this. In the example we use, the satellites have the following positions:

$$\begin{pmatrix} \mathbf{s}_1^T \\ \mathbf{s}_2^T \\ \mathbf{s}_3^T \\ \mathbf{s}_4^T \\ \mathbf{s}_5^T \end{pmatrix} = \begin{pmatrix} (.87299, 1.87293, 6.27191) \\ (2.13601, 0.55425, 3.52644) \\ (2.39458, 0.13330, 3.39908) \\ (2.00895, 0.68732, 3.57732) \\ (2.22186, 0.44388, 3.48882) \end{pmatrix} \tag{6.10}$$

The particular noise vector we are using is

$$\mathbf{n} = \begin{pmatrix} 0.57112950 \\ 0.97354900 \\ 0.43808954 \\ 0.97092952 \\ 0.41250846 \end{pmatrix}. \tag{6.11}$$

Table 6.1 shows the residual, location error, and sensitivity for the three different solutions, and for different amounts of noise levels. In all cases, the solution $\mathbf{z}_M$ has the largest residual, but the smallest location error and sensitivity. Figure 6.1 shows a log-log plot of the error as a function of ϵ for the solutions $\mathbf{z}_M$ and $\mathbf{z}_+$. The plot shows that the error for $\mathbf{z}_M$ is varying linearly with ϵ, and that for $\mathbf{z}_+$ is varying like $\sqrt{\epsilon}$.

TABLE 6.1

This shows the residual, resid $= |\mathbf{f}(\mathbf{z})|$, the location error, e_{loc}, and the sensitivity (as defined in (5.6) for the three different solutions at noise levels of $\epsilon = 1.e-4$, $\epsilon = 1.e-5$, and $\epsilon = 1.e-6$. Note that in all cases, the residual for $\mathbf{z}_M$ is the largest, and the location error and sensitivity are the smallest.

Solution	ϵ	resid	e_{loc}	σ_{sens}
$\mathbf{z}_+$	$1.e-4$	$1.88e-4$	$1.85e-1$	$3.67e7$
$\mathbf{z}_-$	$1.e-4$	$2.07e-4$	$1.86e-1$	$6.16e7$
$\mathbf{z}_M$	$1.e-4$	$2.32e-4$	$1.72e-2$	$1.16e6$
$\mathbf{z}_+$	$1.e-5$	$1.94e-5$	$5.88e-2$	$4.32e8$
$\mathbf{z}_-$	$1.e-5$	$2.00e-5$	$5.89e-2$	$5.09e8$
$\mathbf{z}_M$	$1.e-5$	$2.31e-5$	$1.71e-3$	$1.10e6$
$\mathbf{z}_+$	$1.e-6$	$1.97e-6$	$1.86e-2$	$4.56e9$
$\mathbf{z}_-$	$1.e-6$	$1.98e-6$	$1.86e-2$	$4.79e9$
$\mathbf{z}_M$	$1.e-6$	$2.31e-6$	$2.28e-4$	$2.14e5$

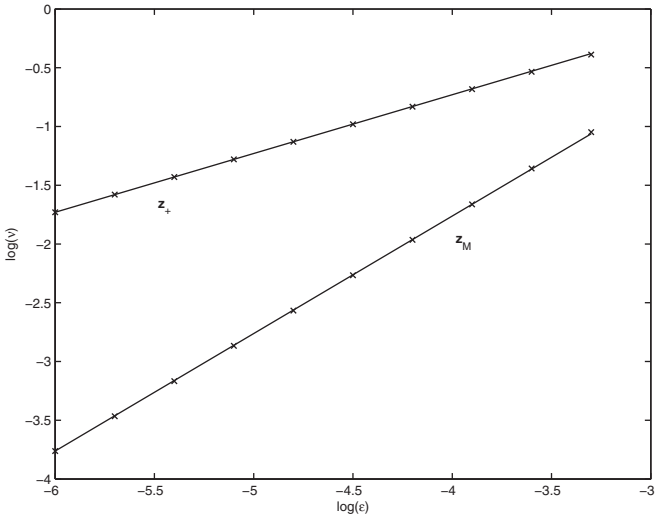


FIG. 6.1. *This gives a log-log plot of the errors of the low residual solution $\mathbf{z}_M$ and one of the larger residual solutions $\mathbf{z}_+$. The continuous lines have slopes of* 1 *and* .5*, and the numerical data are marked by crosses. The plot shows that the error for $\mathbf{z}_M$ is linear in ϵ, and the error for $\mathbf{z}_+$ varies as $\sqrt{\epsilon}$. A similar plot would show that the error in $\mathbf{z}_-$ varies like $\sqrt{\epsilon}$, but we have not shown it here since it would almost completely overlie the plot for $\mathbf{z}_+$.*

As mentioned in the introduction, if we have five satellites, the conditions in (6.8) are generic if we allow two parameters to vary. For example, if we specify the positions of the satellites, the equations (6.8) specifying that we have a degenerate configuration can be considered to be five equations for the three components of $\mathbf{p}/\alpha$. This gives five equations in three unknowns. This shows that generically we do not expect to have a degenerate configuration. However, if we allow the point $\mathbf{x}$ being located to move around on the surface of the earth, this gives us two more unknowns, and we get five equations in five unknowns. For a given configuration of satellites, we are not guaranteed that there will be a point $\mathbf{x}$ on the surface of the earth that gives a degenerate configuration, but it is generic for such a situation to occur. That is, if we arrange for such a situation, and then perturb the satellites, we can find a new point on earth that gives us a degenerate configuration.

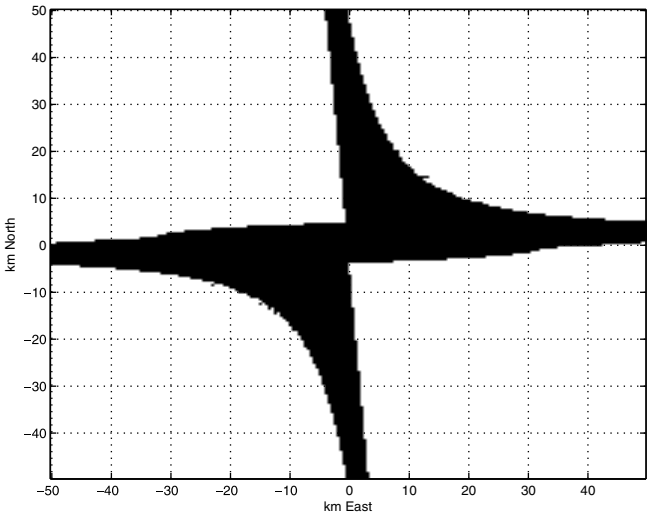


FIG. 6.2. *This shows a plot of the region surrounding a singularity where the most accurate solution does not have the smallest residual. In making this plot, the same noise was added to all of the locations. The noise had a magnitude of* 100 *nanoseconds, which is equivalent to an error of about* 30 *meters. The point where the system is exactly singular is in the middle of the figure.*

We have carried out simulations using true satellite trajectories. In these simulations we once again put four of the satellites at MEO radius, and one of the satellites at the GEO radius. However, rather than purposely putting the satellites in a degenerate configuration, we generated a million different configurations of satellites consistent with actual satellite trajectories. When the noise level was 0.001 earth radii (approximately 21 microseconds), we found that in about one out of every five thousand trials, we would get situations where the solution with the smallest residual was not the solution with the smallest error. When such a situation occurred, we found that we could adjust the position of the point being located (using Newton's method) so that we in fact had a degenerate situation where (6.8) was satisfied.

As another illustration, we once again used true satellite trajectories where one of the satellites was at GEO and the other four satellites were at MEO radius. This time we added 100 nanosecond noise to the data (equivalent to about a 30 meter error for a well-conditioned system). Out of a million trials we found four configurations where the smallest residual did not give the most accurate location. We took one of these solutions and determined the exact location of the nearby singularity. We then surrounded this point by a grid of points, and used the TOA equations with the same fixed satellite positions and fixed TOA noise vector to locate a receiver at each point on this grid. Figure 6.2 shows the region around the singularity where the most accurate solution is not the one with the smallest residual. We see that there is a central core around the singularity where this phenomenon occurs, but it spreads out in some narrow arms to quite large distances.

7. Conclusions. We have proven a general theorem concerning the solutions to nonlinear least squares problems in the neighborhood of a singularity of the equations. We have shown that in the neighborhood of a point where the Hessian has a one-dimensional null space, we will have three solutions $\mathbf{z}_\pm$ and $\mathbf{z}_M$. The general theory shows that the solution $\mathbf{z}_M$ will have the largest residual when evaluating the objective function, but it will have the smallest error when compared to the exact scripted

solution. We have shown that the solution $\mathbf{z}_M$ can be picked out of the three candidate solutions based on the fact that it is less sensitive to perturbations in the data.

REFERENCES

[1] T.M. Apostol, *Mathematical Analysis*, World Student Series Edition, 1974.

[2] M. Golubitsky, I.N. Stewart, and D. Schaeffer, *Singularities and Groups in Bifurcation Theory. Volume* 1, Springer, 1988.

[3] W.J.F. Govaerts, *Numerical Methods for Bifurcations of Dynamical Equilibria*, SIAM, Philadelphia, 2000.

[4] C.L. Lawson and R.J. Hanson, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, NJ, 1974.

[5] J. Mason and L.A. Romero, *TOA/FOA geolocation solutions using multivariate resultants*, Navigation (Washington, DC), 52 (2005), pp. 163–177.

[6] R.C. McOwen, *Partial Differential Equations: Methods and Applications*, Prentice Hall, NJ, 1996.

[7] B.W. Parkinson and J.J. Spilker, *Global Positioning System: Theory and Applications, Vol.* 1, American Institute of Aeronautics and Astronautics, 1996.

[8] L.A. Romero, J. Mason, and D.M. Day, *Analysis and extension of Bancroft's method for TOA geoloaction problems*, SIAM J. Sci. Comput., submitted.

[9] I. Stakgold, *Branching of solutions of nonlinear equations*, SIAM Rev., 13 (1971), pp. 289–332.

[10] G. Strang and K. Borre, *Linear Algebra, Geodesy, and GPS*, Wellesley Cambridge, Wellesley, MA, 1997.

SIAM J. OPTIM.
Vol. 19, No. 2, pp. 756–765

A CHARACTERIZATION OF EFFICIENT AND WEAKLY EFFICIENT POINTS IN CONVEX VECTOR OPTIMIZATION*

KRISTIN WINKLER†

Abstract. We present efficiency criteria for multicriteria problems with $\mathcal{C}(T)$-valued maps via directional derivatives and subdifferentials. It turns out that the essential structure of similar results known for $\mathbb{R}^n$-valued optimization problems remains valid. As an improvement, our formulas allow us to distinguish between efficient and weakly efficient elements.

Key words. vector-valued optimization, efficiency criteria, convex optimization

AMS subject classifications. 46N10, 90C25, 90C29, 90C46

DOI. 10.1137/060674363

1. Introduction.

We consider the vector-valued optimization problem

$$\mathfrak{f}(x) \to \min, \qquad x \in X \subset \mathfrak{X},$$

with $\mathfrak{f} : \mathfrak{X} \to \mathcal{C}(T)$. $\mathcal{C}(T)$ denotes the space of continuous real-valued functions on a compact separated Hausdorff space T. Further, let $\mathfrak{X}$ be a locally convex space, and let X be a nonempty subset of $\mathfrak{X}$. The aim of our paper is to study efficiency criteria for this problem via directional derivatives and subdifferentials known from convex analysis.

There are at least two good reasons for investigating mappings with values in $\mathcal{C}(T)$:

- Location problems frequently consist of identifying an optimal location $x \in \mathbb{R}^2$ for a new facility with respect to n reference facilities $a_1, \dots, a_n \in \mathbb{R}^2$ (for instance customers ...). Modeling them as a multicriteria problem, one gets a goal function $f : \mathbb{R}^2 \to \mathbb{R}^n$,

$$f(x) := \begin{pmatrix} d_1(x, a_1) \\ \vdots \\ d_n(x, a_n) \end{pmatrix},$$

  with distance functions $d_i : \mathbb{R}^2 \to \mathbb{R}$. If we consider a whole service area (for instance, a complete district of a town) instead of a finite number of service points a_i, this problem finds a natural extension in a multicriteria problem with goal function $g(x)(t) := d_t(x, a(t))$. Here, $t \in T$ represents a single point in the service area T. In the simpliest version, for instance, with $d_t(x, a(t)) := \|x - t\|$, g is a function with values in the space of continuous functions over the area T. So, at the end, one has to identify efficient elements in (a subset of) the space $\mathcal{C}(T)$.
- The space $\mathcal{C}(T)$ marks an important stage between the finite-dimensional Euclidean space $\mathbb{R}^n$ and general partially ordered vector spaces. $\mathcal{C}(T)$ with

*Received by the editors November 8, 2006; accepted for publication (in revised form) February 29, 2008; published electronically July 3, 2008.
http://www.siam.org/journals/siopt/19-2/67436.html
†At time of submission, Research Fellow, Institute of Mathematics, Martin Luther University of Halle-Wittenberg, Theodor-Lieser-Strasse 5, D-06099 Halle (Saale), Germany (kristin.winkler@mathematik.uni-halle. de).

its natural componentwise ordering $\mathcal{C}(T)^+$ is not order complete; the cone $\mathcal{C}(T)^+$ does not possess the Daniell property and is not nuclear. Therefore, the space $\mathcal{C}(T)$ lacks essential analytic and order theoretic properties. On the other hand, componentwise calculus—i.e., using realizations $y(t)$, $t \in T$ of functions $y \in \mathcal{C}(T)$, allows us to use definitions and ideas from $\mathbb{R}^n$.

In [22], we proved a characterization of weakly efficient points in $\mathcal{C}(T)$ based on subdifferentials for convex functions. [2] and [11] presented similar results for general distance functions on $\mathbb{R}^n$. But these papers contain only results concerning weak efficiency. Now we add formulas which distinguish efficient points from weak efficient points.

The extensions of the notion of subdifferential (and its applications) from the scalar to the vector-valued case have been investigated for more than 30 years. Early results go back, for instance, to [5, 12, 20, 21]. At a second stage, as the papers of [1, 8, 17, 18, 19] show, the properties of the partial ordering that guarantee existence of subgradients moved into the center of attention. Order completeness and/or Daniell property emerged as essential ingredients for a workable subdifferential theory—the reason why we can not use these results in $\mathcal{C}(T)$. Finally, [9, 14, 13] proposed efficiency criteria on the basis of subdifferentials for a different minimality notion.

This paper is divided as follows: In the next section we repeat some basic notions that we use in this paper, and sections 3–5 contain the main results.

2. Basic notions and preliminaries. Let $\mathfrak{X}$ be a locally convex space, denote by $\mathfrak{X}'$ its topological dual space. As usual, we denote by $\operatorname{int} M$ the topological interior of a subset $M \subseteq \mathfrak{X}$ and by $\operatorname{cl} M$ its closure. $M_1 + M_2$ denotes the pointwise sum of two subsets M_1, M_2 of a vector space, and we write shortly $M + m$ for $M + \{m\}$. Remember that the *algebraic interior* of a set $M \subset \mathfrak{X}$ is given by $\operatorname{core} M := \{x \in M : \forall\, y \in \mathfrak{X}\ \ \exists\, \lambda > 0,\ x + \lambda y \in M\}$.

Let $X \subset \mathfrak{X}$ be a nonempty convex set. The *normal cone* at X in $\bar{x}$ is given by

$$N(X; \bar{x}) = \{p \in \mathfrak{X}' :\ p(x - \bar{x}) \leq 0\ \ \forall\, x \in X\}.$$

Note that $N(X; \bar{x})$ is weak*-closed.

Let T be a compact space. The space $\mathcal{C}(T)$ of all real-valued continuous functions $y : T \to \mathbb{R}$ endowed with the maximum norm

$$\|y\| := \max_{t \in T} |y(t)|, \qquad y \in \mathcal{C}(T),$$

is a Banach space itself. Together with its order cone

$$\mathcal{C}(T)^+ := \{y \in \mathcal{C}(T) :\ y(t) \geq 0\ \ \forall\, t \in T\},$$

$\mathcal{C}(T)$ is a partially ordered vector lattice. It can be shown that

$$\operatorname{int} \mathcal{C}(T)^+ := \{y \in \mathcal{C}(T) :\ y(t) > 0\ \ \forall\, t \in T\}.$$

There exist bases for the cone $\mathcal{C}(T)^+$, but no bounded base. In general, the space $(\mathcal{C}(T), \mathcal{C}(T)^+)$ is not order-complete, does not satisfy the Daniell property, and is not nuclear (for further details on these properties compare, for instance, the monographs of [4] and [10]).

Let $X \subset \mathfrak{X}$ be a nonempty subset, and let $\mathfrak{f} : \mathfrak{X} \to \mathcal{C}(T)$. An element $\bar{x} \in X$ is called *efficient in X with respect to* $\mathfrak{f}$ if

$$\mathfrak{f}(X) \cap (\mathfrak{f}(\bar{x}) - \mathcal{C}(T)^+ \backslash \{0\}) = \emptyset.$$

$\bar{x} \in X$ is called *weakly efficient in X with respect to* $\mathfrak{f}$ if

$$\mathfrak{f}(X) \cap (\mathfrak{f}(\bar{x}) - \operatorname{int} \mathcal{C}(T)^+) = \emptyset.$$

The set of all (weakly) efficient elements in X with respect to $\mathfrak{f}$ is denoted by $\operatorname{Eff}(\mathfrak{f}(X), \mathcal{C}(T)^+)$, and $\operatorname{Eff}_w(\mathfrak{f}(X), \mathcal{C}(T)^+)$, respectively. Obviously $\operatorname{Eff}(\mathfrak{f}(X), \mathcal{C}(T)^+) \subseteq \operatorname{Eff}_w(\mathfrak{f}(X), \mathcal{C}(T)^+)$.

A function $\mathfrak{f} : \mathfrak{X} \to \mathcal{C}(T)$ is said to be *convex* iff

$$\forall\ x, y \in \mathfrak{X},\ \forall\ \lambda \in [0,1], \quad \lambda \mathfrak{f}(x) + (1-\lambda)\mathfrak{f}(y) \in \mathfrak{f}(\lambda x + (1-\lambda)y) + \mathcal{C}(T)^+,$$

i.e., if the functions $x \mapsto \mathfrak{f}(x)(t)$ are convex for all $t \in T$. Note that the classical definitions of generalized derivatives for real-valued convex functions can be applied to such functions $x \mapsto \mathfrak{f}(x)(t)$: For $t \in T$ we denote by

$$\mathfrak{f}^\circ(\bar{x}; d)(t) := \lim_{\tau \downarrow 0} \frac{\mathfrak{f}(\bar{x} + \tau d)(t) - \mathfrak{f}(\bar{x})(t)}{\tau} \tag{2.1}$$

the *directional derivative* of the function $x \mapsto \mathfrak{f}(x)(t)$ in $\bar{x}$ and direction d. Further, we call

$$\partial \mathfrak{f}(\bar{x})(t) := \{p \in \mathfrak{X}' :\ p(x - \bar{x}) \le \mathfrak{f}(x)(t) - \mathfrak{f}(\bar{x})(t)\ \ \forall\ x \in X\} \tag{2.2}$$

the *subdifferential* (in the sense of convex analysis) of $x \mapsto \mathfrak{f}(x)(t)$ in $\bar{x}$. If $\mathfrak{X}$ is a Banach space and if the function $x \mapsto \mathfrak{f}(x)(t)$ is convex and Lipschitz near $\bar{x} \in \mathfrak{X}$, then we have

$$\mathfrak{f}^\circ(\bar{x}; d)(t) := \max\{p(d) :\ p \in \partial \mathfrak{f}(\bar{x})(t)\}, \qquad t \in T. \tag{2.3}$$

Note that $\mathfrak{f}^\circ(x; d)(.)$ may be not continuous. In this point, our concept of generalized derivatives for $\mathcal{C}(T)$-valued functions differs completely from those concepts for vector-valued functions known in literature.

3. Efficiency conditions via directional derivatives. We start with a necessary and sufficient condition for weakly efficient elements.

THEOREM 3.1. *Let $\mathfrak{X}$ be a locally convex space, $X \subset \mathfrak{X}$ a convex subset, and T a compact space. Let $\mathfrak{f} : \mathfrak{X} \to \mathcal{C}(T)$ be a convex map. Then we have for $\bar{x} \in X$*

$$\sup_{t \in T} \mathfrak{f}^\circ(\bar{x}; x - \bar{x})(t) \ge 0\ \ \forall\ x \in X \quad \Longleftrightarrow \quad \bar{x} \in \operatorname{Eff}_w(\mathfrak{f}(X), \mathcal{C}(T)^+). \tag{3.1}$$

Proof. From the definition of weak efficiency, the continuity of the functions $\mathfrak{f}(x)(.)$, and the compactness of T we deduce that

$$\begin{aligned} &\bar{x} \notin \operatorname{Eff}_w(\mathfrak{f}(X), \mathcal{C}(T)^+) \\ &\Longleftrightarrow \quad \exists\ x \in X,\ \exists\ M < 0 :\ \ \mathfrak{f}(x)(t) - \mathfrak{f}(\bar{x})(t) \le M\ \ \forall\ t \in T. \end{aligned} \tag{3.2}$$

With $\tau = 1$, assertion (3.2) can be written as

$$\begin{aligned} &\bar{x} \notin \operatorname{Eff}_w(\mathfrak{f}(X), \mathcal{C}(T)^+) \\ &\Longleftrightarrow \quad \exists\ x \in X,\ \exists\ M < 0 :\ \ \frac{\mathfrak{f}(\bar{x} + \tau(x - \bar{x}))(t) - \mathfrak{f}(\bar{x})(t)}{\tau} \le M\ \ \forall\ t \in T. \end{aligned}$$

The convexity of X yields $\bar{x} + \tau(x - \bar{x}) \in X$ for $0 < \tau \leq 1$. By convexity of $\mathfrak{f}(.)(t)$, the difference quotients are monotone-decreasing as a function of τ for $\tau \downarrow 0$, hence

$$\begin{aligned} & \bar{x} \notin \operatorname{Eff}_w(\mathfrak{f}(X), \mathcal{C}(T)^+) \\ \Longrightarrow \quad & \exists\, x \in X,\ \exists\, M < 0 : \quad \mathfrak{f}^\circ(\bar{x}, x - \bar{x})(t) \leq M \quad \forall\, t \in T \\ \Longrightarrow \quad & \exists\, x \in X : \quad \sup_{t \in T} \mathfrak{f}^\circ(\bar{x}; x - \bar{x})(t) < 0. \end{aligned}$$

The contraposition of this assertion yields

$$\sup_{t \in T} \mathfrak{f}^\circ(\bar{x}; x - \bar{x})(t) \geq 0 \quad \forall\, x \in X \quad \Longrightarrow \quad \bar{x} \in \operatorname{Eff}_w(\mathfrak{f}(X), \mathcal{C}(T)^+),$$

which completes the first part of the proof.

To verify sufficiency, assume $\sup_{t \in T} \mathfrak{f}^\circ(\bar{x}; x - \bar{x}) < 0$ for some $x \in X$, i.e., $\mathfrak{f}^\circ(\bar{x}; x - \bar{x})(t) < 0$ for all $t \in T$. Therefore, for each $t \in T$, there exists some $\varepsilon(t) > 0$ such that

$$\mathfrak{f}(\bar{x} + \tau(x - \bar{x}))(t) < \mathfrak{f}(\bar{x})(t) \qquad \forall\, \tau,\ 0 < \tau < \varepsilon(t),\ \forall\, t \in T.$$

For each $\tau > 0$ with $\bar{x} + \tau(x - \bar{x}) \in X$, we consider the set

$$U_\tau := \{t \in T :\ \mathfrak{f}(\bar{x} + \tau(x - \bar{x}))(t) < \mathfrak{f}(\bar{x})(t)\}.$$

Due to the continuity of the functions, the sets U_τ are open in the relative topology. Further, by the convexity of the functions $\mathfrak{f}(.)(t)$, $t \in U_\tau$ implies that $t \in U_\alpha$ for all $0 < \alpha < \tau$. Therefore $\bigcup_{\{\tau > 0:\ \bar{x} + \tau(x - \bar{x}) \in X\}} U_\tau$ is an open cover of T. Since T is compact, there exist a finite number of $\tau_1, \ldots, \tau_n > 0$, $\bar{x} + \tau_i(x - \bar{x}) \in X$ such that $T = \bigcup_{i=1}^n U_{\tau_i}$. Let $\bar{\tau} := \min\{\tau_1, \ldots, \tau_n\}$. Then we have $\bar{x} + \bar{\tau}(x - \bar{x}) \in X$ and $\mathfrak{f}(\bar{x} + \bar{\tau}(x - \bar{x}))(t) < \mathfrak{f}(\bar{x})(t)$ for all $t \in T$. This yields $\bar{x} \notin \operatorname{Eff}_w(\mathfrak{f}(X), \mathcal{C}(T)^+)$. ∎

The idea of the last part is taken from [11].

Since $\operatorname{Eff}(\mathfrak{f}(X), \mathcal{C}(T)^+) \subseteq \operatorname{Eff}_w(\mathfrak{f}(X), \mathcal{C}(T)^+)$, equivalence (3.1) can be interpreted as a necessary condition for efficient points. In the next theorem, we add a sufficient condition.

THEOREM 3.2. *Under the assumptions of Theorem* 3.1*,*

$$\sup_{t \in T} \mathfrak{f}^\circ(\bar{x}; x - \bar{x})(t) > 0 \quad \forall\, x \in X,\ \mathfrak{f}(x) \neq \mathfrak{f}(\bar{x})$$

is sufficient for $\bar{x} \in \operatorname{Eff}(\mathfrak{f}(X), \mathcal{C}(T)^+)$.

Proof. Using the definition of efficiency, we deduce that

$$\begin{aligned} & \bar{x} \notin \operatorname{Eff}(\mathfrak{f}(X), \mathcal{C}(T)^+) \\ \Longleftrightarrow \quad & \exists\, x \in X,\ \mathfrak{f}(x) \neq \mathfrak{f}(\bar{x}) : \quad \mathfrak{f}(x)(t) - \mathfrak{f}(\bar{x})(t) \leq 0 \quad \forall\, t \in T. \end{aligned} \tag{3.3}$$

We continue as in the first part of the proof of Theorem 3.1 and get

$$\begin{aligned} & \bar{x} \notin \operatorname{Eff}(\mathfrak{f}(X), \mathcal{C}(T)^+) \\ \Longrightarrow \quad & \exists\, x \in X,\ \mathfrak{f}(x) \neq \mathfrak{f}(\bar{x}) : \quad \mathfrak{f}^\circ(\bar{x}, x - \bar{x})(t) \leq 0 \quad \forall\, t \in T \\ \Longrightarrow \quad & \exists\, x \in X,\ \mathfrak{f}(x) \neq \mathfrak{f}(\bar{x}) : \quad \sup_{t \in T} \mathfrak{f}^\circ(\bar{x}; x - \bar{x})(t) \leq 0. \end{aligned}$$

So, if $\bar{x} \in X$ and if for all $x \in X$ either $\mathfrak{f}(x) = \mathfrak{f}(\bar{x})$ or $\sup_{t \in T} \mathfrak{f}^\circ(\bar{x}; x - \bar{x}) > 0$ holds, then $\bar{x} \in \operatorname{Eff}(\mathfrak{f}(X), \mathcal{C}(T)^+)$. ∎

In comparison to Theorem 3.2, the condition for weak efficiency is necessary as well as sufficient. This agrees with corresponding results in real-valued optimization. The strict inequality in Theorem 3.2 is not necessary for efficiency: Consider the example $\mathfrak{f}(x)(t) = x^2$, $t \in T$, at $\bar{x} = 0$.

Example 3.1. We consider the function $\mathfrak{f} : [-1,1] \to \mathcal{C}([-1,1])$ defined by

$$\mathfrak{f}(x)(t) := \max\{x, t\}, \quad t \in T = [-1,1].$$

$\mathfrak{f}$ is $\mathcal{C}(T)^+$-convex, $\mathrm{Eff}\,(\mathfrak{f}([-1,1]), \mathcal{C}(T)^+) = \{-1\}$ and $\mathrm{Eff}_w(\mathfrak{f}([-1,1]), \mathcal{C}(T)^+) = [-1,1]$. The directional derivative in direction $x - \bar{x}$, $x \in [-1,1]$, $x \neq \bar{x}$, can be calculated as

$$\mathfrak{f}^\circ(\bar{x}; x - \bar{x})(t) = \begin{cases} x - \bar{x} & \text{for} \quad -1 \le t < \bar{x} \ \text{ or } \ t = \bar{x} < x, \\ 0 & \text{for} \quad \bar{x} < t \le 1 \ \text{ or } \ t = \bar{x} > x, \end{cases}$$

$$\sup_{t \in T} \mathfrak{f}^\circ(\bar{x}; x - \bar{x})(t) = \begin{cases} 0 & \text{for} \quad -1 \le x < \bar{x}, \\ x - \bar{x} & \text{for} \quad \bar{x} < x \le 1. \end{cases}$$

We deduce that $\sup_{t \in T} \mathfrak{f}^\circ(\bar{x}; x - \bar{x})(t) \ge 0$ for all $\bar{x}, x \in [-1,1]$, but the sufficient condition for efficiency is fulfilled only for $\bar{x} = -1$.

4. Efficiency conditions via subdifferentials—absence of constraints. In a similar way as in the theorems above we can prove assertions based on convex subdifferentials. In the following, $\sigma(\mathfrak{X}', \mathfrak{X})$ represents the weak* topology on $\mathfrak{X}'$.

THEOREM 4.1. *Let $\mathfrak{X}$ be a real Banach space, $\bar{x} \in \mathfrak{X}$, and T a compact space. Let $\mathfrak{f} : \mathfrak{X} \to \mathcal{C}(T)$ be a convex mapping such that the functions $x \mapsto \mathfrak{f}(x)(t)$ are Lipschitz near $\bar{x}$. Then we have*

$$0 \in \mathrm{cl}_{\sigma(\mathfrak{X}', \mathfrak{X})} \mathrm{conv} \bigcup_{t \in T} \partial \mathfrak{f}(\bar{x})(t) \iff \bar{x} \in \mathrm{Eff}_w(\mathfrak{f}(\mathfrak{X}), \mathcal{C}(T)^+),$$

$$0 \in \mathrm{core}\, \mathrm{cl}_{\sigma(\mathfrak{X}', \mathfrak{X})} \mathrm{conv} \bigcup_{t \in T} \partial \mathfrak{f}(\bar{x})(t) \implies \bar{x} \in \mathrm{Eff}\,(\mathfrak{f}(\mathfrak{X}), \mathcal{C}(T)^+).$$

Proof. We claim that

$$\begin{aligned} &\bar{x} \notin \mathrm{Eff}_w(\mathfrak{f}(\mathfrak{X}), \mathcal{C}(T)^+) \\ &\iff \quad \exists\, x \in \mathfrak{X},\ \exists\, M < 0,\ \forall\, p \in \bigcup_{t \in T} \partial \mathfrak{f}(\bar{x})(t) : \quad p(x - \bar{x}) \le M. \end{aligned} \tag{4.1}$$

Indeed, the definition of the subdifferential (2.2), together with the characterization (3.2), yields necessity; sufficiency can be deduced from (2.3) and (3.1). Due to the linearity and the continuity of the functionals $p \mapsto p(x - \bar{x})$, assertion (4.1) remains valid even if we consider all p in the $\sigma(\mathfrak{X}', \mathfrak{X})$-closure of the convex hull of $\bigcup_{t \in T} \partial \mathfrak{f}(\bar{x})(t)$. Hence

$$\begin{aligned} &\bar{x} \notin \mathrm{Eff}_w(\mathfrak{f}(\mathfrak{X}), \mathcal{C}(T)^+) \\ &\iff \quad \exists\, x \in \mathfrak{X},\ \exists\, M < 0,\ \forall\, p \in \mathrm{cl}_{\sigma(\mathfrak{X}', \mathfrak{X})} \mathrm{conv} \bigcup_{t \in T} \partial \mathfrak{f}(\bar{x})(t) : \quad p(x - \bar{x}) \le M. \end{aligned}$$

Now we apply a separation theorem (e.g., [6], Theorem 3.18) for a point and a closed convex set in $\mathfrak{X}'$, equipped with the weak* topology $\sigma(\mathfrak{X}', \mathfrak{X})$, i.e., with dual space $\mathfrak{X}$, and get

$$\bar{x} \notin \mathrm{Eff}_w(\mathfrak{f}(\mathfrak{X}), \mathcal{C}(T)^+) \iff 0 \notin \mathrm{cl}_{\sigma(\mathfrak{X}', \mathfrak{X})} \mathrm{conv} \bigcup_{t \in T} \partial \mathfrak{f}(\bar{x})(t).$$

Contraposition yields the wanted assertion for weakly efficient elements.

On the other hand, inserted in (3.3), the definition of the subdifferential (2.2) yields

$$\begin{aligned} &\bar{x} \notin \operatorname{Eff}(\mathfrak{f}(\mathfrak{X}), \mathcal{C}(T)^+) \\ \Longrightarrow \quad &\exists\, x \in \mathfrak{X}\backslash\{\bar{x}\}: \quad p(x-\bar{x}) \le 0 \ \ \forall\, p \in \bigcup_{t\in T} \partial\mathfrak{f}(\bar{x})(t). \end{aligned} \tag{4.2}$$

Again, due to the linearity and the continuity of the functionals $p \mapsto p(x-\bar{x})$, assertion (4.2) remains valid even if we consider all p in the weak*-closed convex hull of $\bigcup_{t\in T} \partial\mathfrak{f}(\bar{x})(t)$. A restriction to the algebraic interior of $\operatorname{cl}_{\sigma(\mathfrak{X}',\mathfrak{X})}\operatorname{conv} \bigcup_{t\in T} \partial\mathfrak{f}(\bar{x})(t)$, if nonempty, allows us to strengthen the inequality to a strict inequality:

$$\begin{aligned} &\bar{x} \notin \operatorname{Eff}(\mathfrak{f}(\mathfrak{X}), \mathcal{C}(T)^+) \\ \Longrightarrow \quad &\exists\, x \in \mathfrak{X}\backslash\{\bar{x}\}: \ p(x-\bar{x}) < 0 \ \ \forall\, p \in \operatorname{core}\operatorname{cl}_{\sigma(\mathfrak{X}',\mathfrak{X})}\operatorname{conv} \bigcup_{t\in T} \partial\mathfrak{f}(\bar{x})(t). \end{aligned}$$

Therefore 0 cannot be an element in the convex hull of the subdifferentials. Finally, contraposition yields that $0 \in \operatorname{core}\operatorname{cl}_{\sigma(\mathfrak{X}',\mathfrak{X})}\operatorname{conv} \bigcup_{t\in T} \partial\mathfrak{f}(\bar{x})(t)$ is sufficient for $\bar{x} \in \operatorname{Eff}(\mathfrak{f}(\mathfrak{X}), \mathcal{C}(T)^+)$. □

Remark 4.1. By a result of [15], a convex function $f : \mathfrak{X} \to \mathbb{R}$ on a Banach space is Lipschitz near any point x of an open convex set $X \subset \mathfrak{X}$ if it is bounded above on a neighborhood of some point of X. So, boundedness can replace the Lipschitz assumption.

5. Efficiency conditions via subdifferentials—presence of constraints. Including constraints into the problem, we get the following characterization of (weak) efficient elements.

THEOREM 5.1. *Let $\mathfrak{X}$ be a Banach space, $X \subset \mathfrak{X}$ a closed convex set, $\bar{x} \in X$ and T a compact space. Assume $\mathfrak{f} : \mathfrak{X} \to \mathcal{C}(T)$ to be a convex mapping such that the functions $x \mapsto \mathfrak{f}(x)(t)$ are Lipschitz near $\bar{x}$. Then we have*

$$0 \in \operatorname{cl}_{\sigma(\mathfrak{X}',\mathfrak{X})}\operatorname{conv}\left(\bigcup_{t\in T} \partial\mathfrak{f}(\bar{x})(t)\right) + N(X;\bar{x}) \Longleftrightarrow \bar{x} \in \operatorname{Eff}_w(\mathfrak{f}(X), \mathcal{C}(T)^+),$$

$$0 \in \operatorname{core}\left(\operatorname{cl}_{\sigma(\mathfrak{X}',\mathfrak{X})}\operatorname{conv}\left(\bigcup_{t\in T} \partial\mathfrak{f}(\bar{x})(t)\right) + N(X;\bar{x})\right) \Longrightarrow \bar{x} \in \operatorname{Eff}(\mathfrak{f}(X), \mathcal{C}(T)^+).$$

We need two technical results.

LEMMA 5.2. *Under the assumptions of Theorem* 5.1*, the set $\bigcup_{t\in T} \partial\mathfrak{f}(\bar{x})(t)$ is norm bounded in $\mathfrak{X}'$.*

Proof. For $h \in \mathfrak{X}$ we estimate that

$$\begin{aligned} \sup\{p(h): \ p \in \bigcup_{t\in T} \partial\mathfrak{f}(\bar{x})(t)\} &= \sup_{t\in T} \ \sup\{p(h): \ p \in \partial\mathfrak{f}(\bar{x})(t)\} \\ &\le \sup_{t\in T} \ (\mathfrak{f}(\bar{x}+h)(t) - \mathfrak{f}(\bar{x})(t)) \\ &= \max_{t\in T} \ (\mathfrak{f}(\bar{x}+h)(t) - \mathfrak{f}(\bar{x})(t)) \\ &< +\infty. \end{aligned}$$

Hence the set $\bigcup_{t\in T} \partial\mathfrak{f}(\bar{x})(t)$ is $\sigma(\mathfrak{X}',\mathfrak{X})$-bounded and finally (since $\mathfrak{X}'$ is a Banach space) norm bounded in $\mathfrak{X}'$. □

The idea of this proof is from a paper of [11].

LEMMA 5.3. *Let M_1 and M_2 be two subsets of the dual space $\mathfrak{X}'$ of a Banach space $\mathfrak{X}$, and let M_1 be norm bounded. Then we have*

$$\mathrm{cl}_{\sigma(\mathfrak{X}',\mathfrak{X})}(M_1 + M_2) = \mathrm{cl}_{\sigma(\mathfrak{X}',\mathfrak{X})}M_1 + \mathrm{cl}_{\sigma(\mathfrak{X}',\mathfrak{X})}M_2.$$

Proof. Norm boundedness of M_1 implies the $\sigma(\mathfrak{X}',\mathfrak{X})$-compactness of $\mathrm{cl}_{\sigma(\mathfrak{X}',\mathfrak{X})}M_1$. Further it is known that the sum of a compact set and an arbitrary closed set is closed (in the same topology as the compactness is valid in). Therefore, the assertion of the lemma above is a consequence of the formula

$$\mathrm{cl}_{\sigma(\mathfrak{X}',\mathfrak{X})}(M_1 + M_2) = \mathrm{cl}_{\sigma(\mathfrak{X}',\mathfrak{X})}\left(\mathrm{cl}_{\sigma(\mathfrak{X}',\mathfrak{X})}M_1 + \mathrm{cl}_{\sigma(\mathfrak{X}',\mathfrak{X})}M_2\right),$$

which holds for every topological vector space. □

Proof. (Theorem 5.1). To show sufficiency we apply the proof of Theorem 4.1 to the extended real-valued functions

$$\hat{\mathfrak{f}}(x)(t) := \mathfrak{f}(x)(t) + \delta_X(x), \qquad t \in T$$

where $\delta_X(.)$ denotes the indicator function of the set X, $\delta_X(x) = 0$ if $x \in X$ and $\delta_X(x) = +\infty$ else. We have

$$\partial\hat{\mathfrak{f}}(\bar{x})(t) = \partial\mathfrak{f}(\bar{x})(t) + N(X;\bar{x}), \qquad t \in T$$

(compare [3], Proposition 2.3.3 together with Corollary 3 and Proposition 2.3.6). Using the same ideas as in the proof of Theorem 4.1 we can show the following:

$$\bar{x} \notin \mathrm{Eff}_w(\mathfrak{f}(X), \mathcal{C}(T)^+) \Longrightarrow 0 \notin \mathrm{cl}_{\sigma(\mathfrak{X}',\mathfrak{X})}\mathrm{conv}\left(\bigcup_{t\in T}\partial\hat{\mathfrak{f}}(\bar{x})(t)\right),$$

$$\bar{x} \notin \mathrm{Eff}\,(\mathfrak{f}(X), \mathcal{C}(T)^+) \Longrightarrow 0 \notin \mathrm{core}\,\mathrm{cl}_{\sigma(\mathfrak{X}',\mathfrak{X})}\mathrm{conv}\left(\bigcup_{t\in T}\partial\hat{\mathfrak{f}}(\bar{x})(t)\right).$$

Applying the two lemmata above we derive

$$\bar{x} \notin \mathrm{Eff}_w(\mathfrak{f}(X), \mathcal{C}(T)^+) \Longrightarrow 0 \notin \mathrm{cl}_{\sigma(\mathfrak{X}',\mathfrak{X})}\mathrm{conv}\left(\bigcup_{t\in T}\partial\mathfrak{f}(x)(t)\right) + N(X;x),$$

$$\bar{x} \notin \mathrm{Eff}\,(\mathfrak{f}(X), \mathcal{C}(T)^+) \Longrightarrow 0 \notin \mathrm{core}\left(\mathrm{cl}_{\sigma(\mathfrak{X}',\mathfrak{X})}\mathrm{conv}\left(\bigcup_{t\in T}\partial\mathfrak{f}(x)(t)\right) + N(X;x)\right).$$

The contraposition of this implications yields the claimed assertions.

To verify necessity, let $0 \notin \mathrm{cl}_{\sigma(\mathfrak{X}',\mathfrak{X})}\mathrm{conv}\left(\bigcup_{t\in T}\partial\mathfrak{f}(x)(t)\right) + N(X;x)$. By the strong separation theorem there exist $d \in \mathfrak{X}$ and $M < 0$ such that

$$(p_1 + p_2)(d) \le M < 0 \quad \forall\, p_1 \in \mathrm{cl}_{\sigma(\mathfrak{X}',\mathfrak{X})}\mathrm{conv}\bigcup_{t\in T}\partial\mathfrak{f}(x)(t),\ \forall\, p_2 \in N(X;\bar{x}).$$

Since $N(X;\bar{x})$ is a cone, we get

$$\begin{aligned} &p_2(d) \le 0 \qquad \forall\, p_2 \in N(X;\bar{x}),\\ &p_1(d) \le M < 0 \quad \forall\, p_1 \in \mathrm{cl}_{\sigma(\mathfrak{X}',\mathfrak{X})}\mathrm{conv}\bigcup_{t\in T}\partial\mathfrak{f}(x)(t). \end{aligned}$$

The convexity of X yields $d \in T(X;\bar{x}) = \operatorname{cl} \bigcup_{\tau>0} \tau(X-\bar{x})$. Due to Lemma 5.2, the set $\bigcup_{t\in T} \partial \mathfrak{f}(x)(t)$ is bounded; therefore there exist for each $\tilde{d} \in \mathfrak{X}$ some $\varepsilon > 0$ such that

$$p_1(d+\varepsilon\tilde{d}) \le M/2 < 0 \quad \forall\, p_1 \in \bigcup_{t\in T} \partial \mathfrak{f}(x)(t).$$

Further there exist $\bar{d} \in \bigcup_{\tau>0} \tau(X-\bar{x})$ such that $p_1(\bar{d}) \le M/2 < 0$ for all $p_1 \in \bigcup_{t\in T} \partial \mathfrak{f}(x)(t)$. We get $\mathfrak{f}^\circ(\bar{x};\bar{d})(t) \le M/2 < 0$ for all $t \in T$ and $\bar{x} + \tau\bar{d} \in X$ for sufficiently small $\tau > 0$. By Theorem 3.1 this is a contradiction to the weak efficiency of $\bar{x}$. The contraposition of this result yields the wanted assertion. □

Plastria and Carrizosa [11] proved for $\mathcal{C}(T)^+$-convex maps $\mathfrak{f}$ the efficiency criterium

$$\bar{x} \in \operatorname{Eff}_w(\mathfrak{f}(X), \mathcal{C}(T)^+) \iff 0 \in \operatorname{cl}\operatorname{conv}\left(\bigcup_{t\in T} \partial \mathfrak{f}(x)(t)\right) + N(X;x).$$

They choose another idea for their proof: They scalarized the vector-valued optimization problem by the functional

$$F(x) := \sup_{t\in T} [\mathfrak{f}(x)(t) - \mathfrak{f}(\bar{x})(t)] = \max_{t\in T} [\mathfrak{f}(x)(t) - \mathfrak{f}(\bar{x})(t)]$$

($\bar{x} \in X \subset \mathbb{R}^n$ fixed) and proved that

$$\partial F(x) = \operatorname{conv} \bigcup_{t\in T} \partial \mathfrak{f}(x)(t).$$

This idea also influenced our studies. In comparison to [11], we work in a Banach space $\mathfrak{X}$ instead of $\mathbb{R}^n$. Nevertheless, we could have applied the results of [16]; there we found a formula for the subdifferential of the pointwise maxima of infinitely many convex functions. But by this way it is not possible to distinguish between efficient and weakly efficient points.

Finally, let us continue Example 3.1.

Example 5.1. We consider the function $\mathfrak{f} : \mathbb{R} \to \mathcal{C}([-1,1])$ defined by

$$\mathfrak{f}(x)(t) := \max\{x,t\}, \quad t \in T = [-1,1],$$

and add now the constraint set $X = [-1,1]$. Remember $\operatorname{Eff}(\mathfrak{f}(X), \mathcal{C}(T)^+) = \{-1\}$ and $\operatorname{Eff}_w(\mathfrak{f}(X), \mathcal{C}(T)^+) = [-1,1]$. Obviously,

$$N(X;\bar{x}) = \begin{cases} \mathbb{R}^- & \text{for} \quad \bar{x} = -1, \\ \{0\} & \text{for} \quad -1 < \bar{x} < 1, \\ \mathbb{R}^+ & \text{for} \quad \bar{x} = 1. \end{cases}$$

The subdifferentials of the functions $x \mapsto \mathfrak{f}(x)(t)$ for $\bar{x} \in X$ can be calculated as

$$\partial \mathfrak{f}(\bar{x})(t) = \begin{cases} \{1\} & \text{for} \quad t < \bar{x}, \\ [0,1] & \text{for} \quad t = \bar{x}, \\ \{0\} & \text{for} \quad \bar{x} < t, \end{cases}$$

$$\operatorname{cl}\operatorname{conv} \bigcup_{t\in T} \partial \mathfrak{f}(\bar{x})(t) = [0,1].$$

So we verify $0 \in \operatorname{cl}\operatorname{conv} \bigcup_{t\in T} \partial \mathfrak{f}(\bar{x})(t) + N(X;\bar{x})$ for all $\bar{x} \in X$, but 0 is contained in the algebraic interior of $\operatorname{cl}\operatorname{conv} \bigcup_{t\in T} \partial \mathfrak{f}(\bar{x})(t) + N(X;\bar{x})$ only for $\bar{x} = -1$.

6. Conclusions. In this paper, we proved geometrical characterizations of efficient and weakly efficient points based on directional derivatives and subdifferentials in the sense of convex (real-valued) optimization. These assertions extend known results that provide only weakly efficient points.

In this paper, we focused on convex maps. It is possible to prove similar assertions for Lipschitz maps using results of [7], concerning optimality criteria for maximum functions. For more details, compare [23].

Acknowledgments. We are indebted to Prof. C. Zălinescu, University Iaşi (Romania) and the anonymous referees for helpful comments on the presented results.

REFERENCES

[1] J. M. Borwein, *Subgradients of convex operators*, Mathematische Operationsforschung und Statistik Series Optimization, 15 (1984), pp. 179–191.
[2] E. Carrizosa and F. Plastria, *A characterization of efficient points in constrained location problems with regional demand*, Oper. Res. Lett., 19 (1996), pp. 129–134.
[3] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Classics Appl. Math. 5., SIAM, Philadelphia, PA, 1990.
[4] A. Göpfert, H. Riahi, C. Tammer, and C. Zălinescu, *Variational Methods in Partially Ordered Spaces*, CMS Books Math./Ouvrages Math. SMC, 17, G. Di Pillo and A. Murli, eds., Springer-Verlag, New York, 2003.
[5] J.-B. Hiriart-Urruty and L. Thibault, *Existence et caractérisation de différentielles généralisées d'applications localement lipschitziennes d'un espace de Banach séparable dans un espace de Banach réflexif séparable*, Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences, Séries A et B, 290 (1980), pp. 1091–1094.
[6] J. Jahn, *Mathematical Vector Optimization in Partially Ordered Linear Spaces*, Methoden und Verfahren der Mathematischen Physik [Methods and Procedures in Mathematical Physics] 31, Verlag Peter D. Lang, Frankfurt am Main, 1986.
[7] D. V. Luu and W. Oettli, *Necessary optimality conditions for nonsmooth minimax problems*, Z. Anal. Anwendungen, 12 (1993), pp. 709–721.
[8] N. S. Papageorgiou, *Nonsmooth analysis on partially ordered vector spaces*. I. *Convex case*, Pacific J. Math., 107 (1983), pp. 403–458.
[9] N. S. Papageorgiou, *Nonsmooth analysis on partially ordered vector spaces: The subdifferential theory*, Nonlinear Anal., 10 (1986), pp. 615–637.
[10] A. L. Peressini, *Ordered Topological Vector Spaces*, Harper and Row, New York, 1967.
[11] F. Plastria and E. Carrizosa, *Geometrical characterization of weakly efficient points*, J. Optim. Theory Appl., 90 (1996), pp. 217–223.
[12] C. Raffin, *Sur les programmes convexes définis dans des espaces vectoriels topologiques*, Ann. Inst. Fourier (Grenoble), 20 (1970), pp. 457–491.
[13] T. W. Reiland, *Generalized differentiability for a class of nondifferentiable operators with applications to nonsmooth optimization*, Austral. Math. Soc., Ser. A, 47 (1989), pp. 114–132.
[14] T. W. Reiland, *Nonsmooth analysis and optimization on partially ordered vector spaces*, Int. J. Math. Math. Sci., 15 (1992), pp. 65–81.
[15] A. W. Roberts and D. E. Varberg, *Another proof that convex functions are locally Lipschitz*, Amer. Math. Monthly, 81 (1974), pp. 1014–1016.
[16] V. N. Solovev, *On the subdifferential and directional derivatives of the maximum of a family of convex functions*, Izv. Math., 62 (1998), pp. 807–832. Translation from Izv. Ross. Akad. Nauk. Izv., Ser. Mat., 62 (1998), pp. 173–200.
[17] T. Staib, *Notwendige Optimalitätsbedingungen in der Mehrkriteriellen Optimierung mit Anwendung auf Steuerprobleme*, Ph.D. thesis, Friedrich Alexander University of Erlangen-Nürnberg, 1989.
[18] T. Staib, *Necessary optimality conditions for nonsmooth multicriterial optimization problems*, SIAM J. Optim., 2 (1992), pp. 153–171.
[19] M. Théra, *Subdifferential calculus for convex operators*, J. Math. Anal. Appl., 80 (1981), pp. 78–91.
[20] L. Thibault, *Subdifferentials of compactly Lipschitzian vector-valued functions*, Ann. Mat. Pura Appl., 125 (1980), pp. 157–192.

[21] M. Valadier, *Sous-différentiabilité de fonctions convexes à valeurs dans un espace vectoriel ordonné*, Math. Scand., 30 (1972), pp. 65–74.
[22] K. Winkler, *Characterizations of efficient points in convex vector optimization problems*, Math. Methods Oper. Res., 53 (2001), pp. 205–214.
[23] K. Winkler, *Aspekte Mehrkriterieller Optimierung $\mathcal{C}(T)$-Wertiger Abbildungen*, Ph.D. thesis, Martin Luther University of Halle-Wittenberg, 2003.

SIAM J. OPTIM.
Vol. 19, No. 2, pp. 766–785

THE PROXIMAL AVERAGE: BASIC THEORY*

HEINZ H. BAUSCHKE†, RAFAL GOEBEL‡, YVES LUCET§, AND XIANFU WANG†

Abstract. The recently introduced proximal average of two convex functions is a convex function with many useful properties. In this paper, we introduce and systematically study the proximal average for finitely many convex functions. The basic properties of the proximal average with respect to the standard convex-analytical notions (domain, Fenchel conjugate, subdifferential, proximal mapping, epi-continuity, and others) are provided and illustrated by several examples.

Key words. arithmetic average, arithmetic mean, convex analysis, convex function, epi-convergence, epigraphical average, epi-topology, essential smoothness, essential strict convexity, Fenchel conjugate, harmonic mean, Legendre function, Moreau envelope, proximal average, proximal mapping, subdifferential operator

AMS subject classifications. Primary, 90C25; Secondary, 26A51, 26B25, 26E60, 46C05, 47H05, 52A41

DOI. 10.1137/070687542

1. Overview.

Let f_1 and f_2 be two functions that are convex, lower semicontinuous, and proper, and let λ_1 and λ_2 be strictly positive real numbers adding up to 1. How can we average the two functions f_1 and f_2 with respect to the weights λ_1 and λ_2 in a useful way? Perhaps the first approach is to consider the *arithmetic average* $\lambda_1 f_1 + \lambda_2 f_2$. However, functions in convex analysis are allowed to take on the value $+\infty$, for example, to model constraints in optimization problems. Thus, the arithmetic average can turn out to be $+\infty$ everywhere and then carries little information about f_1 and f_2; this happens whenever f_1 and f_2 are nowhere both finite. How could we possibly average such functions? A second thought may suggest to construct the *epigraphical average* $\lambda_1 \star f_1 \mathbin{\blacksquare} \lambda_2 \star f_2$ obtained by forming a convex combination of the epigraphs of f_1 and f_2. Unfortunately, if the functions f_1 and f_2 lack coercivity, then the epigraphical average fails to be helpful: For instance, if f_1 and f_2 are two distinct linear functions, then their epigraphical average is identically equal to $-\infty$, and hence of little use. The *proximal average*, first introduced in [5] in the context of fixed point theory and recently studied in [3, 4, 6, 8, 13] from various viewpoints, avoids the mentioned difficulties and possesses numerous properties that are attractive to convex analysts.

The aim of this paper is to provide the basic theory of the proximal average. In addition, we extend it to more than two functions, and we allow for an additional positive parameter. For the reader's convenience and the sake of completeness, the

*Received by the editors April 9, 2007; accepted for publication (in revised form) March 5, 2008; published electronically July 3, 2008.

http://www.siam.org/journals/siopt/19-2/68754.html

†Department of Mathematics, Irving K. Barber School, University of British Columbia Okanagan, Kelowna, BC V1V 1V7, Canada (heinz.bauschke@ubc.ca, shawn.wang@ubc.ca). The first author's research was partially supported by the Natural Sciences and Engineering Research Council of Canada and by the Canada Research Chair Program. The last author's research was partially supported by the Natural Sciences and Engineering Research Council of Canada.

‡Department of Mathematics and Statistics, Loyola University Chicago, Chicago, IL 60626 (rgoebel1@luc.edu).

§Department of Computer Science, Irving K. Barber School, University of British Columbia Okanagan, Kelowna, BC V1V 1V7, Canada (yves.lucet@ubc.ca). This author's research was partially supported by the Natural Sciences and Engineering Research Council of Canada.

presentation of the theory is largely self-contained. It is shown that the proximal average has many desirable properties in terms of its domain, Fenchel conjugate, Moreau envelope, proximal mapping, subdifferential, epi-continuity, and other convex-analytical notions. Moreover, the arithmetic and epigraphical averages turn out to be limits of the proximal average as the parameter tends to 0 and $+\infty$, respectively. Various examples illustrate our results. An interesting topic for future research is the extension to series and integrals.

The rest of this paper is organized as follows. Section 2 collects the notation used throughout this paper, and section 3 collects and presents results that simplify later proofs. The proximal average is introduced in section 4, where its domain is also characterized. In section 5, we present one very useful result (Theorem 5.1), which states that the Fenchel conjugate of the proximal average is the proximal average of the Fenchel conjugates. An important consequence of this result is that the proximal average is convex, lower semicontinuous, and proper. In section 6 we consider the Moreau envelope and proximal mapping of the proximal average. In section 7 we consider its subdifferential operator as well as essential smoothness and essential strict convexity. In section 8 it is shown that the arithmetic and epigraphical averages are pointwise limits of the proximal average. Epi-convergence properties are discussed in section 9, where the arithmetic and epigraphical averages are shown to be limiting instances of the proximal average with respect to epi-convergence.

2. Standing assumptions and notation. Throughout this paper,

(1) X is a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and corresponding norm $\| \cdot \|$.

Due to its repeated use, we abbreviate the quadratic energy function by

$$\mathfrak{q} = \tfrac{1}{2}\| \cdot \|^2. \tag{2}$$

We set

$$\Gamma(X) = \big\{ f \colon X \to \left]-\infty, +\infty\right] \mid f \text{ is convex, lower semicontinuous, and proper} \big\}. \tag{3}$$

We assume throughout that

$$n \in \{1, 2, 3, \dots\}, \tag{4}$$

that

$$f_1, \dots, f_n \text{ belong to } \Gamma(X), \tag{5}$$

that

$$\lambda_1, \dots, \lambda_n \text{ are nonnegative real numbers such that } \lambda_1 + \cdots + \lambda_n = 1, \tag{6}$$

and that

$$\mu \text{ is a strictly positive real number.} \tag{7}$$

The Fenchel conjugate of a function f is denoted by f^*. It will be convenient to set

$$\boldsymbol{f} = (f_1, \dots, f_n), \quad \boldsymbol{f}^* = (f_1^*, \dots, f_n^*), \quad \text{and} \quad \boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_n). \tag{8}$$

Other notation not explicitly defined here or later is standard in convex analysis and as in, e.g., [18, 19, 20]. Let f be a convex function, and let S be a set. Then we write $\operatorname{dom} f$, $\operatorname{epi} f$, ∂f, $\operatorname{cl} f$, $\inf f$, $\min f$, $\operatorname{argmin} f$, d_S, $\operatorname{conv} S$, $\operatorname{int} S$, ι_S, and N_S to denote the (effective) domain, epigraph, subdifferential operator, lower closure, infimum value, minimum value if the infimum value is attained, the set of minimizers, distance function, convex hull, interior, indicator function, and normal cone operator, respectively. The identity operator is represented by Id.

3. Auxiliary results. We start by reviewing the key notions of epi-multiplication and epi-addition, following the viewpoint taken in [19, section 1.H]. Let $\alpha \geq 0$, $f \in \Gamma(X)$, $g \in \Gamma(X)$, and $h \in \Gamma(X)$. Then

$$\alpha \star f = \begin{cases} \alpha f(\cdot/\alpha) & \text{if } \alpha > 0, \\ \iota_{\{0\}} & \text{if } \alpha = 0. \end{cases} \tag{9}$$

The term "epi-multiplication" stems from the fact that $\operatorname{epi}(\alpha \star f) = \alpha \operatorname{epi}(f)$ when $\alpha > 0$. Epi-addition, or infimal convolution, is defined by

$$f \operatorname{\blacktriangle} g \colon X \to [-\infty, +\infty] : x \mapsto \inf_{y+z=x} \big(f(y) + g(z)\big); \tag{10}$$

and the term "epi-addition" stems from the fact that the strict epigraph of $f \operatorname{\blacktriangle} g$ is the Minkowski sum of the strict epigraphs of f and g, i.e., $\big\{(x,r) \in X \times \mathbb{R} \mid (f \operatorname{\blacktriangle} g)(x) < r\big\} = \big\{(y,s) \in X \times \mathbb{R} \mid f(y) < s\big\} + \big\{(z,t) \in X \times \mathbb{R} \mid g(z) < t\big\}$. The epi-sum of finitely many functions is defined analogously.

To avoid excessive usage of parentheses, epi-multiplication and regular multiplication are given precedence over epi- and regular addition, i.e., $\alpha \star f + g = (\alpha \star f) + g$, $\alpha \star f \operatorname{\blacktriangle} g = (\alpha \star f) \operatorname{\blacktriangle} g$, $\alpha f + g = (\alpha f) + g$, and $\alpha f \operatorname{\blacktriangle} g = (\alpha f) \operatorname{\blacktriangle} g$. It will also be convenient to give epi-addition a higher precedence than regular addition or subtraction, i.e., $f \operatorname{\blacktriangle} g + h = (f \operatorname{\blacktriangle} g) + h$ and $f \operatorname{\blacktriangle} g - h = (f \operatorname{\blacktriangle} g) - h$.

The next three propositions are elementary. Proofs for the finite-dimensional case are in [19]; they extend without difficulty to the present Hilbert space setting.

PROPOSITION 3.1. *Let $f \in \Gamma(X)$, let $\alpha \geq 0$, and let $\beta \geq 0$. Then the following hold.*

(i) $\alpha > 0 \Rightarrow \operatorname{epi}(\alpha \star f) = \alpha(\operatorname{epi} f)$.
(ii) $\operatorname{dom}(\alpha \star f) = \alpha(\operatorname{dom} f)$.
(iii) $f \operatorname{\blacktriangle} \iota_{\{0\}} = f$.
(iv) $\operatorname{dom}(f_1 \operatorname{\blacktriangle} \cdots \operatorname{\blacktriangle} f_n) = (\operatorname{dom} f_1) + \cdots + (\operatorname{dom} f_n)$.
(v) $\alpha \star (f_1 \operatorname{\blacktriangle} \cdots \operatorname{\blacktriangle} f_n) = \alpha \star f_1 \operatorname{\blacktriangle} \cdots \operatorname{\blacktriangle} \alpha \star f_n$.
(vi) $\alpha(f_1 \operatorname{\blacktriangle} \cdots \operatorname{\blacktriangle} f_n) = \alpha f_1 \operatorname{\blacktriangle} \cdots \operatorname{\blacktriangle} \alpha f_n$.
(vii) $\alpha \star (\beta \star f) = (\alpha\beta) \star f$.
(viii) $(\alpha + \beta) \star f = \alpha \star f \operatorname{\blacktriangle} \beta \star f$.
(ix) $\alpha > 0 \Rightarrow \alpha(\beta \star (\alpha^{-1} f)) = \beta \star f$.

Proof. The conclusions all follow readily from the definitions; see also [19, Exercise 1.28(a)] for (i), [19, page 25] for (v) and (vii), and [19, Exercise 2.24(c)] for (viii). ∎

PROPOSITION 3.2. *Let $\alpha \geq 0$. Then the following hold.*

(i) $(\alpha f)^* = \alpha \star f^*$.
(ii) $(\alpha \star f)^* = \alpha f^*$.
(iii) $(f_1 \operatorname{\blacktriangle} \cdots \operatorname{\blacktriangle} f_n)^* = f_1^* + \cdots + f_n^*$.

Proof. The statements are simple consequences of the definitions; see also [19, page 475] for (i) and (ii), and [19, Theorem 11.23(a)] for (iii). ∎

PROPOSITION 3.3. *Let $f \in \Gamma(X)$, and let $\alpha \geq 0$. Then the following hold.*

(i) $\mathfrak{q}^* = \mathfrak{q}$; *in fact,* $\mathfrak{q}$ *is the only function equal to its Fenchel conjugate.*
(ii) $\alpha > 0 \Rightarrow \alpha^{-1} \mathbin{\star} \mathfrak{q} = \alpha\, \mathfrak{q}$.
(iii) $(\alpha \mathbin{\star} \mathfrak{q})^* = \alpha\, \mathfrak{q}$.
(iv) $(\alpha\, \mathfrak{q})^* = \alpha \mathbin{\star} \mathfrak{q}$.
(v) $(f \mathbin{\blacktriangleleft} \mathfrak{q}) + (f^* \mathbin{\blacktriangleleft} \mathfrak{q}) = \mathfrak{q}$.

Proof. (i): See, e.g., [19, Example 11.11]. (ii): An immediate consequence of the definition of $\mathfrak{q}$. (iii): Combine Proposition 3.2(ii) with (i). (iv): Combine Proposition 3.2(i) with (i). (v): See [16] or [19, Example 11.26]. □

The next result is deep and stated as a fact.

FACT 3.4. *The following hold.*

(i) *If* $\operatorname{int}\operatorname{dom} f_1 \cap \cdots \cap \operatorname{int}\operatorname{dom} f_{n-1} \cap \operatorname{dom} f_n \neq \varnothing$, *then* $(f_1 + \cdots + f_n)^* = f_1^* \mathbin{\blacktriangleleft} \cdots \mathbin{\blacktriangleleft} f_n^*$ *and the epi-sum is exact, i.e., the infimum in the definition of the epi-sum is attained.*
(ii) *If* $\operatorname{int}\operatorname{dom} f_1^* \cap \cdots \cap \operatorname{int}\operatorname{dom} f_{n-1}^* \cap \operatorname{dom} f_n^* \neq \varnothing$, *then* $f_1 \mathbin{\blacktriangleleft} \cdots \mathbin{\blacktriangleleft} f_n$ *is exact and* $\operatorname{epi}(f_1 \mathbin{\blacktriangleleft} \cdots \mathbin{\blacktriangleleft} f_n) = (\operatorname{epi} f_1) + \cdots + (\operatorname{epi} f_n)$.

Proof. This is a consequence of [20, Theorem 2.8.7]. □

The following result on the conjugate of the difference will be useful.

FACT 3.5. *Let $g \in \Gamma(X)$, and let $h \in \Gamma(X)$ such that both h and h^* have full domain. Then*

$$(\forall x^* \in X) \quad (g - h)^*(x^*) = \sup_{y^* \in X} \big(g^*(y^*) - h^*(y^* - x^*)\big). \tag{11}$$

Proof. This is a consequence of [9, Theorem 2.2]. □

COROLLARY 3.6. *Let $g \in \Gamma(X)$. Then*

$$(g - \mu \mathbin{\star} \mathfrak{q})^* = \mu(\mathfrak{q} - \mu^{-1} g^*)^* - \mu^{-1} \mathbin{\star} \mathfrak{q}. \tag{12}$$

Proof. Set $h = \mu \mathbin{\star} \mathfrak{q}$. Then $h^* = \mu\, \mathfrak{q}$ by Proposition 3.3(iii), and hence both h and h^* have full domain. Using Fact 3.5, we deduce that for every $x^* \in X$

$$\begin{aligned}
(g - h)^*(x^*) &= \sup_{y^* \in X} \big(g^*(y^*) - \mu\, \mathfrak{q}(y^* - x^*)\big) \\
&= \sup_{y^* \in X} \big(g^*(y^*) - \mu\, \mathfrak{q}(y^*) - \mu\, \mathfrak{q}(x^*) + \mu \langle y^*, x^* \rangle\big) \\
&= -\mu\, \mathfrak{q}(x^*) + \sup_{y^* \in X} \big(\langle y^*, \mu x^* \rangle - \big(\mu\, \mathfrak{q}(y^*) - g^*(y^*)\big)\big) \\
&= -\mu\, \mathfrak{q}(x^*) + \mu \sup_{y^* \in X} \big(\langle y^*, x^* \rangle - \big(\mathfrak{q}(y^*) - \mu^{-1} g^*(y^*)\big)\big) \\
&= -(\mu^{-1} \mathbin{\star} \mathfrak{q})(x^*) + \mu(\mathfrak{q} - \mu^{-1} g^*)^*(x^*). \qquad (13)
\end{aligned}$$

The proof is complete. □

LEMMA 3.7. $\big(\lambda_1 \mathbin{\star} (f_1 + \mu \mathbin{\star} \mathfrak{q}) \mathbin{\blacktriangleleft} \cdots \mathbin{\blacktriangleleft} \lambda_n \mathbin{\star} (f_n + \mu \mathbin{\star} \mathfrak{q})\big)^* = \lambda_1(f_1^* \mathbin{\blacktriangleleft} \mu\, \mathfrak{q}) + \cdots + \lambda_n(f_n^* \mathbin{\blacktriangleleft} \mu\, \mathfrak{q})$.

Proof. Using Proposition 3.2(iii), Proposition 3.2(ii), Fact 3.4(i), and Proposition 3.3(iii), we compute that

$$\begin{aligned}
&\big(\lambda_1 \mathbin{\star} (f_1 + \mu \mathbin{\star} \mathfrak{q}) \mathbin{\blacktriangleleft} \cdots \mathbin{\blacktriangleleft} \lambda_n \mathbin{\star} (f_n + \mu \mathbin{\star} \mathfrak{q})\big)^* \\
&\quad = \big(\lambda_1 \mathbin{\star} (f_1 + \mu \mathbin{\star} \mathfrak{q})\big)^* + \cdots + \big(\lambda_n \mathbin{\star} (f_n + \mu \mathbin{\star} \mathfrak{q})\big)^* \\
&\quad = \lambda_1 (f_1 + \mu \mathbin{\star} \mathfrak{q})^* + \cdots + \lambda_n (f_n + \mu \mathbin{\star} \mathfrak{q})^* \\
&\quad = \lambda_1 \big(f_1^* \mathbin{\blacktriangleleft} (\mu \mathbin{\star} \mathfrak{q})^*\big) + \cdots + \lambda_n \big(f_n^* \mathbin{\blacktriangleleft} (\mu \mathbin{\star} \mathfrak{q})^*\big) \\
&\quad = \lambda_1 (f_1^* \mathbin{\blacktriangleleft} \mu\, \mathfrak{q}) + \cdots + \lambda_n (f_n^* \mathbin{\blacktriangleleft} \mu\, \mathfrak{q}). \qquad (14)
\end{aligned}$$

This completes the proof. □

FACT 3.8. *Let* $(\forall i)$ $x_i \in \operatorname{dom} f_i$, *and set* $x = x_1 + \cdots + x_n$. *Then the following implications hold.*

(i) $(f_1 \mathbin{\blacktriangleleft} \cdots \mathbin{\blacktriangleleft} f_n)(x) = f_1(x_1) + \cdots + f_n(x_n) \Rightarrow \partial(f_1 \mathbin{\blacktriangleleft} \cdots \mathbin{\blacktriangleleft} f_n)(x) = \partial f_1(x_1) \cap \cdots \cap \partial f_n(x_n)$.

(ii) $\partial f_1(x_1) \cap \cdots \cap \partial f_n(x_n) \neq \varnothing \Rightarrow (f_1 \mathbin{\blacktriangleleft} \cdots \mathbin{\blacktriangleleft} f_n)(x) = f_1(x_1) + \cdots + f_n(x_n)$.

Proof. See [20, Corollary 2.4.7]. □

PROPOSITION 3.9. *Let* $f \in \Gamma(X)$, *and let* $\alpha > 0$. *Then* $\partial(0 \mathbin{\star} f) = N_{\{0\}}$ *and* $\partial(\alpha \mathbin{\star} f) = (\partial f) \circ (\alpha^{-1} \operatorname{Id})$.

Proof. Since $0 \mathbin{\star} f = \iota_{\{0\}}$, we deduce that $\partial(0 \mathbin{\star} f) = \partial \iota_{\{0\}} = N_{\{0\}}$. Also, $\partial(\alpha \mathbin{\star} f) = \partial(\alpha f \circ (\alpha^{-1} \operatorname{Id}))$; the formula thus follows from convex calculus (see, e.g., [20, Theorem 2.8.3]). □

4. Definition, reformulations, domain, and exactness. In section 1, we have seen that the idea of computing the averaged Minkowski sum is doomed in general, due to the potential lack of coercivity properties of the terms. The proximal average can be interpreted as a three-step remedy of this idea. First, each function is "coercified" by epi-adding $\mu \mathbin{\star} \mathfrak{q}$. Second, the epi-average of the coercified terms is computed. The third step removes $\mu \mathbin{\star} \mathfrak{q}$ through subtraction. We are now ready to describe the proximal average.

DEFINITION 4.1 (proximal average). *The* $\boldsymbol{\lambda}$-weighted proximal average of $\boldsymbol{f}$ with parameter μ *is*

$$p_\mu(\boldsymbol{f}, \boldsymbol{\lambda}) = \lambda_1 \mathbin{\star} (f_1 + \mu \mathbin{\star} \mathfrak{q}) \mathbin{\blacktriangleleft} \cdots \mathbin{\blacktriangleleft} \lambda_n \mathbin{\star} (f_n + \mu \mathbin{\star} \mathfrak{q}) - \mu \mathbin{\star} \mathfrak{q}, \tag{15}$$

i.e., if $I = \big\{ i \in \{1, \ldots, n\} \mid \lambda_i > 0 \big\}$, *then*

$$(\forall x \in X) \quad p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})(x) = \frac{1}{\mu} \left(-\frac{1}{2}\|x\|^2 + \inf_{\sum_{i \in I} x_i = x} \sum_{i \in I} \lambda_i \left(\mu f_i(x_i/\lambda_i) + \frac{1}{2}\|x_i/\lambda_i\|^2 \right) \right). \tag{16}$$

We also write $p(\boldsymbol{f}, \boldsymbol{\lambda})$ *if* $\mu = 1$, $p_\mu(\boldsymbol{f})$ *if all* λ_i *coincide, and* $p(\boldsymbol{f})$ *if* $\mu = 1$ *and all* λ_i *coincide.*

Remark 4.2. Some immediate consequences of the definition are the following.

(i) $p_\mu(f_1, 1) = f_1$.

(ii) If $I = \big\{ i \in \{1, \ldots, n\} \mid \lambda_i > 0 \big\}$, $\widetilde{\boldsymbol{f}} = (f_i)_{i \in I}$ and $\widetilde{\boldsymbol{\lambda}} = (\lambda_i)_{i \in I}$, then $p_\mu(\boldsymbol{f}, \boldsymbol{\lambda}) = p_\mu(\widetilde{\boldsymbol{f}}, \widetilde{\boldsymbol{\lambda}})$.

(iii) If π is a permutation of $I = \{1, \ldots, n\}$, $\widetilde{\boldsymbol{f}} = (f_{\pi(i)})_{i \in I}$ and $\widetilde{\boldsymbol{\lambda}} = (\lambda_{\pi(i)})_{i \in I}$, then $p_\mu(\boldsymbol{f}, \boldsymbol{\lambda}) = p_\mu(\widetilde{\boldsymbol{f}}, \widetilde{\boldsymbol{\lambda}})$.

(iv) $p_\mu(\boldsymbol{f}, \boldsymbol{\lambda}) = \mu^{-1} p_1(\mu \boldsymbol{f}, \boldsymbol{\lambda})$; equivalently, $p(\mu \boldsymbol{f}, \boldsymbol{\lambda}) = \mu p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})$.

(v) If $\Lambda_{n-1} = \lambda_1 + \cdots + \lambda_{n-1} > 0$, then

$$\begin{aligned} p_1(\boldsymbol{f}, \boldsymbol{\lambda}) &= p_1\big((f_1, \ldots, f_n), (\lambda_1, \ldots, \lambda_n)\big) \\ &= p_1\Big(\Big(p_1\big((f_1, \ldots, f_{n-1}), \Lambda_{n-1}^{-1}(\lambda_1, \ldots, \lambda_{n-1})\big), f_n\Big), (\Lambda_{n-1}, \lambda_n)\Big). \end{aligned} \tag{17}$$

The identities in items (iv) and (v) may be useful if one wishes to develop the theory of results for a general $\mu > 0$ and a general $n \geq 2$ from the simpler case $\mu = 1$ and $n = 2$; however, the direct approach favored in this paper is not only self-contained, but it also yields proofs that we found much more readable. Nonetheless, (iv) and (v) may be convenient for the numerical computation of the proximal average—especially when the simpler case is already implemented [13].

PROPOSITION 4.3 (reformulations).

$$p_\mu(\boldsymbol{f},\boldsymbol{\lambda}) = \big(\lambda_1(f_1^* \mathbin{\blacktriangleleft} \mu\,\mathfrak{q}) + \cdots + \lambda_n(f_n^* \mathbin{\blacktriangleleft} \mu\,\mathfrak{q})\big)^* - \mu^{-1}\,\mathfrak{q} \tag{18}$$

$$= \big(\lambda_1(f_1 + \mu^{-1}\,\mathfrak{q})^* + \cdots + \lambda_n(f_n + \mu^{-1}\,\mathfrak{q})^*\big)^* - \mu^{-1}\,\mathfrak{q} \tag{19}$$

and

$$(\forall x \in X)\quad p_\mu(\boldsymbol{f},\boldsymbol{\lambda})(x) = \inf_{\sum \lambda_i y_i = x} \sum \lambda_i f_i(y_i) + \frac{1}{\mu}\bigg(\Big(\sum \lambda_i\,\mathfrak{q}(y_i)\Big) - \mathfrak{q}(x)\bigg). \tag{20}$$

Proof. By Proposition 3.1(iv), $(\forall i)$ $\operatorname{dom}(f_i^* \mathbin{\blacktriangleleft} \mu\,\mathfrak{q}) = (\operatorname{dom} f_1^*) + (\operatorname{dom} \mu\,\mathfrak{q}) = X$. Fact 3.4(i), Proposition 3.2(i), Proposition 3.2(iii), and Proposition 3.3(iv) imply that

$$\begin{aligned}
\big(\lambda_1(f_1^* \mathbin{\blacktriangleleft} \mu\,\mathfrak{q}) + \cdots + \lambda_n(f_n^* \mathbin{\blacktriangleleft} \mu\,\mathfrak{q})\big)^* &= (\lambda_1(f_1^* \mathbin{\blacktriangleleft} \mu\,\mathfrak{q}))^* \mathbin{\blacktriangleleft} \cdots \mathbin{\blacktriangleleft} (\lambda_n(f_n^* \mathbin{\blacktriangleleft} \mu\,\mathfrak{q}))^* \\
&= \lambda_1 \star (f_1^* \mathbin{\blacktriangleleft} \mu\,\mathfrak{q})^* \mathbin{\blacktriangleleft} \cdots \mathbin{\blacktriangleleft} \lambda_n \star (f_n^* \mathbin{\blacktriangleleft} \mu\,\mathfrak{q})^* \\
&= \lambda_1 \star (f_1^{**} + (\mu\,\mathfrak{q})^*) \mathbin{\blacktriangleleft} \cdots \mathbin{\blacktriangleleft} \lambda_n \star (f_n^{**} + (\mu\,\mathfrak{q})^*) \\
&= \lambda_1 \star (f_1 + \mu \star \mathfrak{q}) \mathbin{\blacktriangleleft} \cdots \mathbin{\blacktriangleleft} \lambda_n \star (f_n + \mu \star \mathfrak{q}). \qquad (21)
\end{aligned}$$

This and Proposition 3.3(ii) yield (18). In turn, Fact 3.4(i) and Proposition 3.3(iv) imply (19). Changing variables, we see that (20) is equivalent to (16). □

Remark 4.4 (some history). In [5], the proximal average was considered for $n = 2$ and $\mu = 1$, and written equivalently as

$$\big(\lambda_1(f_1^* \mathbin{\blacktriangleleft} \mathfrak{q}) + \lambda_2(f_2^* \mathbin{\blacktriangleleft} \mathfrak{q})\big)^* - \mathfrak{q}; \tag{22}$$

see (18). The function (22) was utilized in [5] to explicitly illustrate Moreau's observation [16] that the set of proximal mappings is convex. More recently, the proximal average was considered in [3], again with $n = 2$ and $\mu = 1$, though it was written as (see (19))

$$\big(\lambda_1(f_1 + \mathfrak{q})^* + \lambda_2(f_2 + \mathfrak{q})^*\big)^* - \mathfrak{q}. \tag{23}$$

Example 4.5 (connection to means of numbers). Let $\alpha_1, \ldots, \alpha_n$ be strictly positive numbers and suppose that $(\forall i)$ $f_i = \alpha_i\,\mathfrak{q}$. Using (19), we see that

$$\begin{aligned}
p_{\mu^{-1}}(\boldsymbol{f},\boldsymbol{\lambda}) &= \left(\sum_{i=1}^n \lambda_i(\alpha_i\,\mathfrak{q} + \mu\,\mathfrak{q})^*\right)^* - \mu\,\mathfrak{q} \\
&= \left(\sum_{i=1}^n \frac{\lambda_i}{\alpha_i + \mu}\,\mathfrak{q}\right)^* - \mu\,\mathfrak{q} = \left(\sum_{i=1}^n \frac{\lambda_i}{\alpha_i + \mu}\right)^{-1} \mathfrak{q} - \mu\,\mathfrak{q}, \qquad (24)
\end{aligned}$$

and thus

$$p_\mu(\boldsymbol{f},\boldsymbol{\lambda}) = \left(\left(\sum_{i=1}^n \frac{\lambda_i}{\alpha_i + \mu^{-1}}\right)^{-1} - \mu^{-1}\right)\mathfrak{q}. \tag{25}$$

Denote the coefficient of $\mathfrak{q}$ in (25) by δ. Since δ is the difference of the weighted harmonic mean of $\alpha_1 + \mu^{-1}, \ldots, \alpha_n + \mu^{-1}$ and μ^{-1}, the harmonic-arithmetic mean inequality implies that δ does not exceed the weighted arithmetic mean

$$\sum_{i=1}^n \lambda_i \alpha_i. \tag{26}$$

As $\mu \to +\infty$, we note that δ converges to the weighted harmonic mean

$$\left(\sum_{i=1}^{n} \frac{\lambda_i}{\alpha_i}\right)^{-1}, \tag{27}$$

while a calculus exercise shows that δ approaches, as $\mu \to 0^+$, the weighted arithmetic mean (26). In Remark 8.6, we revisit this example from a more general point of view.

The next result locates the domain of the proximal average exactly; moreover, it strengthens [3, Theorem 4.11], where equality was observed only for the closures and interiors.

THEOREM 4.6 (domain). $\operatorname{dom} p_\mu(\boldsymbol{f}, \boldsymbol{\lambda}) = \lambda_1 \operatorname{dom} f_1 + \cdots + \lambda_n \operatorname{dom} f_n$.

Proof. Using Proposition 3.1(iv) and Proposition 3.1(ii), we obtain $\operatorname{dom} p_\mu(\boldsymbol{f}, \boldsymbol{\lambda}) = \operatorname{dom}(\lambda_1 \star (f_1 + \mu \star \mathfrak{q})) + \cdots + \operatorname{dom}(\lambda_1 \star (f_1 + \mu \star \mathfrak{q})) = \lambda_1 \operatorname{dom}(f_1 + \mu \star \mathfrak{q}) + \cdots + \lambda_n \operatorname{dom}(f_n + \mu \star \mathfrak{q}) = \lambda_1 \operatorname{dom}(f_1) + \cdots + \lambda_n \operatorname{dom}(f_n)$. ∎

COROLLARY 4.7. *Suppose that at least one function f_i has full domain and that $\lambda_i > 0$. Then $p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})$ has full domain.*

Example 4.8. Assume each $\lambda_i > 0$ and each $f_i = \iota_{C_i}$ where C_i is a nonempty closed convex subset of X. In X^n, set $H = \{(z_i) \mid \sum \sqrt{\lambda_i} z_i = 0\}$ and $(\forall x \in X)\ D_x$ is the Cartesian product $\bigtimes \left(\sqrt{\lambda_i} C_i - \sqrt{\lambda_i} x\right)$. Then

$$p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})\colon X \to \left]-\infty, +\infty\right] : x \mapsto \tfrac{1}{2\mu} d^2_{H \cap D_x}(0). \tag{28}$$

Proof. Fix $x \in X$. Using (16), we obtain

$$\begin{aligned}
p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})(x) &= \mu^{-1} \left(-\frac{1}{2}\|x\|^2 + \inf_{\sum_i x_i = x} \sum \lambda_i \left(\mu\, \iota_{C_i}(x_i/\lambda_i) + \frac{1}{2}\|x_i/\lambda_i\|^2 \right) \right) \\
&= \mu^{-1} \inf_{\substack{\text{each } c_i \in C_i \\ \sum \lambda_i c_i = x}} \sum \lambda_i \left(\frac{1}{2}\|c_i\|^2 - \frac{1}{2}\|x\|^2 \right) \\
&= \mu^{-1} \inf_{z=(z_i) \in H \cap D_x} \sum \lambda_i \left(\frac{1}{2}\left\|x + z_i/\sqrt{\lambda_i}\right\|^2 - \frac{1}{2}\|x\|^2 \right) \\
&= \mu^{-1} \inf_{z=(z_i) \in H \cap D_x} \sum \frac{1}{2}\|z_i\|^2, \qquad (29)
\end{aligned}$$

which completes the proof. ∎

Remark 4.9. Consider Example 4.8 with $n = 2$, $\mu = 1$, $\lambda_1 > 0$, and $\lambda_2 > 0$. Then (28) simplifies to

$$p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})\colon X \to \left]-\infty, +\infty\right] : x \mapsto \frac{1}{2\lambda_1\lambda_2} d^2_{(\lambda_1(C_1 - x)) \cap (\lambda_2(x - C_2))}(0), \tag{30}$$

which is a formula first observed in [5, Theorem 6.1].

THEOREM 4.10 (exactness). *For every $x \in \operatorname{dom} p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})$ there exist $y_i \in \lambda_i \operatorname{dom} f_i$ such that $x = y_1 + \cdots + y_n$ and $p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})(x) = (\lambda_1 \star (f_1 + \mu \star \mathfrak{q}))(y_1) + \cdots + (\lambda_n \star (f_n + \mu \star \mathfrak{q}))(y_n) - (\mu \star \mathfrak{q})(x)$.*

Proof. Set $(\forall i)\ g_i = \lambda_i \star (f_i + \mu \star \mathfrak{q})$. If $\lambda_i = 0$, then $g_i = \iota_{\{0\}}$, and hence $g_i^* = \iota_X$ has full domain. If $\lambda_i > 0$, then using Proposition 3.2(i), Fact 3.4(i), and Proposition 3.3(iv), we see that

$$g_i^* = \left(\lambda_i \star (f_i + \mu \star \mathfrak{q})\right)^* = \lambda_i (f_i + \mu \star \mathfrak{q})^* = \lambda_i (f_i^* \mathbin{\blacktriangleright} (\mu \star \mathfrak{q})^*) = \lambda_i (f_i^* \mathbin{\blacktriangleright} \mu\, \mathfrak{q}); \tag{31}$$

thus, g_i^* also has full domain. Therefore, by Fact 3.4(ii), the epi-sum

$$p_\mu(\boldsymbol{f},\boldsymbol{\lambda}) + \mu \star \mathfrak{q} = g_1 \mathbin{\blacklozenge} \cdots \mathbin{\blacklozenge} g_n \tag{32}$$

is *exact*. Since $\operatorname{dom} p_\mu(\boldsymbol{f},\boldsymbol{\lambda}) = \lambda_1 \operatorname{dom} f_1 + \cdots + \lambda_n \operatorname{dom} f_n$ by Theorem 4.6, the existence of the y_i is now clear. □

5. Fenchel conjugate. In this section, we compute the Fenchel conjugate of the proximal average. The explicit form obtained has several interesting consequences. We begin with a reformulation of Lemma 3.7:

$$(p_\mu(\boldsymbol{f},\boldsymbol{\lambda}) + \mu \star \mathfrak{q})^* = \lambda_1(f_1^* \mathbin{\blacklozenge} \mu\,\mathfrak{q}) + \cdots + \lambda_n(f_n^* \mathbin{\blacklozenge} \mu\,\mathfrak{q}). \tag{33}$$

We are now ready for a useful generalization of [5, Theorem 6.1] where $n = 2$ and $\mu = 1$.

THEOREM 5.1 (Fenchel conjugate). $\big(p_\mu(\boldsymbol{f},\boldsymbol{\lambda})\big)^* = p_{\mu^{-1}}(\boldsymbol{f}^*,\boldsymbol{\lambda})$.

Proof. Set

$$g = p_\mu(\boldsymbol{f},\boldsymbol{\lambda}) + \mu \star \mathfrak{q}. \tag{34}$$

By (33), we have

$$g^* = \lambda_1(f_1^* \mathbin{\blacklozenge} \mu\,\mathfrak{q}) + \cdots + \lambda_n(f_n^* \mathbin{\blacklozenge} \mu\,\mathfrak{q}). \tag{35}$$

In view of (6), (35), Proposition 3.1(vi), Proposition 3.3(v), and Proposition 3.2(i), we obtain that

$$\begin{aligned}
\mathfrak{q} - \mu^{-1}g^* &= \lambda_1\big(\mathfrak{q} - \mu^{-1}(f_1^* \mathbin{\blacklozenge} \mu\,\mathfrak{q})\big) + \cdots + \lambda_n\big(\mathfrak{q} - \mu^{-1}(f_1^* \mathbin{\blacklozenge} \mu\,\mathfrak{q})\big)\\
&= \lambda_1\big(\mathfrak{q} - (\mu^{-1}f_1^* \mathbin{\blacklozenge} \mathfrak{q})\big) + \cdots + \lambda_n\big(\mathfrak{q} - (\mu^{-1}f_1^* \mathbin{\blacklozenge} \mathfrak{q})\big)\\
&= \lambda_1\big((\mu^{-1}f_1^*)^* \mathbin{\blacklozenge} \mathfrak{q}\big) + \cdots + \lambda_n\big((\mu^{-1}f_n^*)^* \mathbin{\blacklozenge} \mathfrak{q}\big)\\
&= \lambda_1\big(\mu^{-1} \star f_1 \mathbin{\blacklozenge} \mathfrak{q}\big) + \cdots + \lambda_n\big(\mu^{-1} \star f_n \mathbin{\blacklozenge} \mathfrak{q}\big). \qquad (36)
\end{aligned}$$

Consequently, using Fact 3.4(i), Proposition 3.2(i), Proposition 3.2(iii), Proposition 3.2(ii), Proposition 3.3(i), we see that

$$\begin{aligned}
\big(\mathfrak{q} - \mu^{-1}g^*\big)^* &= \Big(\lambda_1\big(\mu^{-1} \star f_1 \mathbin{\blacklozenge} \mathfrak{q}\big) + \cdots + \lambda_n\big(\mu^{-1} \star f_n \mathbin{\blacklozenge} \mathfrak{q}\big)\Big)^*\\
&= \Big(\lambda_1\big(\mu^{-1} \star f_1 \mathbin{\blacklozenge} \mathfrak{q}\big)\Big)^* \mathbin{\blacklozenge} \cdots \mathbin{\blacklozenge} \Big(\lambda_n\big(\mu^{-1} \star f_n \mathbin{\blacklozenge} \mathfrak{q}\big)\Big)^*\\
&= \lambda_1 \star \big(\mu^{-1} \star f_1 \mathbin{\blacklozenge} \mathfrak{q}\big)^* \mathbin{\blacklozenge} \cdots \mathbin{\blacklozenge} \lambda_n \star \big(\mu^{-1} \star f_n \mathbin{\blacklozenge} \mathfrak{q}\big)^*\\
&= \lambda_1 \star \big((\mu^{-1} \star f_1)^* + \mathfrak{q}^*\big) \mathbin{\blacklozenge} \cdots \mathbin{\blacklozenge} \lambda_n \star \big((\mu^{-1} \star f_n)^* + \mathfrak{q}^*\big)\\
&= \lambda_1 \star \big(\mu^{-1}f_1^* + \mathfrak{q}\big) \mathbin{\blacklozenge} \cdots \mathbin{\blacklozenge} \lambda_n \star \big(\mu^{-1}f_n^* + \mathfrak{q}\big). \qquad (37)
\end{aligned}$$

Now Proposition 3.1(vi), Proposition 3.1(ix), and Proposition 3.3(ii) imply that

$$\begin{aligned}
\mu\big(\mathfrak{q} - \mu^{-1}g^*\big)^* &= \mu\Big(\lambda_1 \star \big(\mu^{-1}(f_1^* + \mu\,\mathfrak{q})\big) \mathbin{\blacklozenge} \cdots \mathbin{\blacklozenge} \lambda_n \star \big(\mu^{-1}(f_n^* + \mu\,\mathfrak{q})\big)\Big)\\
&= \mu\Big(\lambda_1 \star \big(\mu^{-1}(f_1^* + \mu\,\mathfrak{q})\big)\Big) \mathbin{\blacklozenge} \cdots \mathbin{\blacklozenge} \mu\Big(\lambda_n \star \big(\mu^{-1}(f_n^* + \mu\,\mathfrak{q})\big)\Big)\\
&= \lambda_1 \star (f_1^* + \mu\,\mathfrak{q}) \mathbin{\blacklozenge} \cdots \mathbin{\blacklozenge} \lambda_n \star (f_n^* + \mu\,\mathfrak{q})\\
&= \lambda_1 \star (f_1^* + \mu^{-1} \star \mathfrak{q}) \mathbin{\blacklozenge} \cdots \mathbin{\blacklozenge} \lambda_n \star (f_n^* + \mu^{-1} \star \mathfrak{q}). \qquad (38)
\end{aligned}$$

Combining (34), Corollary 3.6, and (38), we conclude that

$$\begin{aligned}\big(p_\mu(\boldsymbol{f},\boldsymbol{\lambda})\big)^* &= (g-\mu\mathbin{\star}\mathfrak{q})^*\\ &= \mu(\,\mathfrak{q}-\mu^{-1}g^*)^*-\mu^{-1}\mathbin{\star}\mathfrak{q}\\ &= \lambda_1\mathbin{\star}(f_1^*+\mu^{-1}\mathbin{\star}\mathfrak{q})\mathbin{\blacktriangleleft}\cdots\mathbin{\blacktriangleleft}\lambda_n\mathbin{\star}(f_n^*+\mu^{-1}\mathbin{\star}\mathfrak{q})-\mu^{-1}\mathbin{\star}\mathfrak{q}\\ &= p_{\mu^{-1}}(\boldsymbol{f}^*,\boldsymbol{\lambda}),\end{aligned}\tag{39}$$

as claimed. □

COROLLARY 5.2 (lower semicontinuity). *$p_\mu(\boldsymbol{f},\boldsymbol{\lambda})$ is convex, lower semicontinuous, and proper.*

Proof. Applying Theorem 5.1 twice, we deduce that $(p_\mu(\boldsymbol{f},\boldsymbol{\lambda}))^{**}=(p_{\mu^{-1}}(\boldsymbol{f}^*,\boldsymbol{\lambda}))^*$ $=p_{(\mu^{-1})^{-1}}(\boldsymbol{f}^{**},\boldsymbol{\lambda})=p_\mu(\boldsymbol{f},\boldsymbol{\lambda})$. □

The next result refines the corresponding two-function version [3, Proposition 4.8].

Example 5.3. $p(\boldsymbol{f},\boldsymbol{f}^*)=\mathfrak{q}$.

Proof. Theorem 5.1 readily implies that the $p(\boldsymbol{f},\boldsymbol{f}^*)$ is equal to its conjugate; consequently, it must be equal to $\mathfrak{q}$ by Proposition 3.3(i). □

THEOREM 5.4 (inequalities). $(\lambda_1f_1^*+\cdots+\lambda_nf_n^*)^*\le p_\mu(\boldsymbol{f},\boldsymbol{\lambda})\le\lambda_1f_1+\cdots+\lambda_nf_n$.

Proof. The right inequality follows from (20) (by setting $y_i=x$). Applying the right inequality to $\boldsymbol{f}^*$ and μ^{-1}, we learn that

$$p_{\mu^{-1}}(\boldsymbol{f}^*,\boldsymbol{\lambda})\le\lambda_1f_1^*+\cdots+\lambda_nf_n^*.\tag{40}$$

Taking the Fenchel conjugate of (40) and utilizing Theorem 5.1, we deduce that $p_\mu(\boldsymbol{f},\boldsymbol{\lambda})=\big(p_{\mu^{-1}}(\boldsymbol{f}^*,\boldsymbol{\lambda})\big)^*\ge\big(\lambda_1f_1^*+\cdots+\lambda_nf_n^*\big)^*$. □

COROLLARY 5.5 (infimum value).

$$\lambda_1\inf f_1+\cdots+\lambda_n\inf f_n\le\inf p_\mu(\boldsymbol{f},\boldsymbol{\lambda})\le\inf(\lambda_1f_1+\cdots+\lambda_nf_n).\tag{41}$$

COROLLARY 5.6 (common minimizers). *Suppose that $\bigcap_{i\colon\lambda_i>0}\operatorname{argmin}(f_i)\neq\varnothing$. Then*

$$\min p_\mu(\boldsymbol{f},\boldsymbol{\lambda})=\sum_{i\colon\lambda_i>0}\lambda_i\min f_i\quad\textit{and}\quad\operatorname{argmin}p_\mu(\boldsymbol{f},\boldsymbol{\lambda})=\bigcap_{i\colon\lambda_i>0}\operatorname{argmin}(f_i).\tag{42}$$

Proof. Combine Theorem 5.4 and Corollary 5.5. □

6. Moreau envelope and proximal mapping.

DEFINITION 6.1. *Let $f\in\Gamma(X)$. The* Moreau envelope *of f with parameter μ is* $e_\mu f=f\mathbin{\blacktriangleleft}\mu\mathbin{\star}\mathfrak{q}$.

Observe that

$$e_\mu f=(f^*+\mu\,\mathfrak{q})^*.\tag{43}$$

THEOREM 6.2 (Moreau envelope and its Fenchel conjugate).

(i) $e_\mu p_\mu(\boldsymbol{f},\boldsymbol{\lambda})=\lambda_1e_\mu f_1+\cdots+\lambda_ne_\mu f_n$.

(ii) $\big(e_\mu p_\mu(\boldsymbol{f},\boldsymbol{\lambda})\big)^*=\lambda_1\mathbin{\star}(e_\mu f_1)^*\mathbin{\blacktriangleleft}\cdots\mathbin{\blacktriangleleft}\lambda_n\mathbin{\star}(e_\mu f_n)^*$.

Proof. Fix $y \in X$, and set $I = \big\{i \in \{1,\ldots,n\} \mid \lambda_i > 0\big\}$. Using (16), we obtain

$$\begin{aligned}
\big(e_\mu p_\mu(\boldsymbol{f},\boldsymbol{\lambda})\big)(y) &= \inf_x p_\mu(\boldsymbol{f},\boldsymbol{\lambda})(x) + \frac{1}{2\mu}\|y-x\|^2 \\
&= \inf_x \inf_{\sum_{i\in I} x_i = x} \sum_{i\in I} \lambda_i\Big(f_i(x_i/\lambda_i) + \frac{1}{2\mu}\|x_i/\lambda_i\|^2\Big) \\
&\qquad + \frac{1}{2\mu}\|y\|^2 - \frac{1}{\mu}\langle x, y\rangle \\
&= \inf_x \inf_{\sum_{i\in I} x_i = x} \sum_{i\in I} \lambda_i\Big(f_i(x_i/\lambda_i) \\
&\qquad + \frac{1}{2\mu}\|x_i/\lambda_i\|^2 + \frac{1}{2\mu}\|y\|^2 - \frac{1}{\mu}\langle x_i/\lambda_i, y\rangle\Big) \\
&= \inf_x \inf_{\sum_{i\in I} x_i = x} \sum_{i\in I} \lambda_i\Big(f_i(x_i/\lambda_i) + \frac{1}{2\mu}\|y - x_i/\lambda_i\|^2\Big) \\
&= \inf_{x_i, i\in I} \sum_{i\in I} \lambda_i\Big(f_i(x_i/\lambda_i) + \frac{1}{2\mu}\|y - x_i/\lambda_i\|^2\Big) \\
&= \sum_{i\in I} \lambda_i \inf_{x_i}\Big(f_i(x_i/\lambda_i) + \frac{1}{2\mu}\|y - x_i/\lambda_i\|^2\Big) \\
&= \sum_{i\in I} \lambda_i \big(e_\mu f_i\big)(y). \qquad (44)
\end{aligned}$$

This implies (i), and (ii) follows by Fenchel conjugation. Alternatively, using Definition 6.1, Proposition 3.2(iii), Theorem 5.1, Proposition 3.3(iv), and Proposition 3.3(ii), one may prove (ii) via $\big(e_\mu p_\mu(\boldsymbol{f},\boldsymbol{\lambda})\big)^* = \big(p_\mu(\boldsymbol{f},\boldsymbol{\lambda}) \mathbin{\blacktriangleleft} \mu\,\mathfrak{q}\big)^* = \big(p_\mu(\boldsymbol{f},\boldsymbol{\lambda})\big)^* + \mu\,\mathfrak{q} = p_{\mu^{-1}}(\boldsymbol{f}^*,\boldsymbol{\lambda}) + \mu^{-1}\star\mathfrak{q} = \lambda_1\star(f_1^* + \mu^{-1}\star\mathfrak{q}) \mathbin{\blacktriangleleft} \cdots \mathbin{\blacktriangleleft} \lambda_n\star(f_n^* + \mu^{-1}\star\mathfrak{q}) = \lambda_1\star(f_1^* + \mu\,\mathfrak{q}) \mathbin{\blacktriangleleft} \cdots \mathbin{\blacktriangleleft} \lambda_n\star(f_n^* + \mu\,\mathfrak{q}) = \lambda_1\star(e_\mu f_1)^* \mathbin{\blacktriangleleft} \cdots \mathbin{\blacktriangleleft} \lambda_n\star(e_\mu f_n)^*$ and then deduces (i) by Fenchel conjugation. □

The following result is well known.

PROPOSITION 6.3. *Let $f \in \Gamma(X)$. Then* $\operatorname{argmin} e_\mu f = \operatorname{argmin} f$.

Proof. $\operatorname{argmin} e_\mu f = \partial(e_\mu f)^*(0) = \partial(f^* + \mu\,\mathfrak{q})(0) = (\partial f^* + \mu\,\mathrm{Id})(0) = \partial f^*(0) = \operatorname{argmin} f$. □

COROLLARY 6.4 (minimizers). $\operatorname{argmin} p_\mu(\boldsymbol{f},\boldsymbol{\lambda}) = \operatorname{argmin}\big(\lambda_1 e_\mu f_1 + \cdots + \lambda_n e_\mu f_n\big)$.

Proof. Combine Proposition 6.3 and Theorem 6.2(i). □

Example 6.5 (least-squares solutions). Let $C_1,\ldots,C_n$ be nonempty closed convex subsets of X, and suppose that $(\forall i)\ f_i = \iota_{C_i}$. Then $\operatorname{argmin} p_\mu(\boldsymbol{f},\boldsymbol{\lambda}) = \operatorname{argmin}(\lambda_1 d_{C_1}^2 + \cdots + \lambda_n d_{C_n}^2)$.

Proof. This is a consequence of Corollary 6.4, since $(\forall i)\ e_\mu f_i = e_\mu \iota_{C_i} = \iota_{C_i} \mathbin{\blacktriangleleft} \mu\star\mathfrak{q} = \mu^{-1}\iota_{C_i} \mathbin{\blacktriangleleft} \mu^{-1}\,\mathfrak{q} = \mu^{-1}(\iota_{C_i} \mathbin{\blacktriangleleft} \mathfrak{q}) = \mu^{-1}\frac{1}{2}d_{C_i}^2$. □

DEFINITION 6.6. *Let $f \in \Gamma(X)$. The* proximal mapping *of f with parameter μ is* $P_\mu f = (\mathrm{Id} + \mu\partial f)^{-1}$.

Observe that

$$\mu^{-1}(P_\mu f)^{-1} = \partial f + \mu^{-1}\,\mathrm{Id}, \qquad (45)$$

that

$$P_\mu f = \big(\nabla(f + \mu^{-1}\,\mathfrak{q})^*\big) \circ (\mu^{-1}\,\mathrm{Id}), \qquad (46)$$

and that

$$(P_\mu f) \circ (\mu \operatorname{Id}) = \nabla(e_{\mu^{-1}}(f^*)). \tag{47}$$

We now show that the proximal mapping of the proximal average is simply the average of the individual proximal mappings. This result, which also explains how the proximal average got its name, was first proved in [5, Theorem 6.1] when $n = 2$ and $\mu = 1$.

THEOREM 6.7 (proximal mapping). $P_\mu\big(p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})\big) = \lambda_1 P_\mu f_1 + \cdots + \lambda_n P_\mu f_n$.

Proof. Theorem 5.1 and Theorem 6.2(i) (the latter applied to $\boldsymbol{f}^*$ and μ^{-1}) show that

$$e_{\mu^{-1}}\big((p_\mu(\boldsymbol{f}, \boldsymbol{\lambda}))^*\big) = e_{\mu^{-1}}\big(p_{\mu^{-1}}(\boldsymbol{f}^*, \boldsymbol{\lambda})\big) = \lambda_1 e_{\mu^{-1}}(f_1^*) + \cdots + \lambda_n e_{\mu^{-1}}(f_n^*); \tag{48}$$

in turn, taking gradients yields

$$\nabla\big(e_{\mu^{-1}}\big((p_\mu(\boldsymbol{f}, \boldsymbol{\lambda}))^*\big)\big) = \lambda_1 \nabla(e_{\mu^{-1}}(f_1^*)) + \cdots + \lambda_n \nabla\big(e_{\mu^{-1}}(f_n^*)\big). \tag{49}$$

Using (47), we see that this is equivalent to

$$\big(P_\mu\big(p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})\big)\big) \circ (\mu \operatorname{Id}) = \lambda_1 (P_\mu f_1) \circ (\mu \operatorname{Id}) + \cdots + \lambda_n (P_\mu f_n) \circ (\mu \operatorname{Id}). \tag{50}$$

The result follows. ∎

7. Subdifferential.

THEOREM 7.1 (subdifferential). *Let* $(\forall i)\ x_i \in \operatorname{dom} f_i$, *and set* $x = \lambda_1 x_1 + \cdots + \lambda_n x_n$. *Then the following hold.*

(i) *If* $p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})(x) = (\lambda_1 \bigstar (f_1 + \mu \star \mathfrak{q}))(\lambda_1 x_1) + \cdots + (\lambda_n \bigstar (f_n + \mu \star \mathfrak{q}))(\lambda_n x_n) - (\mu \star \mathfrak{q})(x)$, *then*

$$\partial p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})(x) = -\mu^{-1} x + \bigcap_i \partial(\lambda_i \bigstar (f_i + \mu \star \mathfrak{q}))(\lambda_i x_i) \tag{51}$$

$$= -\mu^{-1} x + \bigcap_{i\colon \lambda_i > 0} \big(\partial f_i(x_i) + \mu^{-1} x_i\big) \tag{52}$$

$$= -\mu^{-1} x + \bigcap_{i\colon \lambda_i > 0} \big(\mu^{-1}(P_\mu f_i)^{-1}(x_i)\big). \tag{53}$$

(ii) *If* $\bigcap_{i\colon \lambda_i > 0} (P_\mu f_i)^{-1}(x_i) \neq \varnothing$, *then*

$$p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})(x) = (\lambda_1 \bigstar (f_1 + \mu \star \mathfrak{q}))(\lambda_1 x_1) + \cdots + (\lambda_n \bigstar (f_n + \mu \star j))(\lambda_n x_n) - (\mu \star \mathfrak{q})(x). \tag{54}$$

Proof. Set $(\forall i)\ g_i = \lambda_i \bigstar (f_i + \mu \star \mathfrak{q})$. Theorem 4.6, Theorem 4.10, and Proposition 3.3(ii) imply that

$$g_1 \mathbin{\blacktriangleleft} \cdots \mathbin{\blacktriangleleft} g_n = p_\mu(\boldsymbol{f}, \boldsymbol{\lambda}) + \mu \star \mathfrak{q} = p_\mu(\boldsymbol{f}, \boldsymbol{\lambda}) + \mu^{-1}\, \mathfrak{q} \tag{55}$$

is exact on $\operatorname{dom}(g_1 \mathbin{\blacktriangleleft} \cdots \mathbin{\blacktriangleleft} g_n) = \lambda_1 \operatorname{dom} f_1 + \cdots + \lambda_n \operatorname{dom} f_n = \operatorname{dom} p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})$. (i): (51), (52), and (53) follow from Fact 3.8(i), Proposition 3.9, and (45), respectively. (ii): Use Fact 3.8(ii). ∎

COROLLARY 7.2. $(\forall x \in X) \bigcap_{i\colon \lambda_i > 0} \partial f_i(x) \subseteq \partial p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})(x)$.

Proof. Take $x^* \in \bigcap_{i\colon \lambda_i > 0} \partial f_i(x)$. Then $(\forall i)\ \lambda_i > 0 \Rightarrow \mu x^* + x \in \mu \partial f_i(x) + x = (P_\mu f_i)^{-1}(x)$. By Theorem 7.1(ii), $p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})(x) = (\lambda_1 \bigstar (f_1 + \mu \star \mathfrak{q}))(\lambda_1 x) + \cdots +$

$(\lambda_n \star (f_n + \mu \star j))(\lambda_n x) - (\mu \star \mathfrak{q})(x)$. Using Theorem 7.1(i), we deduce that $x^* = -\mu^{-1}x + \mu^{-1}(\mu x^* + x) \in \partial p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})(x)$. ☐

For the following results, it will be convenient to write $x = x_1 \oplus \cdots \oplus x_n$ if $x = x_1 + \cdots + x_n$ and $x_i \perp x_j$ for $i \neq j$. We also write $K_1 \oplus \cdots \oplus K_n = \big\{x_1 \oplus \cdots \oplus x_n \mid$ each $x_i \in K_i$ and $x_i \perp x_j$ for $i \neq j\big\}$.

COROLLARY 7.3. *Let $K_1, \ldots, K_n$ be nonempty closed convex cones, and set $(\forall i)$ $P_i = P_{K_i}$, the orthogonal projector onto K_i. Suppose that*

$$(\forall x = x_1 \oplus \cdots \oplus x_n \in K_1 \oplus \cdots \oplus K_n)(\forall i) \quad P_i x = x_i, \tag{56}$$

that

$$(\forall x \in X) \quad x = P_1 x \oplus \cdots \oplus P_n x, \tag{57}$$

and that $(\forall i)$ $f_i = \iota_{K_i}$ and $\lambda_i > 0$. Then

$$(\forall x \in X) \quad p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})(x) = \frac{1}{2\mu} \sum_i \frac{(1-\lambda_i)}{\lambda_i} \|P_i x\|^2. \tag{58}$$

Proof. Observe that $(\forall i)$ $P_\mu f_i = (\mathrm{Id} + \mu \partial \iota_{K_i})^{-1} = (\mathrm{Id} + \partial \iota_{K_i})^{-1} = P_i$. Take $x \in X$ and set

$$(\forall i) \quad x_i = \frac{1}{\lambda_i} P_i x = P_i\left(\frac{1}{\lambda_i} x\right). \tag{59}$$

Using (57), we obtain that

$$x = \lambda_1 x_1 \oplus \cdots \oplus \lambda_n x_n. \tag{60}$$

Now set

$$z = x_1 \oplus \cdots \oplus x_n. \tag{61}$$

By (56), we have $(\forall i)$ $P_i z = x_i$. Thus $z \in \bigcap_i (P_\mu f_i)^{-1}(x_i)$. Therefore, by (60) and Theorem 7.1(ii),

$$\begin{aligned} p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})(x) &= (\lambda_1 \star (f_1 + \mu \star \mathfrak{q}))(\lambda_1 x_1) + \cdots + (\lambda_n \star (f_n + \mu \star \mathfrak{q}))(\lambda_n x_n) - (\mu \star \mathfrak{q})(x) \\ &= \mu^{-1}\lambda_1\, \mathfrak{q}(x_1) + \cdots + \mu^{-1}\lambda_n\, \mathfrak{q}(x_n) - \mu^{-1}\, \mathfrak{q}(x) \\ &= \frac{1}{2\mu}\big(\lambda_1 \|x_1\|^2 + \cdots + \lambda_n \|x_n\|^2 - \|\lambda_1 x_1 + \cdots + \lambda_n x_n\|^2\big) \\ &= \frac{1}{2\mu} \sum_i \lambda_i (1-\lambda_i) \|x_i\|^2. \end{aligned} \tag{62}$$

The conclusion thus follows from (59). ☐

The following two examples are special cases of Corollary 7.3.

Example 7.4. Let $K_1, \ldots, K_n$ be closed subspaces that are pairwise orthogonal and such that $K_1 \oplus \cdots \oplus K_n = X$, and suppose that $f_i = \iota_{K_i}$. Then $p_\mu(\boldsymbol{f}, \boldsymbol{\lambda}) = \mu^{-1} \sum_i (\lambda_i^{-1} - 1)(\mathfrak{q} \circ P_{K_i})$.

Example 7.5 (See also [3, Example 4.9]). Let K be a nonempty closed convex cone in X, and let $\lambda \in]0,1[$. Then

$$(\forall x \in X) \quad p\big((\iota_K, \iota_{K^\ominus}), (1-\lambda, \lambda)\big)(x) = \frac{1}{2(1-\lambda)\lambda}\big(\lambda^2 \|P_K x\|^2 + (1-\lambda)^2 \|P_{K^\ominus} x\|^2\big), \tag{63}$$

where $K^\ominus$ is the polar cone of K.

Remark 7.6. We are now in a position to show that the inequalities in Theorem 5.4 can be strict. Suppose that $n = 2$, that $f_1 = \iota_K$, and that $f_2 = \iota_{K^\ominus}$ where K is a nonempty closed convex cone in X, and that $\lambda_2 = \lambda \in]0,1[$. Using Example 7.5, we see that Theorem 5.4 becomes

$$(\forall x \in X) \quad \iota_X(x) \leq \frac{1}{2(1-\lambda)\lambda}\big(\lambda^2\|P_K x\|^2 + (1-\lambda)^2\|P_{K^\ominus}x\|^2\big) \leq \iota_{\{0\}}(x). \tag{64}$$

The inequalities are strict for every $x \in X \smallsetminus \{0\}$.

Let $f \in \Gamma(X)$. Following [2, section 5], we say that f is *essentially smooth* if ∂f is at most single valued and $\operatorname{int}\operatorname{dom} f$ is nonempty, that f is *essentially strictly convex* if f^* is essentially smooth, and that f is *Legendre* if f is both essentially smooth and essentially strictly convex. These notions coincide in our (reflexive) Hilbert space setting with the well-known notions of the same name in Euclidean space (see [18, section 26]).

The next three results extend corresponding results in [3, section 6] considerably.

COROLLARY 7.7 (essential smoothness). *Suppose that at least one function f_i is essentially smooth, and that $\lambda_i > 0$. Then $p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})$ is essentially smooth.*

Proof. Since f_i is essentially smooth, the set $\operatorname{dom} f_i$ has a nonempty interior. Thus $\lambda_i \operatorname{dom} f_i$ and $\operatorname{dom} p_\mu(\boldsymbol{f}, \boldsymbol{\lambda}) = \lambda_1 \operatorname{dom} f_1 + \cdots + \lambda_n \operatorname{dom} f_n$ (see Theorem 4.6) both have nonempty interiors as well. Now take $x \in \operatorname{dom} p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})$, and let $y_1, \ldots, y_n$ be as in Theorem 4.10, say, $(\forall i)$ $y_i = \lambda_i x_i$, where $x_i \in \operatorname{dom} f_i$. By Theorem 7.1(i), $\partial p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})(x) \subseteq -\mu^{-1}x + \partial f_i(x_i) + \mu^{-1}x_i$. Because f_i is essentially smooth, the set $\partial f_i(x_i)$ is either empty or singleton. Thus $\partial p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})(x)$ is either empty or singleton. Altogether, $p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})$ is essentially smooth. ∎

COROLLARY 7.8 (essential strict convexity). *Suppose that at least one function f_i is essentially strictly convex, and that $\lambda_i > 0$. Then $p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})$ is essentially strictly convex.*

Proof. Since f_i is essentially strictly convex, its conjugate f_i^* is essentially smooth. By Corollary 7.7, $p_{\mu^{-1}}(\boldsymbol{f}^*, \boldsymbol{\lambda})$ is essentially smooth. Hence $(p_{\mu^{-1}}(\boldsymbol{f}^*, \boldsymbol{\lambda}))^*$ is essentially strictly convex. This last function is equal to $p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})$ (by Theorem 5.1), and the proof is thus complete. ∎

COROLLARY 7.9 (Legendre function). *Suppose that at least one function f_i is essentially smooth, and that $\lambda_i > 0$. Furthermore, suppose that at least one function f_j is essentially strictly convex, and that $\lambda_j > 0$. (It does not matter whether j and i are identical or distinct.) Then $p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})$ is both essentially smooth and essentially strictly convex, i.e., Legendre.*

Proof. Combine Corollary 7.7 and Corollary 7.8. ∎

Before we formulate and prove the last result in this section, we briefly return to the Moreau envelope and the proximal mapping. Let $f \in \Gamma(X)$. Applying Proposition 3.3(v) to μf, we readily deduce that (see also [19, Example 11.26(b)])

$$\mu(e_\mu f) + \mu \star (e_{\mu^{-1}}(f^*)) = \mathfrak{q}. \tag{65}$$

Taking gradients and recalling (47) yields $\operatorname{Id} = P_\mu f + \mu(P_{\mu^{-1}}(f^*)) \circ (\mu^{-1}\operatorname{Id})$; equivalently, $\mu \operatorname{Id} = (P_\mu f) \circ (\mu \operatorname{Id}) + \mu P_{\mu^{-1}}(f^*)$, or

$$\operatorname{Id} = \mu^{-1}(P_\mu f) \circ (\mu \operatorname{Id}) + P_{\mu^{-1}}(f^*). \tag{66}$$

The following result generalizes [4, Theorem 4.22] where $n = 2$, $\lambda_1 = \lambda_2 = \frac{1}{2}$, and $\mu = 1$.

THEOREM 7.10. *Suppose that* $(a, a^*) \in X \times X$ *satisfies* $a^* \in \partial f_1(a) \cap \cdots \cap \partial f_n(a)$, *and that* $\{1, 2, \ldots, n\}$ *is the disjoint union of two sets of indices* I *and* J. *Set* $\lambda_J = \sum_{j\in J} \lambda_j$, *and suppose that* $\lambda_J > 0$. *Then for every* $z \in a + (\bigcap_{i\in I} N_{\operatorname{dom} f_i}(a) \cap \bigcap_{j\in J} N_{\operatorname{dom} f_j^*}(a^*))$, *we have*

$$a^* + \mu^{-1}(\lambda_J^{-1} - 1)(z - a) \in \partial p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})(z). \tag{67}$$

Consequently, $p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})$ *is differentiable on* $a + \operatorname{int}(\bigcap_{i\in I} N_{\operatorname{dom} f_i}(a) \cap \bigcap_{j\in J} N_{\operatorname{dom} f_j^*}(a^*))$, *with gradient* $z \mapsto a^* + \mu^{-1}(\lambda_J^{-1} - 1)(z - a)$.

Proof. Let z be as in the conclusion, and set $y = z - a$. Fix $i \in I$. Now $a^* \in \partial f_i(a)$ and $\lambda_J^{-1} y \in N_{\operatorname{dom} f_i}(a) = \partial \iota_{\operatorname{dom} f_i}(a) = \partial \iota_{\operatorname{dom} \mu f_i}(a)$. Hence $\mu a^* \in \mu \partial f_i(a) = \partial(\mu f_i)(a)$. Thus $\mu a^* + \lambda_J^{-1} y \in \partial(\mu f_i)(a) + \partial(\iota_{\operatorname{dom} \mu f_i})(a) \subseteq \partial(\mu f_i + \iota_{\operatorname{dom} \mu f_i})(a) = \partial(\mu f_i)(a)$. It follows that

$$(\forall i \in I) \quad a = (P_\mu f_i)(\mu a^* + \lambda_J^{-1} y + a). \tag{68}$$

Next, fix $j \in J$. Then $a + \lambda_J^{-1} y \in \partial f_j^*(a^*)$, and $\mu^{-1} a + \mu^{-1} \lambda_J^{-1} y \in \partial(\mu^{-1} f_j^*)(a^*)$. Using (66), we thus have $a^* = (P_{\mu^{-1}} f_j^*)(\mu^{-1} a + \mu^{-1} \lambda_J^{-1} y + a^*) = \mu^{-1} a + \mu^{-1} \lambda_J^{-1} y + a^* - \mu^{-1}(P_\mu f_j)(a + \lambda_J^{-1} y + \mu a^*)$. Hence

$$(\forall j \in J) \quad a + \lambda_J^{-1} y = (P_\mu f_j)(a + \lambda_J^{-1} y + \mu a^*). \tag{69}$$

Now (68), (69), and Theorem 6.7 imply that

$$a + y = (P_\mu p_\mu(\boldsymbol{f}, \boldsymbol{\lambda}))(a + \lambda_J^{-1} y + \mu a^*), \tag{70}$$

equivalently,

$$a^* + \mu^{-1}(\lambda_J^{-1} - 1) y \in \partial p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})(a + y). \tag{71}$$

This verifies (67). Denote the intersection of the n normal cones by N. On $a + \operatorname{int} N$, the mapping $z \mapsto a^* + \mu^{-1}(\lambda_J^{-1} - 1)(z - a)$ is thus a continuous selection of $\partial p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})$; therefore, $\nabla p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})(z) = a^* + \mu^{-1}(\lambda_J^{-1} - 1)(z - a)$ by [17, Proposition 2.8]. □

8. Pointwise limits of the proximal average.

PROPOSITION 8.1. *Let* $f \in \Gamma(X)$. *Then* $e_{\mu^{-1}}\big(f \circ (\mu \operatorname{Id})\big) = (e_\mu f) \circ (\mu \operatorname{Id})$.

Proof. For every $x \in X$, we have $e_{\mu^{-1}}\big(f \circ (\mu \operatorname{Id})\big)(x) = \inf_y \big(f(\mu y) + \mu\, \mathfrak{q}(x - y)\big) = \inf_y \big(f(\mu y) + \mu^{-1}\, \mathfrak{q}(\mu x - \mu y)\big) = \inf_z \big(f(z) + \mu^{-1}\, \mathfrak{q}(\mu x - z)\big) = e_\mu f(\mu x)$. □

PROPOSITION 8.2 ([19, Example 11.26(c)]). *Let* $f\colon X \to [-\infty, +\infty]$. *Then*

$$(f + \mu\, \mathfrak{q})^* = (\mu\, \mathfrak{q} - e_{\mu^{-1}} f) \circ (\mu^{-1} \operatorname{Id}). \tag{72}$$

Proof. For every $x^* \in X$, we obtain that

$$\begin{aligned}
(f + \mu\, \mathfrak{q})^*(x^*) &= \sup_x \big(\langle x, x^*\rangle - f(x) - \mu\, \mathfrak{q}(x)\big) \\
&= \sup_x \big(\langle x, x^*\rangle - f(x) - \mu\, \mathfrak{q}(x - \mu^{-1} x^*) + \mu^{-1}\, \mathfrak{q}(x^*) - \langle x, x^*\rangle\big) \\
&= \mu^{-1}\, \mathfrak{q}(x^*) + \sup_x \big(- f(x) - \mu\, \mathfrak{q}(x - \mu^{-1} x^*)\big) \\
&= \mu^{-1}\, \mathfrak{q}(x^*) - \inf_x \big(f(x) + \mu\, \mathfrak{q}(\mu^{-1} x^* - x)\big) \\
&= \mu^{-1}\, \mathfrak{q}(x^*) - (f \mathbin{\blacktriangledown} \mu\, \mathfrak{q})(\mu^{-1} x^*) \\
&= \mu\, \mathfrak{q}(\mu^{-1} x^*) - (f \mathbin{\blacktriangledown} \mu^{-1} \star \mathfrak{q})(\mu^{-1} x^*) \\
&= \big(\mu\, \mathfrak{q} - e_{\mu^{-1}} f\big)(\mu^{-1} x^*). \qquad (73)
\end{aligned}$$

The result follows. □

The following alternative expression of the proximal average was discovered by Hare for the case when $n = 2$ and $\mu = 1$.

THEOREM 8.3 (see [7]). $p_\mu(\boldsymbol{f}, \boldsymbol{\lambda}) = -e_\mu\big(-(\lambda_1 e_\mu f_1 + \cdots + \lambda_n e_\mu f_n)\big)$.

Proof. Set $g = -(\lambda_1 e_\mu f_1 + \cdots + \lambda_n e_\mu f_n)$. Taking the Fenchel conjugate on both sides of (33) leads to $p_\mu(\boldsymbol{f}, \boldsymbol{\lambda}) = \big(\lambda_1(f_1^* \mathbin{\blacktriangleleft} \mu\,\mathfrak{q}) + \cdots + \lambda_n(f_n^* \mathbin{\blacktriangleleft} \mu\,\mathfrak{q})\big)^* - \mu \star \mathfrak{q}$. On the other hand, $(\forall i)$ $f_i^* \mathbin{\blacktriangleleft} \mu\,\mathfrak{q} = (f_i + \mu \star \mathfrak{q})^*$ by Fact 3.4(i) and Proposition 3.3(iii). Altogether,

$$p_\mu(\boldsymbol{f}, \boldsymbol{\lambda}) = \big(\lambda_1(f_1 + \mu \star \mathfrak{q})^* + \cdots + \lambda_n(f_n + \mu \star \mathfrak{q})^*\big)^* - \mu \star \mathfrak{q}. \tag{74}$$

Using (74), Proposition 3.3(ii), Proposition 8.2, and Proposition 8.1 we deduce that

$$\begin{aligned}
p_\mu(\boldsymbol{f}, \boldsymbol{\lambda}) &= \big(\lambda_1(f_1 + \mu \star \mathfrak{q})^* + \cdots + \lambda_n(f_n + \mu \star \mathfrak{q})^*\big)^* - \mu \star \mathfrak{q} \\
&= \big(\lambda_1(f_1 + \mu^{-1}\,\mathfrak{q})^* + \cdots + \lambda_n(f_n + \mu^{-1}\,\mathfrak{q})^*\big)^* - \mu \star \mathfrak{q}. \\
&= \big(\lambda_1(\mu^{-1}\,\mathfrak{q} - e_\mu f_1) \circ (\mu\,\mathrm{Id}) + \cdots + \lambda_n(\mu^{-1}\,\mathfrak{q} - e_\mu f_n) \circ (\mu\,\mathrm{Id})\big)^* - \mu \star \mathfrak{q} \\
&= \big(\mu\,\mathfrak{q} + g \circ (\mu\,\mathrm{Id})\big)^* - \mu \star \mathfrak{q} \\
&= \big(\mu\,\mathfrak{q} - e_{\mu^{-1}}(g \circ (\mu\,\mathrm{Id}))\big) \circ (\mu^{-1}\,\mathrm{Id}) - \mu \star \mathfrak{q} \\
&= \mu^{-1}\,\mathfrak{q} - \big(e_{\mu^{-1}}(g \circ (\mu\,\mathrm{Id}))\big) \circ (\mu^{-1}\,\mathrm{Id}) - \mu \star \mathfrak{q} \\
&= -\big((e_\mu g) \circ (\mu\,\mathrm{Id})\big) \circ (\mu^{-1}\,\mathrm{Id}) \\
&= -e_\mu g. \qquad (75)
\end{aligned}$$

This verifies the result. □

The *μ-proximal hull* of a function g is defined by $h_\mu g = -e_\mu(-e_\mu g)$; it satisfies $e_\mu g \le h_\mu g \le g$ and $e_\mu(h_\mu g) = e_\mu g$ (see [19, Example 1.44]). Theorem 8.3 shows that $p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})$ can be interpreted as some sort of weighted proximal hull of the functions $f_1, \ldots, f_n$. We now turn to the proximal hull of $p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})$.

COROLLARY 8.4 (proximal hull). $h_\mu p_\mu(\boldsymbol{f}, \boldsymbol{\lambda}) = p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})$.

Proof. By Theorem 6.2(i), $e_\mu p_\mu(\boldsymbol{f}, \boldsymbol{\lambda}) = \lambda_1 e_\mu f_1 + \cdots + \lambda_n e_\mu f_n$. Hence, using Theorem 8.3, $h_\mu(p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})) = -e_\mu(-e_\mu p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})) = -e_\mu(-\lambda_1 e_\mu f_1 - \cdots - \lambda_n e_\mu f_n) = p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})$. Since $p_\mu(\boldsymbol{f}, \boldsymbol{\lambda}) + \mu \star \mathfrak{q}$ is clearly convex and lower semicontinuous (by Corollary 5.2), the result follows alternatively from [19, Example 11.26(d)]. □

Let us now determine the pointwise behavior of $p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})$.

THEOREM 8.5 (pointwise limits). *Let $x \in X$. Then the function*

$$]0, +\infty[\to \,]-\infty, +\infty] : \mu \mapsto p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})(x) \quad \textit{is decreasing.} \tag{76}$$

Consequently, $\lim_{\mu \to 0^+} p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})(x)$ *and* $\lim_{\mu \to +\infty} p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})(x)$ *exist. In fact,*

$$\lim_{\mu \to 0^+} p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})(x) = \sup_{\mu > 0} p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})(x) = \big(\lambda_1 f_1 + \cdots + \lambda_n f_n\big)(x) \tag{77}$$

and

$$\lim_{\mu \to +\infty} p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})(x) = \inf_{\mu > 0} p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})(x) = \big(\lambda_1 \star f_1 \mathbin{\blacktriangleleft} \cdots \mathbin{\blacktriangleleft} \lambda_n \star f_n\big)(x). \tag{78}$$

Proof. The fact that $\mu \mapsto p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})(x)$ is decreasing follows from (20); consequently, the two limits exist and the supremum/infimum descriptions are clear. Now $e_\mu\big(-(\lambda_1 e_\mu f_1 + \cdots + \lambda_n e_\mu f_n)\big) \le -(\lambda_1 e_\mu f_1 + \cdots + \lambda_n e_\mu f_n)$. Thus, using Theorem 8.3,

we deduce that $\lambda_1 e_\mu f_1 + \cdots + \lambda_n e_\mu f_n \leq -e_\mu\big(-(\lambda_1 e_\mu f_1 + \cdots + \lambda_n e_\mu f_n)\big) = p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})$. On the other hand, Theorem 5.4 implies that $p_\mu(\boldsymbol{f}, \boldsymbol{\lambda}) \leq \lambda_1 f_1 + \cdots + \lambda_n f_n$. Altogether,

$$\lambda_1 e_\mu f_1 + \cdots + \lambda_n e_\mu f_n \leq p_\mu(\boldsymbol{f}, \boldsymbol{\lambda}) \leq \lambda_1 f_1 + \cdots + \lambda_n f_n. \tag{79}$$

It is well known that Moreau envelopes converge pointwise to the underlying function as the parameter approaches 0; see, e.g., [1, Theorem 2.64] or [19, Theorem 1.25 and Theorem 2.26]. Thus $(\forall i)\ \lim_{\mu\to 0^+} e_\mu f_i = f_i$ pointwise, and (77) follows from taking the pointwise limit in (79) at x as $\mu \to 0^+$. Using (20), we deduce that

$$\begin{aligned}
\lim_{\mu\to+\infty} p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})(x) &= \inf_{\mu>0} p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})(x) \\
&= \inf_{\mu>0} \inf_{\sum \lambda_i y_i = x} \sum \lambda_i f_i(y_i) + \frac{1}{\mu}\bigg(\Big(\sum \lambda_i\, \mathfrak{q}(y_i)\Big) - \mathfrak{q}(x)\bigg) \\
&= \inf_{\sum \lambda_i y_i = x} \inf_{\mu>0} \sum \lambda_i f_i(y_i) + \frac{1}{\mu}\bigg(\Big(\sum \lambda_i\, \mathfrak{q}(y_i)\Big) - \mathfrak{q}(x)\bigg) \\
&= \inf_{\sum \lambda_i y_i = x} \sum \lambda_i f_i(y_i) \\
&= \inf_{\sum' x_i = x} \sum{}' \lambda_i f_i(x_i/\lambda_i) \\
&= \inf_{\sum' x_i = x} \sum{}' (\lambda_i \mathbin{\triangleright} f_i)(x_i) \\
&= \big(\lambda_1 \mathbin{\triangleright} f_1 \mathbin{\blacktriangleleft} \cdots \mathbin{\blacktriangleleft} \lambda_n \mathbin{\triangleright} f_n\big)(x), \qquad (80)
\end{aligned}$$

where the indices in the $\sum'$ sums range over all i such that $\lambda_i > 0$. □

The following nice observation, which is based on the comments of an anonymous referee, builds a bridge to [15].

Remark 8.6 (parallel sums). Suppose that $X = \mathbb{R}^N$, let $A_1, \ldots, A_n$ be positive definite $N \times N$ matrices, and suppose that $(\forall i)\ f_i(x) = \frac{1}{2}\langle x, A_i x\rangle$, i.e., identify each A_i with its quadratic form. As $\mu \to 0^+$, $p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})$ converges pointwise to $\lambda_1 f_1 + \cdots + \lambda_n f_n$, and as $\mu \to +\infty$, $p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})$ converges pointwise to $\lambda_1 \mathbin{\triangleright} f_1 \mathbin{\blacktriangleleft} \cdots \mathbin{\blacktriangleleft} \lambda_n \mathbin{\triangleright} f_n$. Using [15] (see also [10, Example IV.2.3.8], [12], and [14]), the matrices corresponding to the quadratic forms $\lambda_1 f_1 + \cdots + \lambda_n f_n$, $\lambda_1 \mathbin{\triangleright} f_1 \mathbin{\blacktriangleleft} \cdots \mathbin{\blacktriangleleft} \lambda_n \mathbin{\triangleright} f_n$, and $p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})$ are, respectively, the arithmetic average $\lambda_1 A_1 + \cdots + \lambda_n A_n$; the harmonic average $(\lambda_1 A_1^{-1} + \cdots + \lambda_n A_n^{-1})^{-1}$, i.e., the *parallel sum* of the matrices $\lambda_1^{-1} A_1, \ldots, \lambda_n^{-1} A_n$; and $(\lambda_1 (A_1 + \mu^{-1}\,\mathrm{Id})^{-1} + \cdots + \lambda_n (A_n + \mu^{-1}\,\mathrm{Id})^{-1})^{-1} - \mu^{-1}\,\mathrm{Id}$, i.e., a μ^{-1}-shifted version of the harmonic average (in accordance with the comment before Definition 4.1). Note that this provides another proof of Example 4.5, and that the theory for parallel sum extends to matrices that are only positive semidefinite.

9. Epi-continuity and epi-limits of the proximal average. We now discuss the convergence behavior of the proximal average with respect to the epi-topology. Analogously to [3, section 5], we assume throughout this section that

$$X \text{ is finite-dimensional.} \tag{81}$$

Definition 9.1 (epi-convergence and epi-topology). *(See* [19, *Chapter* 6]*.) Let* g *and* $(g_k)_{k\in\mathbb{N}}$ *be functions from* X *to* $]-\infty, +\infty]$. *Then* $(g_k)_{k\in\mathbb{N}}$ epi-converges *to* g, *in symbols* $g_k \xrightarrow{\mathrm{e}} g$, *if the following hold for every* $x \in X$.

(i) $\big(\forall\, (x_k)_{k\in\mathbb{N}}\big)\ x_k \to x \Rightarrow g(x) \leq \underline{\lim}\, g_k(x_k)$.

(ii) $\big(\exists\, (y_k)_{k\in\mathbb{N}}\big)\ y_k \to x$ *and* $\overline{\lim}\, g_k(y_k) \leq g(x)$.

The epi-topology *is the topology induced by epi-convergence.*

FACT 9.2. *Let g and $(g_k)_{k\in\mathbb{N}}$ be in $\Gamma(X)$ such that $g_k \xrightarrow{\text{e}} g$, and let h and $(h_k)_{k\in\mathbb{N}}$ be in $\Gamma(X)$ such that $h_k \xrightarrow{\text{e}} h$. Let ρ and $(\rho_k)_{k\in\mathbb{N}}$ be in $[0,+\infty[$ such that $\rho_k \to \rho$, and let $q\colon X \to \mathbb{R}$ be continuous. Then the following hold.*

(i) $g_k \pm q \xrightarrow{\text{e}} g \pm q$.
(ii) $\rho > 0 \Rightarrow \rho_k g_k \xrightarrow{\text{e}} \rho g$.
(iii) $\rho = 0$ *and* $\operatorname{dom} g = X \Rightarrow \rho_k g_k \xrightarrow{\text{e}} \rho g$.
(iv) $g_k^* \xrightarrow{\text{e}} g^*$.
(v) $0 \in \operatorname{int}(\operatorname{dom} g - \operatorname{dom} h) \Rightarrow g_k + h_k \xrightarrow{\text{e}} g + h$.

Proof. (i): See [19, Exercise 7.8(a)]. (ii): See [19, Exercise 7.8(d)]. (iii): See [3] or verify this directly. (iv): See [19, Theorem 11.34]. (v): See [19, Exercise 7.47(b)]. □

LEMMA 9.3. *Let $g_1,\ldots,g_n,h$ be in $\Gamma(X)$, and let $(g_{1,k})_{k\in\mathbb{N}},\ldots,(g_{n,k})_{k\in\mathbb{N}},(h_k)_{k\in\mathbb{N}}$ be sequences in $\Gamma(X)$ such that $(\forall i)$ $g_{i,k} \xrightarrow{\text{e}} g_i$ and $h_k \xrightarrow{\text{e}} h$. Let ρ and $(\rho_k)_{k\in\mathbb{N}}$ be in $[0,+\infty[$ such that $\rho_k \to \rho$. Suppose that $\operatorname{dom} g_1^* = \cdots = \operatorname{dom} g_{n-1}^* = \operatorname{dom} h^* = X$ and that $(\forall i \in \{1,\ldots,n-1\})(\forall k)$ $\operatorname{dom} g_{i,k}^* = X$. Then the following hold.*

(i) $g_{1,k} \mathbin{\blacktriangleleft} \cdots \mathbin{\blacktriangleleft} g_{n,k} \xrightarrow{\text{e}} g_1 \mathbin{\blacktriangleleft} \cdots \mathbin{\blacktriangleleft} g_n$.
(ii) $\rho_k \star h_k \xrightarrow{\text{e}} \rho \star h$.

Proof. (i): Fact 9.2(iv)&(v) imply that $g_{1,k}^* + \cdots + g_{n,k}^* \xrightarrow{\text{e}} g_1^* + \cdots + g_n^*$. Using Fact 9.2(iv), we see that $(g_{1,k}^* + \cdots + g_{n,k}^*)^* \xrightarrow{\text{e}} (g_1^* + \cdots + g_n^*)^*$, which is equivalent to $g_{1,k} \mathbin{\blacktriangleleft} \cdots \mathbin{\blacktriangleleft} g_{n,k} \xrightarrow{\text{e}} g_1 \mathbin{\blacktriangleleft} \cdots \mathbin{\blacktriangleleft} g_n$ by Fact 3.4(i). (ii): Fact 9.2(ii)–(iv) imply that $\rho_k h_k^* \xrightarrow{\text{e}} \rho h^*$. Using Fact 9.2(iv) once more, we deduce that $(\rho_k h_k^*)^* \xrightarrow{\text{e}} (\rho h^*)^*$, which is the same as the conclusion in view of Proposition 3.2(i). □

Remark 9.4. Using the horizon functions associated with $g_1,\ldots,g_n$ and [19, Proposition 7.56], one may obtain a stronger version of Lemma 9.3 where the assumption on the functions $g_{i,k}^*$ is less restrictive; however, this is not needed in the sequel.

The next result extends [3, Theorem 5.4].

THEOREM 9.5 (epi-continuity of the proximal average). *Let $(f_{i,k})_{k\in\mathbb{N}}$ be sequences in $\Gamma(X)$ such that $(\forall i)$ $f_{i,k} \xrightarrow{\text{e}} f_i$, let $(\lambda_{i,k})_{k\in\mathbb{N}}$ be sequences in $[0,1]$ such that $(\forall k)$ $\sum_i \lambda_{i,k} = 1$ and $(\forall i)$ $\lambda_{i,k} \to \lambda_i$, and let $(\mu_k)_{k\in\mathbb{N}}$ be a sequence in $]0,+\infty[$ such that $\mu_k \to \mu$. Then*

(82)
$$p_{\mu_k}\big((f_{1,k},\ldots,f_{n,k}),(\lambda_{1,k},\ldots,\lambda_{n,k})\big) \xrightarrow{\text{e}} p_\mu\big((f_1,\ldots,f_n),(\lambda_1,\ldots,\lambda_n)\big) = p_\mu(\boldsymbol{f},\boldsymbol{\lambda}).$$

Proof. By Theorem 9.3(ii),

$$\mu_k \star \mathfrak{q} \xrightarrow{\text{e}} \mu \star \mathfrak{q}. \tag{83}$$

Furthermore,

$$(\forall i) \quad f_{i,k} + \mu_k \star \mathfrak{q} \xrightarrow{\text{e}} f_i + \mu \star \mathfrak{q} \tag{84}$$

by Fact 9.2(v) because $(\mu \star \mathfrak{q})^* = \mu\,\mathfrak{q}$ has full domain. Using (84), Lemma 9.3(ii), and the fact that $(\forall i)$ $(f_i + \mu \star \mathfrak{q})^* = (f_i^* \mathbin{\blacktriangleleft} (\mu \star \mathfrak{q})^*)^{**} = (f_i^* \mathbin{\blacktriangleleft} \mu\,\mathfrak{q})^{**}$ has full domain (and similarly for $(f_{i,k} + \mu_k \star \mathfrak{q})^*$), we deduce that

$$(\forall i) \quad \lambda_{i,k} \star (f_{i,k} + \mu_k \star \mathfrak{q}) \xrightarrow{\text{e}} \lambda_i \star (f_i + \mu \star \mathfrak{q}). \tag{85}$$

Since $(\forall i)\ \big(\lambda_i \mathbin{\star}(f_i+\mu \mathbin{\star} \mathfrak{q})\big)^* = \lambda_i(f_i+\mu \mathbin{\star} \mathfrak{q})^* = \lambda_i(f_i^* \mathbin{\#} \mu \mathbin{\star} \mathfrak{q})$ has full domain (and similarly for $(\lambda_{i,k} \mathbin{\star}(f_{i,k}+\mu_k \mathbin{\star} \mathfrak{q}))^*$), (85) and Lemma 9.3(i) yield

(86)
$$\lambda_{1,k} \mathbin{\star}(f_{1,k}+\mu_k \mathbin{\star} \mathfrak{q}) \mathbin{\#} \cdots \mathbin{\#} \lambda_{n,k} \mathbin{\star}(f_{n,k}+\mu_k \mathbin{\star} \mathfrak{q}) \xrightarrow{\mathrm{e}} \lambda_1 \mathbin{\star}(f_1+\mu \mathbin{\star} \mathfrak{q}) \mathbin{\#} \cdots \mathbin{\#} \lambda_n \mathbin{\star}(f_n+\mu \mathbin{\star} \mathfrak{q}).$$

In turn, (83), (86), and Fact 9.2(i) imply (82). □

We now describe the behavior of $p_\mu(\boldsymbol{f},\boldsymbol{\lambda})$ when μ approaches either 0 or $+\infty$ while $\boldsymbol{f}$ and $\boldsymbol{\lambda}$ are fixed.

COROLLARY 9.6. *$p_\mu(\boldsymbol{f},\boldsymbol{\lambda}) \xrightarrow{\mathrm{e}} \lambda_1 f_1 + \cdots + \lambda_n f_n$ as $\mu \to 0^+$, and $p_\mu(\boldsymbol{f},\boldsymbol{\lambda}) \xrightarrow{\mathrm{e}} \mathrm{cl}(\lambda_1 \mathbin{\star} f_1 \mathbin{\#} \cdots \mathbin{\#} \lambda_n \mathbin{\star} f_n)$ as $\mu \to +\infty$.*

Proof. Theorem 8.5 shows that $\mu \mapsto p_\mu(\boldsymbol{f},\boldsymbol{\lambda})$ is pointwise increasing. In view of (77) and the lower semicontinuity of $p_\mu(\boldsymbol{f},\boldsymbol{\lambda})$ (see Corollary 5.2), an application of [19, Proposition 7.4(d)] yields that $p_\mu(\boldsymbol{f},\boldsymbol{\lambda}) \xrightarrow{\mathrm{e}} \lambda_1 f_1 + \cdots + \lambda_n f_n$ as $\mu \to 0^+$. Combining (78) with [19, Proposition 7.4(e)], we deduce similarly that $p_\mu(\boldsymbol{f},\boldsymbol{\lambda}) \xrightarrow{\mathrm{e}} \mathrm{cl}(\lambda_1 \mathbin{\star} f_1 \mathbin{\#} \cdots \mathbin{\#} \lambda_n \mathbin{\star} f_n)$ as $\mu \to +\infty$. □

Corollary 9.6 and (77) show that, as $\mu \to 0^+$, the pointwise and epigraphical limits of $p_\mu(\boldsymbol{f},\boldsymbol{\lambda})$ coincide. When $\mu \to +\infty$, the pointwise and epigraphical limits of $p_\mu(\boldsymbol{f},\boldsymbol{\lambda})$ may differ as we illustrate next.

Example 9.7. Suppose that $X = \mathbb{R}^2$, that $n = 2$, that $\lambda_1 > 0$, that $\lambda_2 > 0$, that $f_1 = \iota_{C_1}$, and that $f_2 = \iota_{C_2}$ where C_1 and C_2 are nonempty closed convex subsets of X such that $\lambda_1 C_1 + \lambda_2 C_2$ is not closed. Concretely, we may let C_1 and C_2 be the epigraphs of $x \mapsto \exp(x)$ and $x \mapsto \exp(-x)$, respectively. Then the pointwise limit (see (78))

$$\lim_{\mu \to +\infty} p_\mu(\boldsymbol{f},\boldsymbol{\lambda}) = \lambda_1 \mathbin{\star} f_1 \mathbin{\#} \lambda_2 \mathbin{\star} f_2 = \iota_{\lambda_1 C_1 + \lambda_2 C_2} \tag{87}$$

is not lower semicontinuous, and hence different from the epigraphical limit (see Corollary 9.6) $\mathrm{cl}(\lambda_1 \mathbin{\star} f_1 \mathbin{\#} \lambda_2 \mathbin{\star} f_2)$, which is the indicator function of the closure of $\lambda_1 C_1 + \lambda_2 C_2$.

We now show that the limiting behavior as $\mu \to +\infty$ cannot be obtained by conjugation.

Example 9.8. Suppose that $X = \mathbb{R}^2$, that $n = 2$, that $f_1 \colon (x,y) \mapsto -x + \iota_{\{0\}}(y)$, that $f_2 \colon (x,y) \mapsto x + \iota_{\{0\}}(y)$, that $\lambda_1 > 0$, and that $\lambda_2 > 0$. Now fix $(x,y) \in \mathbb{R}^2$. Using (16) and some calculus, we calculate

$$p_\mu(\boldsymbol{f},\boldsymbol{\lambda})(x,y) = (\lambda_2-\lambda_1)x + \iota_{\{0\}}(y) - 2\mu\lambda_1\lambda_2 = (\lambda_1 f_1 + \lambda_2 f_2)(x,y) - 2\mu\lambda_1\lambda_2. \tag{88}$$

Letting $\mu \to 0^+$ in (88) and in accordance with (77), we observe that $p_\mu(\boldsymbol{f},\boldsymbol{\lambda}) \to \lambda_1 f_1 + \lambda_2 f_2$ pointwise. Recalling (78) and letting $\mu \to +\infty$ in (88), we see that

$$(\lambda_1 \mathbin{\star} f_1 \mathbin{\#} \lambda_2 \mathbin{\star} f_2)(x,y) = \lim_{\mu \to +\infty} p_\mu(\boldsymbol{f},\boldsymbol{\lambda})(x,y) = \begin{cases} -\infty & \text{if } y = 0; \\ +\infty & \text{if } y \neq 0. \end{cases} \tag{89}$$

Since $f_1^*(x,y) = \iota_{\{-1\}}(x)$ and $f_2^*(x,y) = \iota_{\{1\}}(x)$, we have $\mathrm{dom}(f_1^*) \cap \mathrm{dom}(f_2^*) = \varnothing$, and thus $\lambda_1 f_1^* + \lambda_2 f_2^* \equiv +\infty$. Altogether,

$$\lambda_1 \mathbin{\star} f_1 \mathbin{\#} \lambda_2 \mathbin{\star} f_2 \neq \big(\lambda_1 f_1^* + \lambda_2 f_2^*\big)^* \equiv -\infty. \tag{90}$$

Therefore, due to the absence of a constraint qualification on f_1^* and f_2^*, the epigraphical convergence of $p_\mu(\boldsymbol{f},\boldsymbol{\lambda})$ to the epigraphical average of f_1 and f_2 as $\mu \to$

$+\infty$ could not have been obtained by conjugating the epigraphical convergence of $p_{\mu^{-1}}(f_1^*, f_2^*, \lambda_1, \lambda_2)$ to $\lambda_1 f_1^* + \lambda_2 f_2^*$ as $\mu \to +\infty$.

In the presence of a constraint qualification, we can use the proximal average to construct a homotopic curve with very nice properties.

Remark 9.9 (epigraphical and arithmetic averages are homotopic). Suppose that $\operatorname{int}\operatorname{dom} f_1^* \cap \cdots \cap \operatorname{int}\operatorname{dom} f_{n-1}^* \cap \operatorname{dom} f_n^* \neq \varnothing$. By Fact 3.4(i) and Proposition 3.2(i), we have $(\lambda_1 f_1^* + \cdots + \lambda_n f_n^*)^* = \lambda_1 \mathbin{\star} f_1 \mathbin{\#} \cdots \mathbin{\#} \lambda_n \mathbin{\star} f_n$. Therefore,

$$\operatorname{cl}(\lambda_1 \mathbin{\star} f_1 \mathbin{\#} \cdots \mathbin{\#} \lambda_n \mathbin{\star} f_n) = \lambda_1 \mathbin{\star} f_1 \mathbin{\#} \cdots \mathbin{\#} \lambda_n \mathbin{\star} f_n, \tag{91}$$

and hence the pointwise and epigraphical limits of $p_\mu(\boldsymbol{f}, \boldsymbol{\lambda})$ as either $\mu \to 0^+$ or $\mu \to +\infty$ coincide by Theorem 8.5 and Corollary 9.6. Now set

$$(\forall \rho \in [0,1]) \quad q_\rho \colon x \mapsto \begin{cases} (\lambda_1 f_1 + \cdots + \lambda_n f_n)(x) & \text{if } \rho = 0; \\ p_{\tan(\rho\pi/2)}(\boldsymbol{f}, \boldsymbol{\lambda})(x) & \text{if } 0 < \rho < 1; \\ (\lambda_1 \mathbin{\star} f_1 \mathbin{\#} \cdots \mathbin{\#} \lambda_n \mathbin{\star} f_n)(x) & \text{if } \rho = 1. \end{cases} \tag{92}$$

Then Theorem 8.5, Corollary 9.5, and Corollary 9.6 show that $(q_\rho)_{\rho\in[0,1]}$ is a decreasing, pointwise convergent, homotopic (with respect to the epi-topology) curve between the arithmetic average $\lambda_1 f_1 + \cdots + \lambda_n f_n$ and the epigraphical average $\lambda_1 \mathbin{\star} f_1 \mathbin{\#} \cdots \mathbin{\#} \lambda_n \mathbin{\star} f_n$.

Acknowledgments. We wish to thank Prof. J.-B. Hiriart-Urruty for lending us his copy of [14], and two anonymous referees for their comments.

REFERENCES

[1] H. Attouch, *Variational Convergence for Functions and Operators*, Pitman, Boston, MA, 1984.

[2] H. H. Bauschke, J. M. Borwein, and P. L. Combettes, *Essential smoothness, essential strict convexity, and Legendre functions in Banach spaces*, Commun. Contemp. Math., 3 (2001), pp. 615–647.

[3] H. H. Bauschke, Y. Lucet, and M. Trienis, *How to transform one convex function continuously into another*, SIAM Rev., 50 (2008), pp. 115–132.

[4] H. H. Bauschke, Y. Lucet, and X. Wang, *Primal-dual symmetric intrinsic methods for finding antiderivatives of cyclically monotone operators*, SIAM J. Control Optim., 46 (2007), pp. 2031–2051.

[5] H. H. Bauschke, E. Matoušková, and S. Reich, *Projection and proximal point methods: Convergence results and counterexamples*, Nonlinear Anal., 56 (2004), pp. 715–738.

[6] H. H. Bauschke and X. Wang, *The kernel average for two convex functions and its applications to the extension and representation of monotone operators*, Trans. Amer. Math. Soc., to appear.

[7] W. L. Hare, private communication, 2006.

[8] W. L. Hare, *The proximal average of nonconvex functions: A proximal stability perspective*, preprint, 2007.

[9] J.-B. Hiriart-Urruty, *A general formula on the conjugate of the difference of functions*, Canad. Math. Bull., 29 (1986), pp. 482–485.

[10] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* I, Springer-Verlag, New York, 1996.

[11] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* II, Springer-Verlag, New York, 1996.

[12] J.-B. Hiriart-Urruty and M.-L. Mazure, *Formulations variationnelles de l'addition parallèle et de la soustraction parallèlle d'opérateurs semi-définis positifs*, C. R. des Séances Acad. Sci., Ser. I. Math., 302 (1986), pp. 527–530.

[13] Y. Lucet, H. H. Bauschke, and M. Trienis, *The piecewise linear-quadratic model for computational convex analysis*, Comput. Optim. Appl., to appear.

[14] M.-L. Mazure, *Analyse variationnelle des formes quadratriques convexes*, Thèse de doctorat de l'Université Paul Sabatier, Toulouse, France, 1986.
[15] M.-L. Mazure, *L'addition parallèle d'opérateurs interprétée comme inf-convolution de formes quadratiques convexes*, RAIRO Mod. Math. Anal. Numer., 20 (1986), pp. 497–515.
[16] J. J. Moreau, *Proximité et dualité dans un espace hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.
[17] R. R. Phelps, *Convex Functions, Monotone Operators, and Differentiability*, 2nd ed., Lecture Notes in Math. 1364, Springer-Verlag, 1993.
[18] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.
[19] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, 1998.
[20] C. Zălinescu, *Convex Analysis in General Vector Spaces*, World Scientific, River Edge, NJ, 2002.

SIAM J. OPTIM.
Vol. 19, No. 2, pp. 786–807

ON THE BEHAVIOR OF SUBGRADIENT PROJECTIONS METHODS FOR CONVEX FEASIBILITY PROBLEMS IN EUCLIDEAN SPACES*

DAN BUTNARIU†, YAIR CENSOR†, PINI GURFIL‡, AND ETHAN HADAR§

Abstract. We study some methods of subgradient projections for solving a convex feasibility problem with general (not necessarily hyperplanes or half-spaces) convex sets in the inconsistent case and propose a strategy that controls the relaxation parameters in a specific self-adapting manner. This strategy leaves enough user flexibility but gives a mathematical guarantee for the algorithm's behavior in the inconsistent case. We present the numerical results of computational experiments that illustrate the computational advantage of the new method.

Key words. convex feasibility problems, projection method, computational algorithms

AMS subject classifications. 49M37, 90C25

DOI. 10.1137/070689127

1. Introduction.

In this paper we consider, in an Euclidean space framework, the method of simultaneous subgradient projections for solving a convex feasibility problem with general (not necessarily linear) convex sets in the consistent and inconsistent cases. To cope with this situation, we propose two algorithmic developments. One uses *steering parameters* instead of *relaxation parameters* in the simultaneous subgradient projection method, and the other is a strategy that controls the relaxation parameters in a specific self-adapting manner that leaves enough user flexibility while yielding some mathematical guarantees for the algorithm's behavior in the inconsistent case. For the algorithm that uses steering parameters there is currently no mathematical theory. We present the numerical results of computational experiments that show the computational advantage of the mathematically-founded algorithm implementing our specific relaxation strategy. In the remainder of this section we elaborate upon the meaning of the above-made statements.

Given m closed convex subsets $Q_1, Q_2, \ldots, Q_m \subseteq R^n$ of the n-dimensional Euclidean space, expressed as

$$Q_i = \{x \in R^n \mid f_i(x) \leq 0\}, \tag{1.1}$$

where $f_i : R^n \to R$ is a convex function, the *convex feasibility problem* (CFP) is

$$\text{find a point } x^* \in Q := \cap_{i=1}^m Q_i. \tag{1.2}$$

*Received by the editors April 23, 2007; accepted for publication (in revised form) March 6, 2008; published electronically July 3, 2008. This work was supported by grant 2003275 of the United States–Israel Binational Science Foundation, by National Institutes of Health grant HL70472, by The Technion–University of Haifa Joint Research Fund, and by grant 522/04 of the Israel Science Foundation at the Center for Computational Mathematics and Scientific Computation in the University of Haifa.

http://www.siam.org/journals/siopt/19-2/68912.html

†Department of Mathematics, University of Haifa, Mt. Carmel, Haifa 31905, Israel (dbutnaru@math.haifa.ac.il, yair@math.haifa.ac.il).

‡Faculty of Aerospace Engineering, Technion–Israel Institute of Technology, Technion City, Haifa 32000, Israel (pgurfil@aerodyne.technion.ac.il).

§CA Labs, Computer Associates International, Inc., Yokneam 20692, Israel (ethan.hadar@ca.com).

As is well known, if the sets are given in any other form, then they can be represented in the form (1.1) by choosing, for f_i, the squared Euclidean distance to the set. Thus, it is required to solve the system of convex inequalities

$$f_i(x) \leq 0, \quad i = 1, 2, \ldots, m. \tag{1.3}$$

A fundamental question is how to approach the CFP in the inconsistent case when $Q = \cap_{i=1}^{m} Q_i = \emptyset$. Logically, algorithms designed to solve the CFP by finding a point $x^* \in Q$ are bound to fail and should, therefore, not be employed. But this is not always the case. Projection methods that are commonly used for the CFP, particularly in some very large real-world applications (see details below), are applied to CFPs without prior knowledge whether or not the problem is consistent. In such circumstances it is imperative to know how would a method, that is originally known to converge for a consistent CFP, behave if consistency is not guaranteed.

We address this question for a particular type of projection methods. In general, sequential projection methods exhibit *cyclic convergence* in the inconsistent case. This means that the whole sequence of iterates does not converge, but it breaks up into m convergent subsequences (see Gubin, Polyak, and Raik [25, Theorem 2] and Bauschke, Borwein and Lewis [3]). In contrast, simultaneous projection methods generally converge, even in the inconsistent case, to a minimizer of a proximity function that "measures" the weighted sum of squared distances to all sets of the CFP provided such a minimizer exists (see Iusem and De Pierro [28] for a local convergence proof and Combettes [17] for a global one).

Therefore, there is an advantage in using simultaneous projection methods from the point of view of convergence. Additional advantages are that (i) they are inherently parallel already at the mathematical formulation level due to the simultaneous nature, and (ii) they allow the user to assign weights (of importance) to the sets of the CFP. However, a severe limitation, common to sequential as well as simultaneous projection methods, is the need to solve an inner-loop distance-minimization step for the calculation of the orthogonal projection onto each individual set of the CFP. This need is alleviated only for convex sets that are simple to project onto, such as hyperplanes or half-spaces.

A useful path to circumvent this limitation is to use subgradient projections that rely on the calculation of subgradients at the current (available) iteration points; see Censor and Lent [12] or [13, section 5.3]. Iusem and Moledo [32] studied the simultaneous projection method with subgradient projections but only for consistent CFPs. To the best of our knowledge, there does not exist a study of the simultaneous projection method with subgradient projections for the inconsistent case. Our present results are a contribution towards this goal.

The CFP is a fundamental problem in many areas of mathematics and the physical sciences; see, e.g., Combettes [16, 18] and references therein. It has been used to model significant real-world problems in image reconstruction from projections; see, e.g., Herman [26], in radiation therapy treatment planning; see Censor, Altschuler, and Powlis [11] and Censor [9], and in crystallography; see Marks, Sinkler and Landree [33], to name but a few, and has been used under additional names such as *set-theoretic estimation* or the *feasible set approach*. A common approach to such problems is to use projection algorithms; see, e.g., Bauschke and Borwein [2], which employ *orthogonal projections* (i.e., nearest-point mappings) onto the individual sets Q_i. The orthogonal projection $P_\Omega(z)$ of a point $z \in R^n$ onto a closed convex set $\Omega \subseteq R^n$ is defined by

$$P_\Omega(z) := \operatorname{argmin}\{\| z - x \| \mid x \in \Omega\}, \tag{1.4}$$

where, throughout this paper, $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ denote the Euclidean norm and inner product, respectively, in R^n. Frequently a *relaxation parameter* is introduced so that

$$P_{\Omega,\lambda}(z) := (1-\lambda)z + \lambda P_\Omega(z) \tag{1.5}$$

is the *relaxed projection* of z onto Ω with relaxation λ. Many iterative projection algorithms for the CFP were developed; see subsection 1.1 below.

1.1. Projection methods: Advantages and earlier work. The reason why the CFP is looked at from the viewpoint of projection methods can be appreciated by the following brief comments, that we made in earlier publications, regarding projection methods in general. Projections onto sets are used in a variety of methods in optimization theory, but not every method that uses projections really belongs to the class of projection methods. *Projection methods* are iterative algorithms which use projections onto sets. They rely on the general principle that projections onto the given individual sets are easier to perform than projections onto other sets derived from the given individual sets (intersections, image sets under some transformation, etc.).

A projection algorithm reaches its goal, related to the whole family of sets, by performing projections onto the individual sets. Projection algorithms employ projections onto convex sets in various ways. They may use different kinds of projections and, sometimes, even use different types of projections within the same algorithm. They serve to solve a variety of problems, which are of either the feasibility or of the optimization types. They have different algorithmic structures, of which some are particularly suitable for parallel computing, and they demonstrate nice convergence properties and/or good initial behavior patterns.

Apart from theoretical interest, the main advantage of projection methods, which makes them successful in real-world applications, is computational. They commonly have the ability to handle huge-size problems that are beyond the ability of more sophisticated and currently available methods. This is so because the building blocks of a projection algorithm are the projections onto the given individual sets (which are easy to perform), and the algorithmic structure is either sequential or simultaneous (or in-between).

The field of projection methods is vast, and we mention here only a few recent works that can give the reader some good starting points. Such a list includes, among many others, the works of Crombez [20, 21], the connection with variational inequalities; see, e.g., Noor [34] and Yamada [38], which is motivated by real-world problems of signal processing, and the many contributions of Bauschke and Combettes; see, e.g., Bauschke, Combettes, and Kruk [4] and references therein. Bauschke and Borwein [2] and Censor and Zenios [13, Chapter 5] provide reviews of the field.

Systems of linear equations, linear inequalities, or convex inequalities are all encompassed by the CFP, which has broad applicability in many areas of mathematics and the physical and engineering sciences. These include, among others, optimization theory (see, e.g., Eremin [24], Censor and Lent [12], and Chinneck [14]), approximation theory (see, e.g., Deutsch [22] and references therein), image reconstruction from projections in computerized tomography (see, e.g., Herman [26, 27]), and control theory (see, e.g., Boyd et al. [5].)

Combettes [19] and Kiwiel [31] have studied the subgradient projection method for consistent CFPs. Their work presents more general algorithmic steps and is formulated in Hilbert space. Some work has already been done on detecting infeasibility with certain subgradient projection methods by Kiwiel [29, 30]. However, our ap-

proach differs from the latter in that it aims at a subgradient projection method that "will work" regardless of the feasibility of the underlying CFP and which does not require the user to study in advance whether or not the CFP is consistent. Further questions arise such as that of combining our work, or the above quoted results, with Pierra's [36] product space formalism, as extended to handle inconsistent situations by Combettes [17]. These questions are currently under investigation.

2. Simultaneous subgradient projections with steering parameters. Subgradient projections have been incorporated in iterative algorithms for the solution of CFPs. The cyclic subgradient projections (CSP) method for the CFP was given by Censor and Lent [12] as follows.

ALGORITHM 2.1 (The CSP method).

Initialization: $x^0 \in R^n$ *is arbitrary.*

Iterative step: *Given* x^k, *calculate the next iterate* x^{k+1} *by*

$$x^{k+1} = \begin{cases} x^k - \alpha_k \dfrac{f_{i(k)}(x^k)}{\| t^k \|^2} t^k & \text{if } f_{i(k)}(x^k) > 0, \\ x^k & \text{if } f_{i(k)}(x^k) \leq 0, \end{cases} \tag{2.1}$$

where $t^k \in \partial f_{i(k)}(x^k)$ *is a subgradient of* $f_{i(k)}$ *at the point* x^k, *and the relaxation parameters* $\{\alpha_k\}_{k=0}^{\infty}$ *are confined to an interval* $\epsilon_1 \leq \alpha_k \leq 2 - \epsilon_2$, *for all* $k \geq 0$, *with some arbitrarily small* $\epsilon_1, \epsilon_2 > 0$.

Control: *Denoting* $I := \{1, 2, \ldots, m\}$, *the sequence* $\{i(k)\}_{k=0}^{\infty}$ *is an almost cyclic control sequence on* I. *This means (see, e.g.,* [13, *Definition* 5.1.1]*) that* $i(k) \in I$ *for all* $k \geq 0$, *and there exists an integer* $C \geq m$ *such that, for all* $k \geq 0$, $I \subseteq \{i(k+1), i(k+2), \ldots, i(k+C)\}$.

Observe that if $t^k = 0$, then $f_{i(k)}$ takes its minimal value at x^k, implying, by the nonemptiness of Q, that $f_{i(k)}(x^k) \leq 0$ so that $x^{k+1} = x^k$. The relations of the CSP method to other iterative methods for solving the convex feasibility problem and to the relaxation method for solving linear inequalities can be found, e.g., in [13, Chapter 5]; see also, Bauschke and Borwein [2, section 7]. Since sequential projection methods for CFPs commonly have fully simultaneous counterparts, the simultaneous subgradient projections (SSP) method of Dos Santos [23] and Iusem and Moledo [32] is a natural algorithmic development.

ALGORITHM 2.2 (The SSP method).

Initialization: $x^0 \in R^n$ *is arbitrary.*

Iterative step: (i) *Given* x^k, *calculate, for all* $i \in I = \{1, 2, \ldots, m\}$, *intermediate iterates* $y^{k+1,i}$ *by*

$$y^{k+1,i} = \begin{cases} x^k - \alpha_k \dfrac{f_i(x^k)}{\| t^k \|^2} t^k & \text{if } f_i(x^k) > 0, \\ x^k & \text{if } f_i(x^k) \leq 0, \end{cases} \tag{2.2}$$

where $t^k \in \partial f_i(x^k)$ *is a subgradient of* f_i *at the point* x^k, *and the relaxation parameters* $\{\alpha_k\}_{k=0}^{\infty}$ *are confined to an interval* $\epsilon_1 \leq \alpha_k \leq 2 - \epsilon_2$, *for all* $k \geq 0$, *with some arbitrarily small* $\epsilon_1, \epsilon_2 > 0$.

(ii) *Calculate the next iterate* x^{k+1} *by*

$$x^{k+1} = \sum_{i=1}^{m} w_i y^{k+1,i}, \tag{2.3}$$

where w_i *are fixed, user-chosen, positive weights with* $\sum_{i=1}^{m} w_i = 1$.

The convergence analysis for this algorithm is currently available only for consistent ($Q \neq \emptyset$) CFPs; see [23, 32]. In our experimental work, reported in what follows, we applied Algorithm 2.2 to CFPs without knowing whether or not they are consistent. Convergence is diagnosed by performing the plots of a *proximity function* that measures, in some manner, the infeasibility of the system. We used the weighted proximity function of the form

$$p(x) := (1/2) \sum_{i=1}^{m} w_i \parallel P_i(x) - x \parallel^2, \tag{2.4}$$

where $P_i(x)$ is the orthogonal projection of the point x onto Q_i. To combat instabilities in those plots that appeared occasionally in our experiments, we used *steering parameters* σ_k instead of the relaxation parameters α_k in Algorithm 2.2. To this end we need the following definition.

DEFINITION 2.3. *A sequence* $\{\sigma_k\}_{k=0}^{\infty}$ *of real numbers* $0 \leq \sigma_k < 1$ *is called a* steering sequence *if it satisfies the following conditions:*

$$\lim_{k \to \infty} \sigma_k = 0, \tag{2.5}$$

$$\sum_{k=0}^{\infty} \sigma_k = +\infty, \tag{2.6}$$

$$\sum_{k=0}^{\infty} \mid \sigma_k - \sigma_{k+m} \mid < +\infty. \tag{2.7}$$

A historical and technical discussion of these conditions can be found in [1]. The sequential and simultaneous Halpern–Lions–Wittmann–Bauschke algorithms discussed in Censor [10] employ the parameters of a steering sequence to "force" (steer) the iterates towards the solution of the best approximation problem. This steering feature of the steering parameters has a profound effect on the behavior of any sequence of iterates $\{x^k\}_{k=0}^{\infty}$. We return to this point in section 6.

ALGORITHM 2.4 (The SSP method with steering).

Initialization: $x^0 \in R^n$ *is arbitrary.*

Iterative step: (i) *Given* x^k, *calculate, for all* $i \in I = \{1, 2, \ldots, m\}$, *intermediate iterates* $y^{k+1,i}$ *by*

$$y^{k+1,i} = \begin{cases} x^k - \sigma_k \dfrac{f_i(x^k)}{\parallel t^k \parallel^2} t^k & \text{if } f_i(x^k) > 0, \\ x^k & \text{if } f_i(x^k) \leq 0, \end{cases} \tag{2.8}$$

where $t^k \in \partial f_i(x^k)$ *is a subgradient of* f_i *at the point* x^k, *and* $\{\sigma_k\}_{k=0}^{\infty}$ *is a sequence of steering parameters.*

(ii) *Calculate the next iterate* x^{k+1} *by*

$$x^{k+1} = \sum_{i=1}^{m} w_i y^{k+1,i}, \tag{2.9}$$

where w_i *are fixed, user-chosen, positive weights with* $\sum_{i=1}^{m} w_i = 1$.

3. Subgradient projections with strategical relaxation: Preliminaries. Considering the CFP (1.2), the *envelope* of the family of functions $\{f_i\}_{i=1}^m$ is the function

$$f(x) := \max\{f_i(x) \mid i = 1, 2, \ldots, m\}, \tag{3.1}$$

which is also convex. Clearly, the consistent CFP is equivalent to finding a point in

$$Q = \cap_{i=1}^m Q_i = \{x \in R^n \mid f(x) \leq 0\}. \tag{3.2}$$

The subgradient projections algorithmic scheme that we propose here employs a strategy for controlling the relaxation parameters in a specific manner, leaving enough user flexibility while giving some mathematical guarantees for the algorithm's behavior in the inconsistent case. It is described as follows.

ALGORITHM 3.1.

Initialization: *Let M be a positive real number, and let $x^0 \in R^n$ be any initial point.*

Iterative step: *Given the current iterate x^k, set*

$$I(x^k) := \{i \mid 1 \leq i \leq m \quad \textit{and} \quad f_i(x^k) = f(x^k)\}, \tag{3.3}$$

and choose a nonnegative vector $w^k = (w_1^k, w_2^k, \ldots, w_m^k) \in R^m$ such that

$$\sum_{i=1}^m w_i^k = 1 \quad \textit{and} \quad w_i^k = 0 \quad \textit{if} \quad i \notin I(x^k). \tag{3.4}$$

Let λ_k be any nonnegative real number such that

$$\max\left(0,\ f(x^k)\right) \leq \lambda_k M^2 \leq 2\max\left(0,\ f(x^k)\right) \tag{3.5}$$

and calculate

$$x^{k+1} = x^k - \lambda_k \sum_{i \in I(x^k)} w_i^k \xi_i^k, \tag{3.6}$$

where, for each $i \in I(x^k)$, we take a subgradient $\xi_i^k \in \partial f_i(x^k)$.

It is interesting to note that any sequence $\{x^k\}_{k=0}^\infty$ generated by this algorithm is well defined, no matter how x^0 and M are chosen. Similarly to other algorithms described above, Algorithm 3.1 requires computing the subgradients of convex functions. In case a function is differentiable, this reduces to gradient calculations. Otherwise, one can use the subgradient computing procedure presented in Butnariu and Resmerita [8].

The procedure described above was previously studied in Butnariu and Mehrez [7]. The main result there shows that the procedure converges to a solution of the CFP under two conditions: (i) that the solution set Q has a nonempty interior, and (ii) that the envelope f is uniformly Lipschitz on R^n, that is, there exists a positive real number L such that

$$|f(x) - f(y)| \leq L\,\|x - y\| \quad \text{for all} \quad x, y \in R^n. \tag{3.7}$$

Both conditions (i) and (ii) are restrictive, and it is difficult to verify their validity in practical applications. In the following we show that this method converges to the

solutions of consistent CFPs under less demanding conditions. In fact, we show that if the solution set Q of the given CFP has a nonempty interior, then the convergence of Algorithm 3.1 to a point in Q is ensured even if the function f is not uniformly Lipschitz on R^n (i.e., even if f does not satisfy condition (ii) above). However, verifying whether $\operatorname{int} Q \neq \emptyset$ prior to solving a CFP may be difficult or even impossible. Therefore, it is desirable to have alternative conditions, which may be easier to verify in practice, that can ensure convergence of our algorithm to solutions of the CFP, provided that such solutions exist. This is why we prove the convergence of Algorithm 3.1 to the solutions of consistent CFPs whenever the envelope f of the functions f_i involved in the given CFP is strictly convex. The strict convexity of the envelope function f associated with a consistent CFP implies that either the solution set Q of the CFP is a singleton, in which case $\operatorname{int} Q = \emptyset$, or that Q contains (at least) two different solutions of the CFP implying that $\operatorname{int} Q \neq \emptyset$. The verification of whether Q is a singleton or not is as difficult as deciding whether $\operatorname{int} Q \neq \emptyset$. By contrast, since f is strictly convex whenever each f_i is strictly convex, the verification of the strict convexity of f may be relatively easily done in some situations of practical interest, such as when each f_i is a quadratic convex function. In the latter case, strict convexity of f_i amounts to the positive definiteness of the matrix of its purely quadratic part.

It is interesting to note in this context that, when the envelope f of the CFP is not strictly convex, one may consider a "regularized" CFP in which each f_i, which is not strictly convex, is replaced by

$$\overline{f_i}(x) := f_i(x) + \alpha\|x\|^2 \tag{3.8}$$

for some positive real number α. Clearly, all $\overline{f_i}$ are strictly convex, and thus so is the envelope $\overline{f}$ of the regularized problem. Therefore, if the regularized problem has solutions, then our Algorithm 3.1 will produce the approximations of such solutions. Moreover, any solution of the regularized problem is a solution of the original problem, and thus by solving the regularized problem, we implicitly solve the original problem. The difficult part of this approach is that, even if the original CFP is consistent, then the regularized version of it may be inconsistent for all, or for some, values $\alpha > 0$. How to decide whether an $\alpha > 0$ exists such that the corresponding regularized CFP is consistent, and how to compute such an α (if any) are questions whose answers we do not know.

4. Subgradient projections with strategical relaxation: Convergence analysis. In order to discuss the convergence behavior of the subgradient projections method with strategical relaxation, recall that convex functions defined on the whole space R^n are continuous and, consequently, are bounded on bounded sets in R^n. Therefore, the application of Butnariu and Iusem [6, Proposition 1.1.11] or Bauschke and Borwein [2, Proposition 7.8] to the convex function f shows that it is Lipschitz on bounded subsets of R^n, i.e., for any nonempty bounded subset $S \subseteq R^n$, there exists a positive real number $L(S)$, called *a Lipschitz constant of f over the set S*, such that

$$|f(x) - f(y)| \leq L(S)\,\|x - y\| \text{ for all } x, y \in S. \tag{4.1}$$

Our next result is a convergence theorem for Algorithm 3.1 when applied to a consistent CFP. It was noted in the previous section that Algorithm 3.1 is well defined regardless of how the initial point x^0 or the positive constant M involved in the algorithm are chosen. However, this is no guarantee that a sequence $\{x_k\}_{k=0}^{\infty}$ generated by Algorithm 3.1 for the random choices of x^0 and M will converge to the solutions

of the CFP, even if such solutions exist. The theorem below shows a way of choosing x^0 and M, which ensures that, under some additional conditions for the problem data, the sequence $\{x_k\}_{k=0}^{\infty}$ generated by Algorithm 3.1 will necessarily approximate a solution of the CFP (provided that solutions exist). As shown in section 5 below, determining x^0 and M as required in the next theorem can be quite easily done for practically significant classes of CFPs. Also, as shown in section 6, determining x^0 and M in this manner enhances the self-adaptability of the procedure to the problem data and makes Algorithm 3.1 produce the approximations of the solutions of the CFP which, in many cases, are more accurate than those produced by other CFP solving algorithms. We denote with $B(x, r)$ the ball centered at x with radius r.

THEOREM 4.1. *If a positive number M and an initial point x^0 in Algorithm* 3.1 *are chosen so that $M \geq L(B(x^0, r))$ for some positive real number r satisfying the condition*

$$B(x^0, r/2) \cap Q \neq \emptyset \tag{4.2}$$

and if at least one of the following conditions holds:

(i) $B(x^0, r/2) \cap \operatorname{int} Q \neq \emptyset$,

(ii) *the function f is strictly convex,*

then any sequence $\{x^k\}_{k=0}^{\infty}$, generated by Algorithm 3.1*, converges to an element of Q.*

We present the proof of Theorem 4.1 as a sequence of lemmas. To do so, note that, if $\{x^k\}_{k=0}^{\infty}$ is generated by Algorithm 3.1, then for each integer $k \geq 0$, we have

$$x^{k+1} = x^k - \lambda_k \nu^k, \tag{4.3}$$

where

$$\nu^k := \sum_{i \in I(x^k)} w_i^k \xi_i^k \in \operatorname{conv} \cup_{i \in I(x^k)} \partial f_i(x^k). \tag{4.4}$$

Using (4.3), for any $z \in R^n$, we have

$$\left\|x^{k+1} - z\right\|^2 = \left\|x^k - z\right\|^2 + \lambda_k \left(\lambda_k \left\|\nu^k\right\|^2 - 2\left\langle \nu^k, x^k - z\right\rangle\right). \tag{4.5}$$

By Clarke [15, Proposition 2.3.12] we deduce that

$$\partial f(x^k) = \operatorname{conv} \cup_{i \in I(x^k)} \partial f_i(x^k), \tag{4.6}$$

and this implies that $\nu^k \in \partial f(x^k)$ because of (4.4). Therefore,

$$\left\langle \nu^k, z - x^k \right\rangle \leq f'_+(x^k; z - x^k), \tag{4.7}$$

where $f'_+(u; v)$ denotes the right-sided directional derivative at u in the direction v. Now suppose that M, r, and x^0 are chosen according to the requirements of Theorem 4.1, that is,

$$r > 0, M \geq L(B(x^0, r)) \text{ and } B(x^0, r/2) \cap Q \neq \emptyset. \tag{4.8}$$

Next we prove the following basic fact.

LEMMA 4.2. *If* (4.8) *is satisfied and if $z \in B(x^0, r/2) \cap Q$, then, for all $k \geq 0$, we have, for any sequence $\{x^k\}_{k=0}^{\infty}$, generated by Algorithm* 3.1*,*

$$x^{k+1} \in B(x^0, r) \text{ and } \left\|x^{k+1} - z\right\| \leq \left\|x^k - z\right\| \leq r/2. \tag{4.9}$$

Proof. We first show that, if for some integer $k \geq 0$,

$$x^k \in B(x^0, r) \text{ and } \left\|x^k - z\right\| \leq r/2, \tag{4.10}$$

then (4.9) holds. If $\lambda_k = 0$ or $\nu^k = 0$, then, by (4.3), we have $x^{k+1} = x^k$, which combined with (4.10), implies (4.9). Assume now that $\lambda_k \neq 0$ and $\nu^k \neq 0$. Since, by (4.10), $x^k \in B(x^0, r)$ by (4.8) and by [15, Proposition 2.1.2(a)], we deduce that

$$M \geq L(B(x^0, r)) \geq \left\|\nu^k\right\|. \tag{4.11}$$

According to (3.5), we also have $f(x^k) > 0$ (otherwise $\lambda_k = 0$). Since $f(z) \leq 0$ we obtain from the subgradient inequality

$$\left\langle \nu^k, x^k - z \right\rangle \geq f(x^k) - f(z) \geq f(x^k) > 0. \tag{4.12}$$

This and (4.11) imply that

$$2\left\langle \nu^k, x^k - z \right\rangle \geq 2f(x^k) \geq \lambda_k M^2 \geq \lambda_k \left\|\nu^k\right\|^2 \tag{4.13}$$

showing that the quantity inside the parentheses in (4.5) is nonpositive. Thus, we deduce that

$$\left\|x^{k+1} - z\right\| \leq \left\|x^k - z\right\| \leq r/2 \tag{4.14}$$

in this case too. This proves that if (4.10) is true for all $k \geq 0$, then so is (4.9). Now, we prove by induction that (4.10) is true for all $k \geq 0$. If $k = 0$, then (4.10) obviously holds. Suppose that (4.10) is satisfied for some $k = p$. As shown above, this implies that condition (4.9) is satisfied for $k = p$, and thus we have that

$$x^{p+1} \in B(x^0, r) \text{ and } \left\|x^{p+1} - z\right\| \leq r/2. \tag{4.15}$$

Hence, condition (4.10) also holds for $k = p + 1$. Consequently, condition (4.9) holds for $k = p + 1$, and this completes the proof. ∎

Observe that, according to Lemma 4.2, if $\{x^k\}_{k=0}^{\infty}$ is a sequence generated by Algorithm 3.1 and if the conditions (4.8) are satisfied, then there exists $z \in B(x^0, r/2) \cap Q$ and for any such z the sequence $\{\|x^k - z\|\}_{k=0}^{\infty}$ is nonincreasing and bounded from below and therefore convergent. Since the sequence $\{\|x^k - z\|\}_{k=0}^{\infty}$ is convergent, it is also bounded, and consequently, the sequence $\{x^k\}_{k=0}^{\infty}$ is bounded too. This shows that the next result applies to any sequence $\{x^k\}_{k=0}^{\infty}$ generated by Algorithm 3.1 under the assumptions of Theorem 4.1.

LEMMA 4.3. *If $\{x^k\}_{k=0}^{\infty}$ is a bounded sequence generated by Algorithm* 3.1*, then the sequence $\{x^k\}_{k=0}^{\infty}$ has accumulation points, and, for each accumulation point x^* of $\{x^k\}_{k=0}^{\infty}$, there exists a sequence of natural numbers $\{k_s\}_{s=0}^{\infty}$ such that the following limits exist:*

$$x^* = \lim_{s\to\infty} x^{k_s},\ \lambda_* = \lim_{s\to\infty} \lambda_{k_s}, \tag{4.16}$$

$$\xi_i^* = \lim_{s\to\infty} \xi_i^{k_s},\ w_i^* = \lim_{s\to\infty} w_i^{k_s} \text{ for all } i = 1, 2, \ldots, m, \tag{4.17}$$

$$\nu^* = \lim_{s\to\infty} \nu^{k_s}, \tag{4.18}$$

and we have

$$w^* := (w_1^*, w_2^*, \ldots, w_m^*) \in R_+^m \text{ and } \sum_{i\in I(x^*)} w_i^* = 1 \tag{4.19}$$

and

$$\nu^* = \sum_{i \in I(x^*)} w_i^* \xi_i^* \in \partial f(x^*). \tag{4.20}$$

Moreover, if $\lambda_* = 0$, then x^* is a solution of the CFP.

Proof. The sequence $\{x^k\}_{k=0}^{\infty}$ is bounded, and thus has accumulation points. Let x^* be an accumulation point of $\{x^k\}_{k=0}^{\infty}$, and let $\{x^{p_s}\}_{s=0}^{\infty}$ be a convergent subsequence of $\{x^k\}_{k=0}^{\infty}$ such that $x^* = \lim_{s\to\infty} x^{p_s}$. The function f is continuous (since it is real-valued and convex on R^n); hence, it is bounded on bounded subsets of R^n. Therefore, the sequence $\{f(x^{p_s})\}_{s=0}^{\infty}$ converges to $f(x^*)$, and the sequence $\{f(x^k)\}_{k=0}^{\infty}$ is bounded. By (3.5), the boundedness of $\{f(x^k)\}_{k=0}^{\infty}$ implies that the sequence $\{\lambda_k\}_{k=0}^{\infty}$ is bounded. Since, for every $i = 1, 2, \ldots, m$, the operator $\partial f_i : R^n \to 2^{R^n}$ is monotone, it is locally bounded (cf. Pascali and Sburlan [35, Theorem on p. 104]).

Consequently, there exists a neighborhood U of x^* on which all ∂f_i, $i = 1, 2, \ldots, m$ are bounded. Clearly, since $x^* = \lim_{s\to\infty} x^{p_s}$, the neighborhood U contains all but finitely many terms of the sequence $\{x^{p_s}\}_{s=0}^{\infty}$. This implies that the sequences $\{\xi_i^{p_s}\}_{s=0}^{\infty}$ are uniformly bounded, and therefore the sequence $\{\nu^{p_s}\}_{s=0}^{\infty}$ is bounded too.

Therefore, there exists a subsequence $\{k_s\}_{s=0}^{\infty}$ of $\{p_s\}_{s=0}^{\infty}$ such that the limits in (4.16)–(4.18) exist. Obviously, the vector $w^* = (w_1^*, w_2^*, \ldots, w_m^*) \in R_+^m$, and according to [7, Lemma 1], we also have $\sum_{i\in I(x^*)} w_i^* = 1$. This and (4.4) imply that $\nu^* = \sum_{i\in I(x^*)} w_i^* \xi_i^*$.

Observe that, since $\nu^{k_s} \in \partial f(x^{k_s})$ for all $s \geq 0$ and since ∂f is a closed mapping (cf. Phelps [37, Proposition 2.5]), we have that $\nu^* \in \partial f(x^*)$. Now, if $\lambda_* = 0$, then according to (3.5) and the continuity of f, we deduce that

$$0 \leq \max\{0, f(x^*)\} = \lim_{s\to\infty} \max\{0, f(x^{k_s})\} \leq \lim_{s\to\infty} \lambda_{k_s} M^2 = \lambda_* M^2 = 0, \tag{4.21}$$

which implies that $f(x^*) \leq 0$, that is, $x^* \in Q$. □

LEMMA 4.4. *Let $\{x^k\}_{k=0}^{\infty}$ be a sequence generated by Algorithm* 3.1. *If* (4.8) *is satisfied and if at least one of the conditions* (i) *or* (ii) *of Theorem* 4.1 *holds, then the sequence $\{x^k\}_{k=0}^{\infty}$ has accumulation points, and any such point belongs to Q.*

Proof. As noted above, when (4.8) is satisfied, then the sequence $\{x^k\}_{k=0}^{\infty}$ is bounded, and hence it has accumulation points. Let x^* be such an accumulation point, and let $\{k_s\}_{s=0}^{\infty}$ be the sequence of natural numbers associated with x^* whose existence is guaranteed by Lemma 4.3. Since, for any $z \in C \cap B(x^0, r/2)$, the sequence $\{\|x^k - z\|\}_{k=0}^{\infty}$ is convergent (cf. Lemma 4.2), we deduce that

$$\begin{aligned} \|x^* - z\| &= \lim_{s\to\infty} \left\|x^{k_s} - z\right\| = \lim_{k\to\infty} \left\|x^k - z\right\| = \lim_{s\to\infty} \left\|x^{k_s+1} - z\right\| \\ &= \|x^* - \lambda_* \nu^* - z\|. \end{aligned} \tag{4.22}$$

This implies that

$$\|x^* - z\|^2 = \|x^* - z\|^2 + \lambda_* \left(\lambda_* \|\nu^*\|^2 - 2\langle \nu^*, x^* - z\rangle\right). \tag{4.23}$$

If $\lambda_* = 0$, then $x^* \in Q$ by Lemma 4.3. Suppose that $\lambda_* > 0$. Then, by (4.23), we have

$$\lambda_* \|\nu^*\|^2 - 2\langle \nu^*, x^* - z\rangle = 0 \tag{4.24}$$

for all $z \in C \cap B(x^0, r/2)$. We distinguish now between two possible cases.

Case I: Assume that condition (i) of Theorem 4.1 is satisfied. According to (4.24), the set $Q \cap B(x^0, r/2)$ is contained in the hyperplane

$$H := \left\{ x \in R^n \mid \langle \nu^*, x \rangle = (1/2) \left(2 \langle \nu^*, x^* \rangle - \lambda_* \|\nu^*\|^2 \right) \right\}. \tag{4.25}$$

By condition (i) of Theorem 4.1, it follows that $\operatorname{int}(Q \cap B(x^0, r/2)) \neq \emptyset$, and this is an open set contained in $\operatorname{int} H$. So, unless $\nu^* = 0$ (in which case $H = R^n$), we have reached a contradiction because $\operatorname{int} H = \emptyset$. Therefore, we must have $\nu^* = 0$. According to Lemma 4.3, we have $0 = \nu^* \in \partial f(x^*)$, which implies that x^* is a global minimizer of f. Consequently, for any $z \in Q$, we have $f(x^*) \leq f(z) \leq 0$, that is, $x^* \in Q$.

Case II: Assume that condition (ii) of Theorem 4.1 is satisfied. According to (4.24), we have

$$\lambda_* \|\nu^*\|^2 = 2 \langle \nu^*, x^* - z \rangle . \tag{4.26}$$

By (3.5), the definition of M, and [15, Proposition 2.1.2] we deduce that

$$2f(x^{k_s}) \geq \lambda_{k_s} M^2 \geq \lambda_{k_s} \left\| \nu^{k_s} \right\|^2 \tag{4.27}$$

for all integers $s \geq 0$. Letting $s \to \infty$ we get

$$2f(x^*) \geq \lambda_* M^2 \geq \lambda_* \|\nu^*\|^2 = 2 \langle \nu^*, x^* - z \rangle , \tag{4.28}$$

where the last equality follows from (4.26). Consequently, we have

$$f(x^*) \geq \langle \nu^*, x^* - z \rangle \quad \text{for all} \quad z \in Q \cap B(x^0, r/2). \tag{4.29}$$

The convexity of f implies that, for all $z \in Q \cap B(x^0, r/2)$,

$$-f(x^*) \leq \langle \nu^*, z - x^* \rangle \leq f(z) - f(x^*) \leq -f(x^*). \tag{4.30}$$

Therefore, we have that

$$-f(x^*) = \langle \nu^*, z - x^* \rangle = f(z) - f(x^*) \text{ for all } \quad z \in Q \cap B(x^0, r/2). \tag{4.31}$$

Thus $f(z) = 0$, for all $z \in Q \cap B(x^0, r/2)$. Hence, using again the convexity of f, we deduce that, for all $\quad z \in Q \cap B(x^0, r/2)$,

$$f'_+(x^*; z - x^*) \leq f(z) - f(x^*) = -f(x^*) = \langle \nu^*, z - x^* \rangle \leq f'_+(x^*; z - x^*). \tag{4.32}$$

This implies that

$$f'_+(x^*; z - x^*) = \langle \nu^*, z - x^* \rangle = f(z) - f(x^*) \text{ for all } \quad z \in Q \cap B(x^0, r/2). \tag{4.33}$$

Since, by condition (ii) of Theorem 4.1, f is strictly convex, we also have (see [6, Proposition 1.1.4]) that

$$f'_+(x^*; z - x^*) < f(z) - f(x^*) \text{ for all } \quad z \in \left(Q \cap B(x^0, r/2) \right) \setminus \{x^*\}. \tag{4.34}$$

Hence, the equalities in (4.33) cannot hold unless $Q \cap B(x^0, r/2) = \{x^*\}$, and thus $x^* \in Q$. $\square$

The previous lemmas show that if (4.8) holds and if one of the conditions (i) or (ii) of Theorem 4.1 is satisfied, then the sequence $\{x^k\}_{k=0}^{\infty}$ is bounded, and all of its accumulation points are in Q. In fact, the results above say something more. Namely, in view of Lemma 4.2, they show that if (4.8) holds and if one of the conditions (i) or (ii) of Theorem 4.1 is satisfied, then all of the accumulation points x^* of $\{x^k\}_{k=0}^{\infty}$ are contained in $Q \cap B(x^0, r)$ because all x^k are in $B(x^0, r)$ by (4.9). In order to complete the proof of Theorem 4.1, it remains to show that the following result is true.

LEMMA 4.5. *Under the conditions of Theorem* 4.1 *any sequence* $\{x^k\}_{k=0}^{\infty}$*, generated by Algorithm* 3.1*, has at most one accumulation point.*

Proof. Observe that, under the conditions of Theorem 4.1, the conditions (4.8) are satisfied, and therefore the sequence $\{x^k\}_{k=0}^{\infty}$ is bounded. Let x^* be an accumulation point of $\{x^k\}_{k=0}^{\infty}$. By Lemma 4.4 we deduce that $x^* \in Q$, i.e., $f(x^*) \leq 0$. Consequently, for any natural number k we have

$$\langle \nu^k, x^k - x^* \rangle \geq f(x^k) - f(x^*) \geq f(x^k).$$

Now, using this fact, a reasoning similar to that which proves (4.13) but made with x^* instead of z leads to

$$2 \langle \nu^k, x^k - x^* \rangle \geq \lambda_k \left\| \nu^k \right\|^2$$

for all natural numbers k. This and (4.5) combined imply that the sequence $\{\|x^k - x^*\|\}_{k=0}^{\infty}$ is nonincreasing and therefore convergent. Consequently, if $\{x^{k_p}\}_{p=0}^{\infty}$ is a subsequence of $\{x^k\}_{k=0}^{\infty}$ such that $\lim_{p\to\infty} x^{k_p} = x^*$, we have

$$\lim_{k\to\infty} \left\| x^k - x^* \right\| = \lim_{p\to\infty} \left\| x^{k_p} - x^* \right\| = 0$$

showing that any accumulation point x^* of $\{x^k\}_{k=0}^{\infty}$ is exactly the limit of $\{x^k\}_{k=0}^{\infty}$. ∎

The application of Theorem 4.1 depends on our ability to choose numbers M and r and a vector x^0 such that condition (4.8) is satisfied. We show below that this can be done when the functions f_i of the CFP (1.2) are quadratic or affine, and there is some a priori known ball which intersects Q. In actual applications it may be difficult to a priori decide whether the CFP (1.2) has or does not have solutions. However, as noted above, Algorithm 3.1 is well defined and will generate sequences $\{x^k\}_{k=0}^{\infty}$ no matter how the initial data M, r, and x^0 are chosen. This leads to the question whether it is possible to decide if Q is empty or not by simply analyzing the behavior of sequences $\{x^k\}_{k=0}^{\infty}$ generated by Algorithm 3.1. A partial answer to this question is contained in the following result.

COROLLARY 4.6. *Suppose that the CFP* (1.2) *has no solution, and that the envelope* f *is strictly convex. Then, no matter how the initial vector* x^0 *and the positive number* M *are chosen, any sequence* $\{x^k\}_{k=0}^{\infty}$ *generated by Algorithm* 3.1*, has the following properties:*

(i) *If* $\{x^k\}_{k=0}^{\infty}$ *is bounded and*

$$\lim_{k\to\infty} \left\| x^{k+1} - x^k \right\| = 0, \tag{4.35}$$

then f *has a (necessarily unique) minimizer, and* $\{x^k\}_{k=0}^{\infty}$ *converges to that minimizer, while*

$$\lim_{k\to\infty} f(x^k) = \inf\{f(x) \mid x \in R^n\}. \tag{4.36}$$

(ii) *If f has no minimizer, then the sequence $\{x^k\}_{k=0}^{\infty}$ is unbounded, or the sequence $\{\|x^{k+1}-x^k\|\}_{k=0}^{\infty}$ does not converge to zero.*

Proof. Clearly, (ii) is a consequence of (i). In order to prove (i) observe that, since the CFP (1.2) has no solution, all values of f are positive. Also, if f has a minimizer, then this minimizer is unique because f is strictly convex.

If $\{x^k\}_{k=0}^{\infty}$ is bounded, then it has an accumulation point, say, x^*. By Lemma 4.3 there exists a sequence of positive integers $\{k_s\}_{s=0}^{\infty}$ such that (4.16) and (4.19)–(4.20) are satisfied. Using Lemma 4.3 again, we deduce that, if the limit λ_* in (4.16) is zero, then the vector $x^* = \lim_{s\to\infty} x^{k_s}$ is a solution of the CFP (1.2), i.e., $f(x^*) \leq 0$, contradicting the assumption that the CFP (1.2) has no solution. Hence, $\lambda_* > 0$. By (4.3), (4.35), and (4.16) we have that

$$0 = \lim_{s\to\infty} \lambda_{k_s} \nu^{k_s} = \lambda_* \nu^*. \tag{4.37}$$

Thus, we deduce that $\nu^* = 0$. From (4.19)–(4.20) and [15, Proposition 2.3.12] we obtain

$$0 = \nu^* = \sum_{i\in I(x^*)} w_i^* \xi_i^* \in \partial f(x^*) \tag{4.38}$$

showing that x^* is a minimizer of f. So, all of the accumulation points of $\{x^k\}_{k=0}^{\infty}$ coincide because f has no more than one minimizer. Consequently, the bounded sequence $\{x^k\}_{k=0}^{\infty}$ converges, and its limit is the unique minimizer of f. ∎

Remark 4.7. Checking numerically a condition such as (ii) in Corollary 4.6 or the condition in Corollary 4.9 below seems virtually impossible. But there is no escape from such situations in such mathematically oriented results. Condition (ii) in Corollary 4.6 is meaningful in the inconsistent case in which a feasible point does not exist, but a proximity function that "measures" the feasibility violation of the limit point can be minimized. An easy adaptation of the proof of Corollary 4.6 shows that, if the sequence $\{x^k\}_{k=0}^{\infty}$ has a bounded subsequence $\{x^{k_t}\}_{t=0}^{\infty}$ such that the limit $\lim_{t\to\infty}(x^{k_t+1} - x^{k_t}) = 0$, then all of the accumulation points of $\{x^{k_t}\}_{t=0}^{\infty}$ are the minimizers of f (even if f happens to be not strictly convex).

Remark 4.8. The fact that, for some choice of x^0 and M, a sequence $\{x^k\}_{k=0}^{\infty}$, generated by Algorithm 3.1, has the property that $\lim_{k\to\infty} f(x^k) = 0$ does not imply that the CFP (1.2) has a solution. For example, take in (1.2) $m = n = 1$ and $f_1(x) = e^{-x}$. Clearly, in this case (1.2) has no solution, and $f = f_1$. However, for $x^0 = 0$, $M = 1$, and $\lambda_k = (3/2)f(x^k)$, we have $\lim_{k\to\infty} f(x^k) = 0$.

A meaningful implication of Corollary 4.6 is the following result.

COROLLARY 4.9. *Suppose that the CFP* (1.2) *has no solution, and that f is strictly convex. Then, no matter how the initial vector x^0 and the positive number M are chosen in Algorithm* 3.1*, the following holds: If the series $\sum_{k=0}^{\infty} \|x^k - x^{k+1}\|$ converges, then the function f has a unique global minimizer and the sequence $\{x^k\}_{k=0}^{\infty}$, generated by Algorithm* 3.1*, converges to that minimizer, while the sequence $\{f(x^k)\}_{k=0}^{\infty}$ converges to* $\inf\{f(x) \mid x \in R^n\}$.

Proof. When $\sum_{k=0}^{\infty} \|x^k - x^{k+1}\|$ converges to some number S we have

$$\left\|x^0 - x^{k+1}\right\| \leq \sum_{\ell=0}^{k} \left\|x^\ell - x^{\ell+1}\right\| \leq S \tag{4.39}$$

for all integers $k \geq 0$. This implies that the sequence $\{x^k\}_{k=0}^{\infty}$ is bounded, and $\lim_{k\to\infty} \|x^k - x^{k+1}\| = 0$. Hence, by applying Corollary 4.6, we complete the proof. ∎

Remark 4.10. Finding an initial vector x^0, the radius r and a positive number M satisfying condition $M \geq L(B(x^0, r))$ (and satisfying (4.2) provided that Q is nonempty) when there is no a priori knowledge about the existence of a solution of the CFP can be quite easily done when at least one of the sets Q_i, say Q_{i_0}, is bounded and the functions f_i are differentiable. In this case it is sufficient to determine a vector x^0 and a positive number r large enough so that the ball $B(x^0, r/2)$ contains Q_{i_0}. Clearly, for such a ball, if Q is nonempty, then condition (4.2) holds. Once the ball $B(x^0, r)$ is determined, finding a number $M \geq L(B(x^0, r))$ can be done by taking into account that the gradients of the differentiable convex functions $f_i : R^n \to R$ are necessarily continuous, and therefore the numbers

$$L_i = \sup\{\|\nabla f_i(x)\| \mid x \in B(x^0, r)\} \tag{4.40}$$

are necessarily finite. Since $L := \max\{L_i \mid 1 \leq i \leq m\}$ is necessarily a Lipschitz constant of f over $B(x^0, r)$, one can take $M = L$.

Remark 4.11. The method of choosing x^0, r, and M presented in Remark 4.10 does not require a priori knowledge of the existence of a solution of the CFP and can be applied even when Q is empty. In such a case one should compute, along the iterative procedure of Algorithm 3.1, the sums $S_k = \sum_{\ell=0}^{k} \|x^\ell - x^{\ell+1}\|$. Theorem 4.1 and Corollary 4.9 then provide the following insights and tools for solving the CFP, provided that f is strictly convex:

- If along the computational process the sequence S_k remains bounded from above by some number S^*, while the sequence $\{f(x^k)\}_{k=0}^{\infty}$ stabilizes itself asymptotically at some *positive* value, then the given CFP has no solution, but the sequence $\{x^k\}_{k=0}^{\infty}$ still approximates a global minimum of f, which may be taken as a surrogate solution of the given CFP.
- If along the computational process the sequence S_k remains bounded from above by some number S^*, while the sequence $\{f(x^k)\}_{k=0}^{\infty}$ stabilizes itself asymptotically at some *nonpositive* value, then the given CFP has a solution, and the sequence $\{x^k\}_{k=0}^{\infty}$ approximates such a solution.

5. Implementation of Algorithm 3.1 for linear or quadratic functions. The application of Algorithm 3.1 does not require a priori knowledge of the constant r. However, in order to implement this algorithm so that the conditions for convergence will be guaranteed, we have to determine numbers r and M as required by Theorem 4.1. The method proposed in Remark 4.10 might yield a very large value of r. This is due to the mathematical generality of Remark 4.10. The quadratic and affine cases treated next seem to be restrictive from the theoretical/mathematical point of view, but their importance lies in the fact that they cover many significant real-world applications.

We deal first with the problem of determining a number M such that

$$M \geq L(B(x^0, r)) \tag{5.1}$$

provided that an $r > 0$ is given. Recall that, if $g : R^n \to R$ is a continuously differentiable function, then by Taylor's formula, we have that, whenever $x, y \in B(x^0, r)$, there exists a $u \in [x, y]$ such that

$$\begin{aligned} \mid g(y) - g(x) \mid &= \mid \langle \nabla g(u), y - x \rangle \mid \leq \|\nabla g(u)\| \, \|y - x\| \\ &\leq \|y - x\| \max\{\|\nabla g(u)\| \mid u \in B(x^0, r)\}. \end{aligned} \tag{5.2}$$

This shows that

$$\max\{\|\nabla g(u)\| \mid u \in B(x^0, r)\} \tag{5.3}$$

is a Lipschitz constant for g on $B(x^0, r)$. Suppose now that each function f_i is either linear or quadratic. Denote $I_1 = \{i \mid 1 \le i \le m,\ f_i \text{ is linear}\}$ and $I_2 = \{i \mid 1 \le i \le m,\ f_i \text{ is quadratic}\}$. Namely,

$$f_i(x) = \langle a^i, x\rangle + b_i \ \text{ for all } \ i \in I_1, \tag{5.4}$$

with $a^i \in R^n \backslash \{0\}$ and $b_i \in R$, and

$$f_i(x) = \langle x, U_i x\rangle + \langle a^i, x\rangle + b_i \text{ for all } \ i \in I_2, \tag{5.5}$$

where $U_i = (u^i_{\ell,k})$ is a symmetric positive semidefinite $n \times n$ matrix, $a^i \in R^n$, and $b_i \in R$. We have, of course,

$$\nabla f_i(x) = \begin{cases} a^i & \text{if } \ i \in I_1, \\ 2U_i x + a^i & \text{if } \ i \in I_2, \end{cases} \tag{5.6}$$

so that (5.3) can give us Lipschitz constants for each f_i over $B(x^0, r)$. Denote

$$L_i := \begin{cases} \|a^i\| & \text{if } i \in I_1, \\ 2\|U_i\|_\infty \left(\|x^0\| + r\right) + \|a^i\| & \text{if } i \in I_2, \end{cases} \tag{5.7}$$

where $\|U_i\|_\infty$ is the operator norm of U_i. Due to (4.6), this implies that $\cup_{x \in B(x^0, r)} \partial f(x) \subseteq B(0, L)$, where

$$L := \max\{L_i \mid 1 \le i \le m\}. \tag{5.8}$$

Taking $\xi \in \partial f(x)$ and $\zeta \in \partial f(y)$, for some $x, y \in B(x^0, r)$, we have

$$\begin{aligned} L\|x - y\| \ge \|\zeta\|\,\|x - y\| \ge \langle \zeta, y - x\rangle \ge f(y) - f(x) \\ \ge \langle \xi, y - x\rangle \ge -\|\xi\|\,\|x - y\| \ge -L\|x - y\|, \end{aligned} \tag{5.9}$$

which implies that

$$|f(y) - f(x)| \le L\|x - y\| \ \text{ for all } \ x, y \in B(x^0, r). \tag{5.10}$$

In other words, L is a Lipschitz constant of f over $B(x^0, r)$. Thus, given an $r > 0$, we can take M to be any number such that $M \ge L$. Note that choosing x^0 such that the corresponding r is small may speed up the computational process by reducing the number of iterations needed to reach a reasonably good approximate solution of the CFP. In general, determining a number r is straightforward when one has some information about the range of the variation of the coordinates of some solutions to the CFP.

For instance, if one knows a priori that the solutions of the CFP are vectors $x = (x_j)_{j=1}^n$ such that

$$\ell_j \le x_j \le u_j, \ 1 \le j \le n, \tag{5.11}$$

where, $\ell_j, u_j \in R$ for all j, then the set Q is contained in the hypercube of edge length $\delta = u_{\max} - \ell_{\min}$, whose faces are parallel to the axes of the coordinates, and centered at the point x^0 whose coordinates are $x^0_j = \frac{1}{2}(\ell_{\min} + u_{\max})$, where

$$\ell_{\min} := \min\{\ell_j \mid 1 \le j \le n\} \text{ and } u_{\max} := \max\{u_j \mid 1 \le j \le n\}. \tag{5.12}$$

Therefore, by choosing this x^0 as the initial point for Algorithm 3.1 and choosing $r = \sqrt{n}\delta$, condition (4.2) holds.

6. Computational results. In this section, we compare the performance of Algorithms 2.2, 2.4, and 3.1 by examining a few test problems. There are a number of degrees-of-freedom used to evaluate and compare the performance of the algorithms. These are the maximum number of iterations, the number of constraints, the lower and upper bounds of the box constraints, the values of the relaxation parameters, the initial values of the steering parameters, and the steering sequence. In all our experiments, the steering sequence of Algorithm 2.4 assumed the form

$$\sigma_k = \frac{\sigma}{k+1}, \tag{6.1}$$

with a fixed user-chosen constant σ. The main performance measure is the value of $f(x^k)$ plotted as a function of the iteration index k.

6.1. Test problem description. There are three types of constraints in our test problems: box constraints, linear constraints, and quadratic constraints. Some of the numerical values used to generate the constraints are uniformly distributed random numbers lying in the interval $\tau = [\tau_1, \tau_2]$, where τ_1 and τ_2 are user-chosen predetermined values.

The n box constraints are defined by

$$\ell_j \le x_j \le u_j, \quad j = 1, 2, \ldots, n, \tag{6.2}$$

where $\ell_j, u_j \in \tau$ are the lower and upper bounds, respectively. Each of the N_q quadratic constrains is generated according to

$$G_i(x) = \langle x, U_i x \rangle + \langle v^i, x \rangle + \beta_i, \quad i = 1, 2, \ldots, N_q. \tag{6.3}$$

Here U_i are the $n \times n$ matrices defined by

$$U_i = W_i \Lambda_i W_i^T, \tag{6.4}$$

where the $n \times n$ matrices Λ_i are diagonal and positive definite, given by

$$\Lambda_i = \operatorname{diag}\left(\delta_1^i, \delta_2^i, \ldots, \delta_n^i\right), \tag{6.5}$$

where $0 < \delta_1^i \le \delta_2^i \le \cdots \le \delta_n^i \in \tau$ are generated randomly. The matrices W_i are generated by orthonormalizing an $n \times n$ random matrix whose entries lie in the interval τ. Finally, the vector $v^i \in R^n$ is constructed so that all of its components lie in the interval τ, and similarly, the scalar $\beta_i \in \tau$. The N_ℓ linear constraints are constructed in a similar manner according to

$$L_i(x) = \langle y^i, x \rangle + \gamma_i, \quad i = 1, 2, \ldots, N_\ell. \tag{6.6}$$

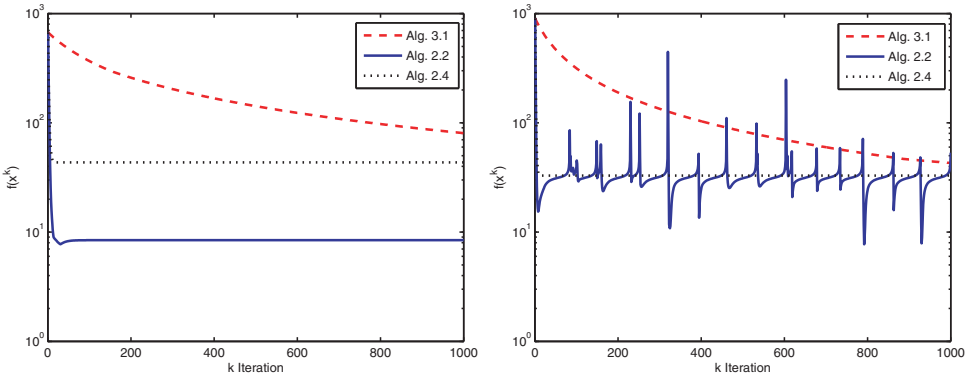
Thus, the total number of constraints is $n + N_q + N_\ell$.

Table 6.1 summarizes the test cases used to evaluate and compare the performance of Algorithms 2.2, 2.4, and 3.1. In these eight experiments, we modified the value of the constant σ in (6.1), the interval τ, the number of constraints, the number of iterations, and the relative tolerance ε, used as a termination criterion between subsequent iterations.

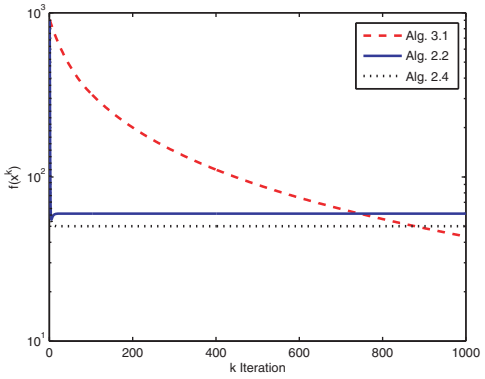
In Table 6.1, Cases 1 and 2 represent small-scale problems with a total of 13 constraints, whereas Cases 4–6 represent midscale problems with a total of 130 constraints. Cases 6–8 examine the case of overrelaxation, wherein the initial steering (relaxation) parameter is at least 2.

TABLE 6.1
Test cases for performance evaluation.

Case	$\alpha/\sigma/\lambda$	τ	n	N_q	N_ℓ	Iterations	ε
1	1.1		3	5	5	1,000	
2	1.1		3	5	5	1,000	
3	1.98		3	5	5	1,000	
4	1.98	$[-0.1, 0.1]$	30	50	50	1,000	0.1
5	1.98	$[-10, 10]$	30	50	50	100,000	0.1
6	2	$[-0.1, 0.1]$	30	50	50	1,000	0.1
7	3	$[-10, 10]$	3	5	5	1,000	0.1
8	5	$[-0.1, 0.1]$	3	5	5	1,000	0.1



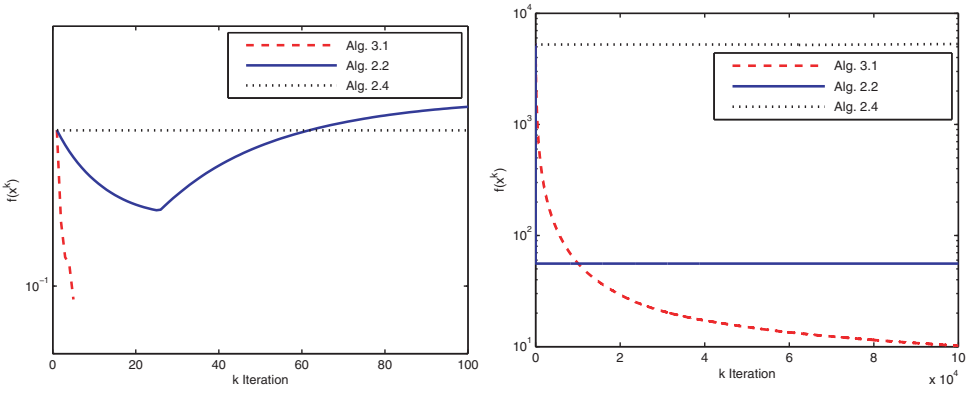
(a) Case 1. (b) Case 2.



(c) Case 3.

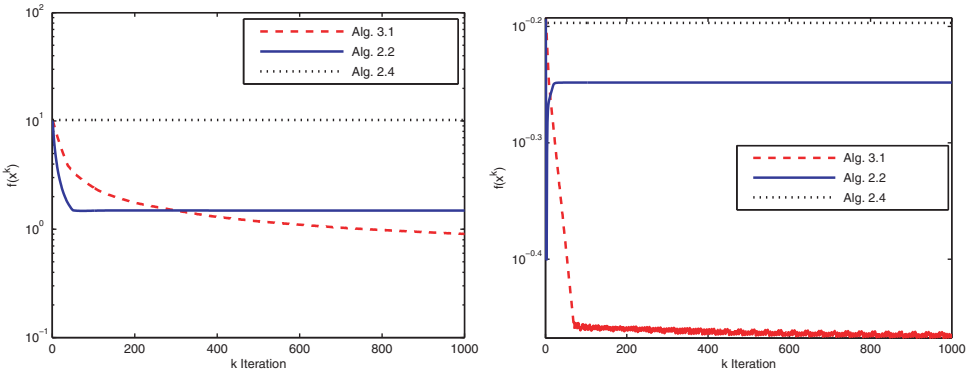
FIG. 1. *Simulation results for a small-scale problem, comparing Algorithms* 2.2, 2.4, *and* 3.1.

6.2. Results. The results of our experiments are depicted in Figures 1–3. The results of Cases 1–3 are shown in Figures 1(a)–1(c), respectively. It is seen that, in Case 1, Algorithm 2.2 has better initial convergence than Algorithms 2.4 and 3.1. However, in Case 2, Algorithm 2.4 yields fast and smooth initial behavior, while Algorithm 2.2 oscillates chaotically. Algorithm 3.1 exhibits slow initial convergence, similarly to Case 1. In Case 3, Algorithm 3.1 supersedes the performance of the other two algorithms, since it continues to converge toward zero. However, none of the algo-

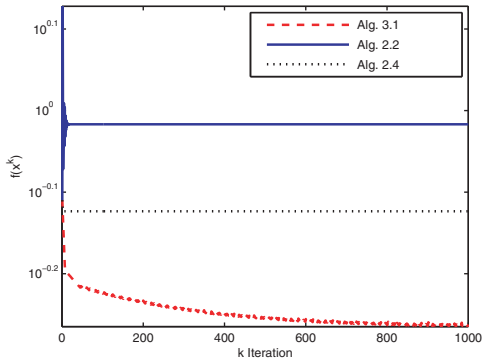


(a) Case 4. (b) Case 5.

FIG. 2. *Simulation results for a midscale problem, comparing Algorithms* 2.2, 2.4, *and* 3.1.



(a) Case 6. (b) Case 7.



(c) Case 8.

FIG. 3. *Simulation results for small-scale and midscale problems with overrelaxation, comparing Algorithms* 2.2, 2.4, *and* 3.1.

rithms detects a feasible solution, since none converged to the tolerance threshold after the maximum number of iterations.

The midsized problems of Cases 4 and 5 are depicted by Figures 2(a) and 2(b). Figure 2(a) shows that Algorithm 3.1 detects a feasible solution, while both Algorithms 2.2 and 2.4 fail to detect such a solution. The curve of Algorithm 3.1 in Figure 2(a) stops when it reaches the feasible point detection tolerance, which is 0.1. Once the point is detected, there is no need to further iterate, and the process stops. The curve for Algorithm 2.2 in this figure shows irregular behavior since it searches for a feasible solution without reaching the detection threshold of 0.1, and accumulated numerical errors start to affect it. Figure 2(b) shows a phenomenon similar to the one observed in the small-scale problem: Algorithm 3.1 continues to seek a feasible solution, while Algorithms 2.2 and 2.4 converge to a steady state, indicating the failure to detect a feasible solution.

In the experiments Cases 6–8, Algorithm 3.1 outperforms the other algorithms, arriving very close to finding feasible solutions. It should be observed that the behavior of Algorithm 3.1 observed above is the result of the way in which the relaxation parameters λ_k are self-regulating their sizes. In Algorithm 3.1 the relaxation parameter λ_k can be chosen (see (3.6)) to be any number of the form

$$\lambda_k = \beta_k \frac{\max(0, f(x^k))}{M^2} + 2(1-\beta_k)\frac{\max(0, f(x^k))}{M^2} = (2-\beta_k)\frac{\max(0, f(x^k))}{M^2}, \tag{6.7}$$

where β_k runs over the interval $[0, 1]$. Consequently, the size of λ_k can be very close to zero when x^k is close to a feasible solution (no matter how β_k is chosen in $[0, 1]$). Also, λ_k may happen to be much larger than 2 when x^k is far from a feasible solution, and the number $f(x^k)$ is large enough (note that $2 - \beta_k$ stays between 1 and 2). So, Algorithm 3.1 is naturally underrelaxing or overrelaxing the computational process according to the relative position of the current iterate x^k to the feasibility set of the problem. As our experiments show, in some circumstances, this makes Algorithm 3.1 behave better than the other procedures we compare with it. At the same time, the self-regulation of the relaxation parameters, which is essential in Algorithm 3.1, may happen to reduce the initial speed of convergence of this procedure, that is, Algorithm 3.1 may require more computational steps in order to reach a point x^k, which is close enough to the feasibility set such that its self-regulatory features are to be really advantageous for providing a very precise solution of the given problem (which the other procedures may fail to do since they may become stationary in the vicinity of the feasibility set). Another interesting feature of Algorithm 3.1, which differentiates it from the other algorithms we compare it with, is its essentially nonsimultaneous character: Algorithm 3.1 does not necessarily ask for $w_i^k > 0$ for all $i \in \{1, \ldots, m\}$. The set of positive weights w_i^k, which condition the progress of the algorithm at step k, essentially depends on the current iterate x^k (see (3.4)) and allows reducing the number of subgradients needed to be computed at each iterative step (in fact, one can content himself with only one $w_i^k > 0$, and thus with a single subgradient ξ_i^k). This may be advantageous in cases when computing subgradients is difficult and therefore time consuming.

The main observations can be summarized as follows:

1. Algorithm 3.1 exhibits faster initial convergence than the other algorithms in the vicinity of points with very small $f(x^k)$. When the algorithms reach points with small $f(x^k)$ values, then Algorithm 3.1 tends to further reduce the value of $f(x^k)$, while the other algorithms tend to converge onto a constant steady-state value.

2. The problem dimensions in our experiments have little impact on the behavior of the algorithms.
3. All the examined small-scale problems have no feasible solutions. This can be seen from the fact that all three algorithms stabilize around $f(x^k) = 50$.
4. The chaotic oscillations of Algorithm 2.2 in the underrelaxed case is due to the fact that this algorithm has no internal mechanism to self-adapt its progress to the distance between the current iterates and the sets whose intersections are to be found. This phenomenon can hardly happen in Algorithm 3.1 because its relaxation parameters are self-adapting to the size of the current difference between successive iterations. This is an important feature of this algorithm. However, this feature also renders it somewhat slower than the other algorithms.
5. In some cases, Algorithms 2.2 and 2.4 indicate that the problem has no solution. In contrast, Algorithm 3.1 continues to make progress and seems to indicate that the problem has a feasible solution. This phenomenon is again due to the self-adaptation mechanism and can be interpreted in one of the following ways: (a) The problem indeed has a solution, but Algorithms 2.2 and 2.4 are unable to detect it (because they stabilize too fast). Algorithm 3.1 detects a solution provided that it is given enough running time; (b) The problem has no solution, and then Algorithm 3.1 will stabilize close to zero, indicating that the problem has no solution, but this may be due to computing (round-off) errors. Thus, a very small perturbation of the functions involved in the problem may render the problem feasible.

7. Conclusions. We have studied here mathematically and experimentally subgradient projections methods for the convex feasibility problem. The behavior of the fully simultaneous subgradient projections method in the inconsistent case is not known. Therefore, we studied and tested two options. One is the use of steering parameters instead of relaxation parameters, and the other is a variable relaxation strategy, which is self-adapting. Our small-scale and midscale experiments are not, decisive in all aspects and call for further research. But one general feature of the algorithm with the self-adapting strategical relaxation is its stability (nonoscillatory) behavior, and its relentless improvement of the iterations towards a solution in all cases. At this time we have not yet refined enough our experimental setup. For example, by the iteration index k on the horizontal axes of our plots we consider a whole sweep through all the sets of the convex feasibility problem, regardless of the algorithm. This is a good first approximation by which to compare the different algorithms. More accurate comparisons should use actual run times. Also, several numerical questions still remain unanswered in this report. These include the effect of various values of the constant σ as well as algorithmic behavior for higher iteration indices. In light of the applications mentioned in section 1, higher dimensional problems must be included. These and other computational questions are currently investigated.

Acknowledgments. We gratefully acknowledge the constructive comments of two anonymous referees, which helped us improve an earlier version of this paper.

REFERENCES

[1] H.H. Bauschke, *The approximation of fixed points of compositions of nonexpansive mappings in Hilbert space*, J. Math. Anal. Appl., 202 (1996), pp. 150–159.

[2] H.H. Bauschke and J.M. Borwein, *On projection algorithms for solving convex feasibility problems*, SIAM Rev., 38 (1996), pp. 367–426.
[3] H.H. Bauschke, J.M. Borwein, and A.S. Lewis, *The method of cyclic projections for closed convex sets in Hilbert space*, in Recent Developments in Optimization Theory and Nonlinear Analysis, Contemp. Math., 204, Y. Censor and S. Reich, eds., 1997, pp. 1–38.
[4] H.H. Bauschke, P.L. Combettes, and S.G. Kruk, *Extrapolation algorithm for affine-convex feasibility problems*, Numer. Algorithms, 41 (2006), pp. 239–274.
[5] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM, Philadelphia, 1994.
[6] D. Butnariu and A.N. Iusem, *Totally Convex Functions for Fixed Points Computation and Infinite Dimensional Optimization*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000.
[7] D. Butnariu and A. Mehrez, *Convergence criteria for generalized gradient methods of solving locally Lipschitz feasibility problems*, Comput. Optim. Appl., 1 (1992), pp. 307–326.
[8] D. Butnariu and E. Resmerita, *Averaged subgradient methods for optimization and Nash equilibria computation*, Optimization, 51 (2002), pp. 863–888.
[9] Y. Censor, *Mathematical optimization for the inverse problem of intensity modulated radiation therapy*, in Intensity-Modulated Radiation Therapy: The State of the Art, American Association of Physicists in Medicine, Medical Physics Monograph 29, J.R. Palta and T.R. Mackie, eds., Medical Physics Publishing, Madison, Wisconsin, 2003, pp. 25–49.
[10] Y. Censor, *Computational acceleration of projection algorithms for the linear best approximation problem*, Linear Algebra Appl., 416 (2006), pp. 111–123.
[11] Y. Censor, M.D. Altschuler, and W.D. Powlis, *On the use of Cimmino's simultaneous projections method for computing a solution of the inverse problem in radiation therapy treatment planning*, Inverse Problems, 4 (1988), pp. 607–623.
[12] Y. Censor and A. Lent, *Cyclic subgradient projections*, Math. Program., 24 (1982), pp. 233–235.
[13] Y. Censor and S.A. Zenios, *Parallel Optimization: Theory, Algorithms, and Applications*, Oxford University Press, New York, 1997.
[14] J.W. Chinneck, *The constraint consensus method for finding approximately feasible points in nonlinear programs*, INFORMS J. Comput., 16 (2004), pp. 255–265.
[15] F.H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley & Sons, New York, 1983.
[16] P.L. Combettes, *The foundations of set-theoretic estimation*, Proc. IEEE, 81 (1993), pp. 182–208.
[17] P.L. Combettes, *Inconsistent signal feasibility problems: Least-squares solutions in a product space*, IEEE Trans. Signal Process., SP-42 (1994), pp. 2955–2966.
[18] P.L. Combettes, *The convex feasibility problem in image recovery*, Adv. Imaging and Electron Phys., 95 (1996), pp. 155–270.
[19] P.L. Combettes, *Convex set theoretic image recovery by extrapolated iterations of parallel subgradient projections*, IEEE Trans. Image Process., 6 (1997), pp. 493–506.
[20] G. Crombez, *Non-monotoneous parallel iteration for solving convex feasibility problems*, Kybernetika, 39 (2003), pp. 547–560.
[21] G. Crombez, *A sequential iteration algorithm with non-monotoneous behavior in the method of projections onto convex sets*, Czechoslovak Math. J., 56 (2006), pp. 491–506.
[22] F. Deutsch, *Best Approximation in Inner Product Spaces*, Springer-Verlag, New York, 2001.
[23] L.T. Dos Santos, *A parallel subgradient method for the convex feasibility problem*, J. Comput. Appl. Math., 18 (1987), pp. 307–320.
[24] I.I. Eremin, *Fejér mappings and convex programming*, Siberian Math. J., 10 (1969), pp. 762–772.
[25] L. Gubin, B. Polyak, and E. Raik, *The method of projections for finding the common point of convex sets*, USSR Comput. Math. Math. Phys., 7 (1967), pp. 1–24.
[26] G.T. Herman, *Image Reconstruction From Projections: The Fundamentals of Computerized Tomography*, Academic Press, New York, 1980.
[27] G.T. Herman and L.B. Meyer, *Algebraic reconstruction techniques can be made computationally efficient*, IEEE Trans. Medical Image, 12 (1993), pp. 600–609.
[28] A.N. Iusem and A.R. De Pierro, *Convergence results for an accelerated nonlinear Cimmino algorithm*, Numer. Math., 49 (1986), pp. 367–378.
[29] K.C. Kiwiel, *The efficiency of subgradient projection methods for convex optimization* I: *General level methods*, SIAM J. Control Optim., 34 (1996), pp. 660–676.
[30] K.C. Kiwiel, *The efficiency of subgradient projection methods for convex optimization* II: *Implementations and extensions*, SIAM J. Control Optim., 34 (1996), pp. 677–697.
[31] K.C. Kiwiel and B. Lopuch, *Surrogate projection methods for finding fixed points of firmly nonexpansive mappings*, SIAM J. Optim., 7 (1997), pp. 1084–1102.

[32] A.N. Iusem and L. Moledo, *A finitely convergent method of simultaneous subgradient projections for the convex feasibility problem*, Comput. Appl. Math., 5 (1986), pp. 169–184.
[33] L.D. Marks, W. Sinkler, and E. Landree, *A feasible set approach to the crystallographic phase problem*, Acta Crystallogr., A55 (1999), pp. 601–612.
[34] M. A. Noor, *Some developments in general variational inequalities*, Appl. Math. Comput., 152 (2004), pp. 197–277.
[35] D. Pascali and S. Sburlan, *Nonlinear Mappings of Monotone Type*, Sijthoff & Noordhoff, Alphen aan den Rijn, the Netherlands, 1978.
[36] G. Pierra, *Decomposition through formalization in a product space*, Math. Program., 28 (1984), pp. 96–115.
[37] R.R. Phelps, *Convex Functions, Monotone Operators and Differentiability*, 2nd ed., Springer-Verlag, Berlin, 1993.
[38] I. Yamada, *Hybrid steepest descent method for variational inequality problem over the fixed point set of certain quasi-nonexpansive mappings*, Numer. Funct. Anal. Optim., 25 (2004), pp. 619–655.

SIAM J. OPTIM.
Vol. 19, No. 2, pp. 808–823

AN AUGMENTED PRIMAL-DUAL METHOD FOR LINEAR CONIC PROGRAMS*

FLORIAN JARRE† AND FRANZ RENDL‡

Abstract. We propose a new iterative approach for solving linear programs over convex cones. Assuming that Slater's condition is satisfied, the conic problem is transformed to the minimization of a convex differentiable function in the primal-dual space. This function shows similarities with the augmented Lagrangian function and is called "augmented primal-dual function" or "apd-function". The evaluation of the function and its derivative is cheap if the projection of a given point onto the cone can be computed cheaply, and if the projection of a given point onto the affine subspace defining the primal problem can be computed cheaply. For the special case of a semidefinite program, a certain regularization of the apd-function is analyzed. Numerical examples minimizing the apd-function with a conjugate gradient method illustrate the potential of the approach.

Key words. conic program, linear convergence, augmented primal-dual function

AMS subject classifications. 90C22, 90C25, 49M29

DOI. 10.1137/070687128

1. Introduction.

We present a new method for solving convex conic programs. The method is based on minimizing a convex differentiable "augmented primal-dual function" (apd-function) that is related to the augmented Lagrangian but less problem dependent and does not require any penalty parameter. If Slater's condition is satisfied, the problem of solving the conic program is equivalent to minimizing the apd-function. The evaluation of function value and gradient of the apd-function requires two conic projections and two projections on an affine subspace. If these projections are cheap, it is possible to minimize the apd-function by any descent method such as a conjugate gradient type method or a limited memory BFGS method. In our numerical examples we report results obtained with a conjugate gradient approach.

When applying this algorithm to the apd-function of a linear program in standard form with a system matrix $A \in \mathbb{R}^{m\times n}$, a factorization of the matrix AA^T can be computed in a preprocessing phase. Given this factorization the cost per iteration is of order $O(mn)$ arithmetic operations. When minimizing the apd-function of a linear program by Newton's method—at a cost of order $O(n^3)$ arithmetic operations per iteration—this algorithm converges in a finite number of iterations. We therefore anticipate that a conjugate gradient or BFGS approach will converge rapidly as well.

When applied to a semidefinite program having a unique and strictly complementary solution the algorithm is sublinearly convergent. We therefore derive a simple modification of the apd-function for which Newton's method is locally quadratically convergent. Generalizations of this modification to Cartesian products of the semidefinite cone, the second-order cone, and the positive orthant are straightforward.

*Received by the editors April 1, 2007; accepted for publication (in revised form) March 24, 2008; published electronically July 3, 2008. The authors acknowledge partial financial support by the Marie Curie Training Network Algorithmic Discrete Optimization (ADONET), MRTN-CT-2003-504438, financed by the European Union, and by the GICOLAG program at the International Erwin Schrödinger Institute for Mathematical Physics (ESI) Vienna, Austria.

http://www.siam.org/journals/siopt/19-2/68712.html

†Institut für Mathematik, Universität Düsseldorf, D–40225 Düsseldorf, Germany (jarre@opt.uni-duesseldorf.de).

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

Our approach is related to projection methods such as considered, for example, in [1]. New elements of this paper are a transformation of a conic problem into an affine-convex pair with cheap projections onto the affine set, a conjugate gradient acceleration, a regularization for the case of semidefinite programs, and promising numerical examples.

Several methods have been proposed in the literature to overcome the limits of interior point methods for solving large-scale semidefinite programs. We recall the spectral bundle method [8] which uses eigenvalue optimization. The low-rank factorization approach of Burer and Monteiro [2] treats semidefinite programs using nonlinear optimization techniques. The currently strongest computational results are reported in the papers by Toh [22] and by Kocvara and Stingl [11]. Toh uses an iterative solver for the augmented KKT system, and Kocvara and Stingl apply an iterative solver to a modified barrier problem. The approach presented in the current paper is closely related to the "boundary point method" from [18] and the regularization approaches in [13].

2. Linear conic programs. We consider linear conic programs of the form

$$(P) \qquad \text{minimize } \langle c, x\rangle \ \text{ s.t. } \ x \in K \cap (\mathcal{L} + b),$$

where K is a closed convex cone in a finite dimensional Euclidean space E, $\mathcal{L}$ is a linear subspace, and $b, c \in E$ are given data. We always assume that the dimension of E is denoted by n. Our practical applications refer to the cases where K is the positive orthant in $E = \mathbb{R}^n$ and where K is the cone of symmetric positive semidefinite matrices in $E = \mathcal{S}^l = \{X \in \mathbb{R}^{l\times l} \mid X = X^T\}$. Here, $n = l(l+1)/2$.

We always assume that K has a nonempty interior (no hidden equality constraints) and that K is pointed (in LP-notation this assumption means there are no free variables).

Often the set $\mathcal{L}$ is given in the form

$$(1) \qquad \mathcal{L} = \{\tilde{x} \mid A\tilde{x} = 0\} \qquad \text{and} \qquad \mathcal{L} + b = \{x \mid Ax = Ab\},$$

where A is a matrix or some other representation of a linear operator. In particular, our analysis yields an algorithm applicable to linear programs of the form

$$\text{minimize } c^T x \ \text{ s.t. } \ Ax = \bar{b}, \ \ x \geq 0,$$

where $\bar{b} := Ab$.

We use the dual program as introduced in [14], section 4.2,

$$(D) \qquad \text{minimize } \langle b, s\rangle \ \text{ s.t. } \ s \in K^D \cap (\mathcal{L}^\perp + c),$$

where $\mathcal{L}^\perp$ is the orthogonal complement of $\mathcal{L}$ and

$$K^D = \{s \in E \mid \langle s, x\rangle \geq 0 \ \forall x \in K\}$$

is the dual cone of K.

It is easily verified that weak duality holds, namely

$$\langle b, c\rangle \leq \langle c, x\rangle + \langle b, s\rangle$$

for all x, s that are feasible for (P) and (D). When (P) satisfies Slater's condition and (D) has a feasible solution, then strong duality holds; see [14], Theorem 4.2.1. In

this case, a point x is optimal for (P) if, and only if, there exists a point s feasible for (D) with

$$\langle b, c\rangle = \langle c, x\rangle + \langle b, s\rangle. \tag{2}$$

We denote such x and s by x^{opt} and s^{opt}.

3. Decomposing the conic program. The linear constraints of (P) and (D) (including (2)) are satisfied for all points in the affine space

$$\boldsymbol{A} := (\mathcal{L} + b) \times (\mathcal{L}^{\perp} + c) \cap \{(x; s) \mid \ \langle c, x\rangle + \langle b, s\rangle = \langle b, c\rangle\} \subset E \times E,$$

and the conic constraints are satisfied for all points in the cone

$$\boldsymbol{C} := K \times K^D \subset E \times E.$$

By the assumption on K, it follows that $\boldsymbol{C}$ is full dimensional, $\dim(\boldsymbol{C}) = 2n$. We assume that $\boldsymbol{A}$ is of dimension $n-1$. (In the case that $b \in \mathcal{L}$ and $c \in \mathcal{L}^{\perp}$ the set $\boldsymbol{A}$ is of dimension n. As we do not provide solutions in the relative interior of the solution set, this case is trivial with optimal solution $x^{opt} = s^{opt} = 0$.)

Solving (P) is equivalent to finding $z := (x; s) \in \boldsymbol{A} \cap \boldsymbol{C}$ where $\boldsymbol{A}$ is an affine subspace and $\boldsymbol{C}$ a convex cone. Moreover, as we will see, projections onto $\boldsymbol{A}$ and $\boldsymbol{C}$ are easily computable for the case of linear or semidefinite programming.

For a closed set $\mathcal{C}$ and a vector $\bar{z}$ we denote the distance of $\bar{z}$ to $\mathcal{C}$ by

$$d(\bar{z}, \mathcal{C}) := \min\{\|z - \bar{z}\|_2 \mid z \in \mathcal{C}\}.$$

Thus, solving (P) is also equivalent to finding z such that

$$\phi(z) := \frac{1}{2}(d(z, \boldsymbol{A})^2 + d(z, \boldsymbol{C})^2) = 0, \tag{3}$$

i.e., such that the differentiable convex function ϕ is minimized. (Differentiability of ϕ is shown in Lemma 1 below.)

When (P) is a linear program in standard form, the function $d(z, \boldsymbol{C})^2$ is of the form $\sum_i ((z_i)^-)^2$ where $(z_i)^- := \min\{0, z_i\}$. Thus, ϕ is closely related to the augmented Lagrangian function. We therefore call ϕ an augmented primal-dual function. It differs from the augmented Lagrangian in that the representation of the linear subspace $\mathcal{L}$ (i.e., the matrix A when $\mathcal{L}$ is of the form (1)) does not enter the representation of ϕ. In other words, ϕ is less "data dependent" than the augmented Lagrangian, and it depends on a larger number of unknowns. As we will see, however, the dependence on a large number of unknowns does not imply that computations with ϕ are numerically expensive.

LEMMA 1. *For a closed convex set $\mathcal{C}$ let*

$$\Pi_{\mathcal{C}} \text{ be the orthogonal projection}$$

(with respect to the Euclidean norm) onto $\mathcal{C}$. Then, given an algorithm for the evaluation of $\Pi_{\mathcal{C}}$, the distance $d(z, \mathcal{C}) = \|z - \Pi_{\mathcal{C}}(z)\|_2$ is easily computed. Moreover, a steepest descent step of step length 1 starting at z for minimizing the differentiable function $\phi_{\mathcal{C}}$ with

$$\phi_{\mathcal{C}}(z) := \frac{1}{2} d(z, \mathcal{C})^2$$

will lead to the nearest point minimizing d (i.e., to $\Pi_{\mathcal{C}}(z)$).

Proof. For completeness we provide an elementary proof of Lemma 1. It is easy to verify that $\Pi_{\mathcal{C}}$ is well defined (single-valued) and Lipschitz continuous with Lipschitz constant 1. We show that $\phi_{\mathcal{C}}(z)$ is a differentiable function and $\nabla\phi_{\mathcal{C}}(z) = z - \Pi_{\mathcal{C}}(z)$. Let $\hat{z} := \Pi_{\mathcal{C}}(z)$. Let Δz be arbitrary. We show that

$$\phi_{\mathcal{C}}(z + \lambda\Delta z) = \phi_{\mathcal{C}}(z) + \lambda\Delta z^T(z - \hat{z}) + o(|\lambda|).$$

First note that

$$2\phi_{\mathcal{C}}(z + \lambda\Delta z) \le \|\hat{z} - (z + \lambda\Delta z)\|_2^2 = \|\hat{z} - z\|_2^2 - 2\lambda(\hat{z} - z)^T\Delta z + O(\lambda^2).$$

On the other hand, let $\hat{z}(\lambda) := \Pi_{\mathcal{C}}(z+\lambda\Delta z)$. As $\hat{z}(\lambda) \in \mathcal{C}$ it follows $(\hat{z}(\lambda)-\hat{z})^T(z-\hat{z}) \le 0$, and $\|\hat{z}(\lambda) - \hat{z}\|_2 \le \|\lambda\Delta z\|_2$. It follows

$$2\phi_{\mathcal{C}}(z + \lambda\Delta z) = \|\hat{z}(\lambda) - (z + \lambda\Delta z)\|_2^2 = \|(\hat{z}(\lambda) - \hat{z}) + (\hat{z} - z) - \lambda\Delta z\|_2^2$$

$$\ge \|\hat{z} - z\|_2^2 - 2\lambda(\hat{z} - z)^T\Delta z - O(\lambda^2).$$

This completes the proof of Lemma 1. ∎

Unfortunately, the one-step convergence of the steepest descent method in Lemma 1 is lost when minimizing the sum $\phi(z) = \frac{1}{2}(d(z, \boldsymbol{A})^2 + d(z, \boldsymbol{C})^2)$ in (3).

Note that the projections onto $\boldsymbol{A}$ and $\boldsymbol{C}$—and thus the function ϕ—are easy to compute for the case of linear and semidefinite programming:

For the case of linear programming, $\boldsymbol{C}$ is the positive orthant in $\mathbb{R}^{2n}$, and the projection onto $\boldsymbol{C}$ can be performed in $O(n)$ arithmetic operations. In the case of semidefinite programming, $\boldsymbol{C}$ is the Cartesian product of the semidefinite cone with itself, and the projection onto $\boldsymbol{C}$ can be computed by performing the eigenvalue decompositions of two symmetric matrices (order l^3 operations).

In section 6 it is shown that also the projection onto $\boldsymbol{A}$ can be done efficiently for these two examples. If $\mathcal{L}$ is given as in (1) with $Ab \in \mathbb{R}^m$ and the Cholesky factorization of AA^T is computed in a preprocessing step before starting the algorithm, then the projection can be evaluated in $O(mn)$ operations.

4. Solving (*P*) and (*D*). As shown in the previous section, solving (P) and (D) is reduced to finding a point in the intersection of the two convex sets $\boldsymbol{A}$ and $\boldsymbol{C}$, both of which are explicitly given. In this section we assume that the intersection of $\boldsymbol{A}$ and $\boldsymbol{C}$ is nonempty.

4.1. Minimizing the distance between *A* and *C*. Standard projection methods solve the problem of finding a point in $\boldsymbol{A} \cap \boldsymbol{C}$ by the following *simple algorithm:*

ALGORITHM 1 (Alternating projections).

Initialization: Let $z^0 \in \boldsymbol{A}$ be given. Set $k = 0$.

1. *Set $\hat{z}^k := \Pi_{\boldsymbol{C}}(z^k)$.*
2. *Set $z^{k+1} := \Pi_{\boldsymbol{A}}(\hat{z}^k)$. Set $k = k + 1$. Go to Step 1.*

By Lemma 1, one iteration of Algorithm 1 can be interpreted as one steepest descent step for minimizing $\frac{1}{2}d(\,.\,, \boldsymbol{C})^2$ followed by a steepest descent step for minimizing $\frac{1}{2}d(\,.\,, \boldsymbol{A})^2$. In general, such method converges very slowly. We therefore consider an acceleration minimizing the sum of both functions by a conjugate gradient scheme:

4.2. Minimizing ϕ. The first simple approach for minimizing ϕ is a conjugate gradient type method with Polak–Ribière type update or Fletcher–Reeves type update of the search direction. For descent methods it is important to understand the behavior of the second derivative of the objective function.

For linear and semidefinite programming, the function ϕ is twice differentiable almost everywhere. (It is differentiable everywhere.) For linear programming the eigenvalues of the Hessian H of ϕ at any point z such that $H(z)$ exists are at most 2 (as each of the Hessians of $\frac{1}{2}d(\ .\ ,\boldsymbol{A})^2$ and $\frac{1}{2}d(\ .\ ,\boldsymbol{C})^2$ only has the eigenvalues zero and one.) The eigenvalues of H are nonnegative, but unfortunately, they may be zero or arbitrarily close to zero. This makes the application of descent methods for minimizing ϕ difficult. Before continuing our analysis of the function ϕ we reduce the number of degrees of freedom by restricting ϕ to a lower dimensional subspace:

Note that $\boldsymbol{A}$ is an affine subspace. We restrict ϕ to $\boldsymbol{A}$ and define the function $\tilde{\phi}$ by

$$\tilde{\phi}(\tilde{z}) := \phi(\tilde{z}) = \frac{1}{2}d(\tilde{z},\boldsymbol{C})^2 \qquad \text{for } \tilde{z} \in \boldsymbol{A}. \tag{4}$$

We stress that $\tilde{\phi}$ is not defined outside $\boldsymbol{A}$. To emphasize this fact we also denote the variable by $\tilde{z}$ rather than just z.

LEMMA 2. *The function $\tilde{\phi}$ is differentiable, and for $\tilde{z} \in \boldsymbol{A}$ its gradient is given by*

$$\nabla\tilde{\phi}(\tilde{z}) = \tilde{z} - \Pi_{\boldsymbol{A}}(\Pi_{\boldsymbol{C}}(\tilde{z})).$$

Proof. Let $\boldsymbol{A} = \boldsymbol{z} + \boldsymbol{L}$ where $\boldsymbol{z}$ is a fixed vector and $\boldsymbol{L}$ a linear subspace. Note that

$$\tilde{z} - \Pi_{\boldsymbol{A}}(\Pi_{\boldsymbol{C}}(\tilde{z})) = \Pi_{\boldsymbol{L}}(\tilde{z} - \Pi_{\boldsymbol{C}}(\tilde{z})).$$

By Lemma 1 it therefore suffices to recall the following more general simple statement:

If $\varphi : E \times E \to \mathbb{R}$ is a differentiable function, then the gradient of the restriction $\tilde{\varphi}$ of φ to $\boldsymbol{A}$ is given by

$$\nabla\tilde{\varphi}(\tilde{z}) = \Pi_{\boldsymbol{L}}(\nabla\varphi(\tilde{z})).$$

The gradient of $\tilde{\varphi}$ at $\tilde{z} \in \boldsymbol{A}$ is a vector $\tilde{w} \in \boldsymbol{L}$ such that

$$\tilde{\varphi}(\tilde{z} + \Delta\tilde{z}) = \tilde{\varphi}(\tilde{z}) + \tilde{w}^T\Delta\tilde{z} + o(\Delta\tilde{z})$$

for all sufficiently small $\Delta\tilde{z} \in \boldsymbol{L}$. The vector $w := \Pi_{\boldsymbol{L}}(\nabla\varphi(\tilde{z}))$ certainly lies in $\boldsymbol{L}$. For $\Delta\tilde{z} \in \boldsymbol{L}$ it follows from symmetry of $\Pi_{\boldsymbol{L}}$ that

$$\begin{aligned}
\tilde{\varphi}(\tilde{z}) + w^T\Delta\tilde{z} &= \varphi(\tilde{z}) + (\Pi_{\boldsymbol{L}}(\nabla\varphi(\tilde{z})))^T\Delta\tilde{z} \\
&= \varphi(\tilde{z}) + (\nabla\varphi(\tilde{z}))^T\Pi_{\boldsymbol{L}}\Delta\tilde{z} \\
&= \varphi(\tilde{z}) + (\nabla\varphi(\tilde{z}))^T\Delta\tilde{z} \\
&= \varphi(\tilde{z} + \Delta\tilde{z}) + o(\Delta\tilde{z}) \\
&= \tilde{\varphi}(\tilde{z} + \Delta\tilde{z}) + o(\Delta\tilde{z}).
\end{aligned}$$

This completes the proof. ∎

A steepest descent step with line search for minimizing $\tilde{\phi}$ starting at a point $\tilde{z} = z^k \in \boldsymbol{A}$ is the same as the computation of z^{k+1} with Algorithm 1 followed by an extrapolation along the line $z^k + \lambda(z^{k+1} - z^k)$. We briefly list a conjugate gradient acceleration of the steepest descent approach:

ALGORITHM 2 (CG-method for minimizing $\tilde{\phi}$).
Let $\tilde{z}^0 \in \boldsymbol{A}$ be given. Let $\Delta\tilde{z}^0 := -\nabla\tilde{\phi}(\tilde{z}^0)$. Set $k = 0$.

1. *Let $\lambda_k := argmin\{\tilde{\phi}(\tilde{z}^k + \lambda\Delta\tilde{z}^k) \mid \lambda > 0\}$.*
2. *Set $\tilde{z}^{k+1} := \tilde{z}^k + \lambda_k\Delta\tilde{z}^k$.*
3. *Compute $\Delta\tilde{z}^{k+1}$ from $\Delta\tilde{z}^k$ and $\nabla\tilde{\phi}(\tilde{z}^{k+1})$ using an update formula such as Polak–Ribière.*
4. *If k is a multiple of $(n-1)$, set $\Delta\tilde{z}^{k+1} := -\nabla\tilde{\phi}(\tilde{z}^{k+1})$ (restart).*
5. *Set $k := k+1$. Go to Step 1.*

The Polak–Ribière update introduced in [16] is analyzed in [4, 3]. Global convergence of suitable modifications of the algorithm can be established under rather weak conditions; see e.g. section 5.2 of [15].

Remark 1. The concept of Algorithm 2 is in some sense "complementary" to the boundary point method of [18]. The latter algorithm generates iterates within the primal-dual cone approaching the set of linear constraints, while the iterates in Algorithm 2 always satisfy the linear constraints and approach the primal-dual cone. The restriction to an affine space (rather than the nonlinear primal-dual cone) opens the door for CG- or limited memory BFGS-accelerations.

Remark 2. When $\boldsymbol{C}$ is polyhedral, and (P), (D) have a unique optimal solution z^{opt}, then the Hessian of $\tilde{\phi}$ is piecewise linear and positive definite near z^{opt} (since z^{opt} is necessarily strictly complementary!), and thus, Newton's method for minimizing $\tilde{\phi}$ converges in a finite number of iterations; see e.g., [7].

Now consider the case where $\boldsymbol{C}$ is not polyhedral. Below we give a very simple example with a unique, strictly complementary optimal solution z^{opt} of (P) and (D) such that there are directions $z^{opt} + \lambda\Delta\tilde{z}$ through z^{opt} along which the intersection of $\boldsymbol{A}$ and $\boldsymbol{C}$ is "tangential" (the function $\tilde{\phi}$ in (4) growing in the order of λ^4) and other directions along which the intersection of $\boldsymbol{A}$ and $\boldsymbol{C}$ is "transversal" ($\tilde{\phi}$ growing in the order of λ^2). This implies that the condition number of the Hessian of $\tilde{\phi}$ near z^{opt} is unbounded and the conjugate gradient method is likely to converge *sublinearly*! For the case of semidefinite programs, we therefore derive an acceleration for this situation.

5. Application to semidefinite programs. In this section we use the following notation common for semidefinite programs: The space of real symmetric $l \times l$-matrices is denoted by $\mathcal{S}^l$. The dimension of $\mathcal{S}^l$ is $n := l(l+1)/2$. The notation $X \succeq 0$ ($X \succ 0$) is used to indicate that $X \in \mathcal{S}^l$ is positive semidefinite (positive definite). The standard scalar product on the space of $l \times l$-matrices is given by

$$\langle C, X\rangle := C \bullet X := \operatorname{trace}(C^T X) = \sum_{i,j=1}^{l} C_{i,j}X_{i,j}.$$

For given matrices $A^{(i)} \in \mathcal{S}^l$, $i = 1, 2, \ldots, m$, we define a linear map $\mathcal{A} : \mathcal{S}^l \to \mathbb{R}^m$ by

$$\mathcal{A}(X) := \begin{bmatrix} A^{(1)} \bullet X \\ \vdots \\ A^{(m)} \bullet X \end{bmatrix}, \quad X \in \mathcal{S}^l.$$

The adjoint operator $\mathcal{A}^* : \mathbb{R}^m \to \mathcal{S}^l$ is given by

$$\mathcal{A}^*(y) = \sum_{i=1}^{m} y_i A^{(i)}, \quad y \in \mathbb{R}^m.$$

With these definitions, the standard pair of primal and dual linear semidefinite programs can now be stated as follows:

$$(P) \qquad \text{minimize} \quad C \bullet X \quad \text{subject to} \quad \mathcal{A}(X) = \bar{b}, \quad X \succeq 0$$

and

$$(D') \qquad \text{maximize} \quad \bar{b}^T y \quad \text{subject to} \quad \mathcal{A}^*(y) + S = C, \quad S \succeq 0.$$

The dual program is equivalent (in the sense that the optimal solutions coincide) to

$$(D) \qquad \text{minimize} \quad B \bullet S \quad \text{subject to} \quad S \in \mathcal{L}^\perp + C, \quad S \succeq 0,$$

where $B \in \mathcal{S}^l$ is such that $\mathcal{A}(B) = \bar{b}$ and $\mathcal{L} = \{X \in \mathcal{S}^l \mid \mathcal{A}(X) = 0\}$.

ASSUMPTION 1. *Throughout this section we assume that the matrices $A^{(i)}$ are linearly independent and that (P) and (D) are strictly feasible and that there is a unique and strictly complementary solution $Z^{opt} = (X^{opt}, S^{opt})$ of (P) and (D) satisfying $X^{opt} + S^{opt} \succ 0$.*

Simple example: We give a simple example of a pair of semidefinite programs (P) and (D) satisfying Assumption 1 such that the Hessian of $\tilde{\phi}$ (see (4)) has an unbounded condition number for $\bar{Z}$ near Z^{opt}. (The Hessian is not defined at Z^{opt}.) Let $m = 1$ and the data of (P) and (D) be given by

$$C := \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \qquad B := \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \text{and} \quad A^{(1)} := \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}.$$

The primal-dual optimal solution $Z^{opt} = (X^{opt}, S^{opt}) = (B, C)$ is unique and strictly complementary. The space $\boldsymbol{L} := \mathcal{L} \times \mathcal{L}^\perp \cap \{(\Delta X, \Delta S) \mid C \bullet \Delta X + B \bullet \Delta S = 0\}$ is given by

$$\boldsymbol{L} = \left\{ \Delta Z = (\Delta X, \Delta S) = \left(\begin{pmatrix} 2a & -a \\ -a & b \end{pmatrix}, \begin{pmatrix} -b & -b \\ -b & 0 \end{pmatrix} \right) \mid \ a, b \in \mathbb{R} \right\}.$$

By construction, $Z^{opt} + \Delta Z \in \boldsymbol{A}$ for $\Delta Z \in \boldsymbol{L}$, and for small $|a|, |b|$ it is easily verified that

$$d(Z^{opt} + \Delta Z, \boldsymbol{C}) = O(|b|) \quad \text{if } a = 0, \qquad d(Z^{opt} + \Delta Z, \boldsymbol{C}) = O(a^2) \quad \text{if } b = 0.$$

Thus, the second directional derivative of $\tilde{\phi}$ is zero Z^{opt} along the direction $b = 0$ and positive along the direction $a = 0$. Minimizing $\tilde{\phi}$ by some conjugate gradient scheme will result in a very slow algorithm.

Discussion: Of course, the above example is not surprising. We have given a convex characterization of the optimal solution of a convex program as the intersection of two convex sets $\boldsymbol{A}$ and $\boldsymbol{C}$, each of which is easily computable. We do not have the property that this characterization is well conditioned under "reasonable assumptions." So far, a computable characterization of the optimal solution of a convex program with both properties—convexity and well conditionedness—is unknown. (The KKT conditions are well conditioned under suitable assumptions, see, e.g., [5], but the complementarity part of the KKT conditions is nonconvex.) This lack of a convex *and* well conditioned characterization of the optimal solution is responsible for the fact that most polynomial-time methods for convex programs use some homotopy approach to compute an optimal solution.

5.1. A local acceleration. We propose an acceleration that can be applied locally near the optimal solution $Z^{opt} = (X^{opt}, S^{opt})$ of (P) and (D), e.g., when the minimization of $\tilde{\phi}$ is turning slow.

Let $\hat{f}(Z) = \hat{f}(X,S) := \|XS - SX\|_F^2$. The nonconvex function $\hat{f}$ is minimized at Z^{opt}. It is differentiable and the derivative

$$\nabla_Z \hat{f}(Z) = 2\begin{pmatrix} S^2X + XS^2 - 2SXS \\ X^2S + SX^2 - 2XSX \end{pmatrix}$$

can be computed in order n^3 operations. More precisely, by exploiting the fact that $XS = (SX)^T$, it can be evaluated with three matrix-matrix multiplications: two for evaluating $2XSX - XXS - SXX = X(SX - XS) + (X(SX - XS))^T$, and one more for the second block of $\nabla\hat{f}(Z)$. Also the derivative of the restriction of $\hat{f}$ to $\boldsymbol{A}$ can be computed as in the proof of Lemma 2.

We therefore propose to solve (P) and (D) in two stages, the first one minimizing $\tilde{\phi}$ for $\tilde{Z} \in \boldsymbol{A}$, and when convergence of this stage is slow, starting a second stage minimizing $\tilde{\phi} + \hat{f}$ for $\tilde{Z} \in \boldsymbol{A}$. For both stages we may use a nonlinear CG-method as in Algorithm 2. The CG-method is n-step locally quadratically convergent if the objective function is three times differentiable near Z^{opt} and if the Hessian at Z^{opt} is positive definite; see e.g. [4, 3, 17]. In Theorem 1 below, we show a slightly weaker statement.

Note: In the following we will consider only points in $\boldsymbol{A}$. For compactness of notation we omit the additional identification $\tilde{Z}$ to indicate that $\tilde{Z} \in \boldsymbol{A}$ and shortly write $Z \in \boldsymbol{A}$. The restriction of $\phi + \hat{f}$ to $\boldsymbol{A}$ will be denoted by Ψ,

$$\Psi(Z) := \phi(Z) + \hat{f}(Z) \qquad \text{for } Z \in \boldsymbol{A}.$$

Again, we emphasize the restriction to $\boldsymbol{A}$.

THEOREM 1. *The gradient of Ψ is strongly semismooth and the generalized Hessian is positive definite at Z^{opt}.*

In short, $G(\ .\ ) := \nabla_Z \Psi(\ .\ )$ is strongly semismooth at the point Z, if for small $\|H\|$ the relation $G(Z+H) = G(Z) + \partial G(Z+H)[H] + O(\|H\|^2)$ holds true, where ∂G denotes any element of the generalized Jacobian of G in the sense of Clarke; see, e.g., Section 2.1 in [21]. By Theorem 3.2 in [19], Theorem 1 implies quadratic convergence of Newton's method for minimizing Ψ. We therefore anticipate that also conjugate gradient-type algorithms or limited-memory BFGS algorithms will converge rapidly.

Proof. Strong semismoothness of the gradient of Ψ at Z^{opt} follows from [20]; here, we prove positive definiteness of the generalized Hessian.

We start by noting that in spite of $\hat{f}$ not being convex, the eigenvalues of the Hessian of $\hat{f}$ at Z^{opt} are nonnegative since Z^{opt} is a minimizer of $\hat{f}$. Hence it suffices to show that either ϕ or $\hat{f}$ has a positive curvature along any given direction through Z^{opt}.

Let a perturbation $\Delta Z = (\Delta X, \Delta S)$ with $Z^{opt} + \Delta Z \in \boldsymbol{A}$ and $\|\Delta X\|_F^2 + \|\Delta S\|_F^2 = 1$ be given. It suffices to show that there exists a $\rho > 0$ independent of ΔZ such that

$$\phi(Z^{opt} + \lambda\Delta Z) + \hat{f}(Z^{opt} + \lambda\Delta Z) \geq \lambda^2\rho$$

for sufficiently small $\lambda > 0$. To this end we proceed in four steps.

Step 1: We apply ΔZ to the linearized primal dual system and relate the norm of ΔZ to the norm of the right-hand side:

By complementarity, $X^{opt}S^{opt} = 0 = S^{opt}X^{opt}$, and thus the matrices $X^{opt} \succeq 0$ and $S^{opt} \succeq 0$ commute. This guarantees that there exists a unitary matrix U and

diagonal matrices

$$\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_l) \succeq 0 \quad \text{and} \quad \Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_l) \succeq 0 \tag{5}$$

such that

$$X^{opt} = U\Lambda U^T \quad \text{and} \quad S^{opt} = U\Sigma U^T. \tag{6}$$

By strict complementarity we may assume without loss of generality that there exists a $k \leq l$ such that

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k > 0 = \lambda_{k+1} = \cdots = \lambda_l$$

and

$$\sigma_1 = \sigma_2 = \cdots = \sigma_k = 0 < \sigma_{k+1} \leq \cdots \leq \sigma_l.$$

As shown in Corollary 1 in [6], the following system of $m + 2n$ linear equations for $2n + m$ unknowns $(\Delta X, \Delta S, \Delta y)$:

$$\begin{aligned} \mathcal{A}(\Delta X) &= p, \\ \mathcal{A}^*(\Delta y) + \Delta S &= Q, \\ \Pi_{\text{up}}(U^T(\Delta X S^{opt} + X^{opt}\Delta S)U) &= r, \end{aligned} \tag{7}$$

is nonsingular. Here, $\Pi_{\text{up}}(U^T(\Delta X S^{opt} + X^{opt}\Delta S)U)$ denotes the upper triangular part of the matrix $U^T(\Delta X S^{opt} + X^{opt}\Delta S)U$; the right-hand side of (7) consists of $p \in \mathbb{R}^m$, $Q \in \mathcal{S}^l$, and the upper triangular part r of an $l \times l$-matrix. For brevity we write $r \in \mathbb{R}^n$.

We eliminate the variable Δy from the second equation of (7). To this end let $\mathcal{F} : \mathcal{S}^l \to \mathbb{R}^{n-m}$ be a linear operator of full rank such that $\mathcal{F}(\mathcal{A}^*(y)) = 0$ for all $y \in \mathbb{R}^m$. Let $q := \mathcal{F}(Q)$. By construction of $\mathcal{F}$, also the linear system

$$M \begin{pmatrix} \Delta X \\ \Delta S \end{pmatrix} := \begin{pmatrix} \mathcal{A}(\Delta X) \\ \mathcal{F}(\Delta S) \\ \Pi_{\text{up}}(U^T(\Delta X S^{opt} + X^{opt}\Delta S)U) \end{pmatrix} = \begin{pmatrix} p \\ q \\ r \end{pmatrix} \tag{8}$$

has full rank. Here, $(p^T, q^T)^T \in \mathbb{R}^n$ and $r \in \mathbb{R}^n$.

First note that $\|(p^T, q^T, r^T)^T\| = \|M\Delta Z\| \geq 1/\|M^{-1}\|_2$ since $\|\Delta Z\| = 1$. From $Z^{opt} + \Delta Z \in \boldsymbol{A}$ it follows that $p = 0$ and $q = 0$. Hence, $\|r\|_2 \geq 1/\|M^{-1}\|_2$.

Step 2: We now perform a change of basis.

Note that problems (P) and (D) remain invariant when replacing B with X^{opt} and C with S^{opt}. Hence, from $Z^{opt} + \Delta Z \in \boldsymbol{A}$ it follows that

$$\begin{aligned} 0 &= C \bullet \Delta X + B \bullet \Delta S \\ &= S^{opt} \bullet \Delta X + X^{opt} \bullet \Delta S. \end{aligned} \tag{9}$$

Let $\widetilde{\Delta X} := U^T \Delta X U$ and $\widetilde{\Delta S} := U^T \Delta S U$. The last equation in (8) then states that

$$\Pi_{\text{up}}(\widetilde{\Delta X}\Sigma + \Lambda\widetilde{\Delta S}) = r, \tag{10}$$

while relation (9) and $UU^T = I$ imply that

$$0 = \widetilde{\Delta X} \bullet \Sigma + \widetilde{\Delta S} \bullet \Lambda. \tag{11}$$

As U is unitary, $\|\widetilde{\Delta X}\|_F = \|\Delta X\|_F$, $\|\widetilde{\Delta S}\|_F = \|\Delta S\|_F$.

We partition $\widetilde{\Delta X}$ conforming with the zero-structure of Λ and Σ,

$$\widetilde{\Delta X} = \begin{pmatrix} \widetilde{\Delta X}_{11} & \widetilde{\Delta X}_{12} \\ \widetilde{\Delta X}_{12}^T & \widetilde{\Delta X}_{22} \end{pmatrix},$$

where $\widetilde{\Delta X}_{11} \in \mathcal{S}^k$ and $\widetilde{\Delta X}_{22} \in \mathcal{S}^{l-k}$. Likewise we partition $\widetilde{\Delta S}$.

Step 3: Let

$$\epsilon := \min\{\lambda_k, \sigma_{k+1}, 1/\lambda_1, 1/\sigma_l\} \qquad \text{and} \qquad \mu := \epsilon/(2\|M^{-1}\|_2).$$

We show that $\phi(Z^{opt}+\lambda\Delta Z)$ grows at least as $\frac{1}{2}(\lambda\epsilon^2\mu/2n^2)^2$ whenever $\|\widetilde{\Delta X}_{22}\|_F \geq \mu$ or $\|\widetilde{\Delta S}_{11}\|_F \geq \mu$.

Assume that $\|\widetilde{\Delta X}_{22}\|_F \geq \mu$. We distinguish two cases:

1. The maximum diagonal element of $\widetilde{\Delta X}_{22}$ is at least $\mu/2n$. It then follows from (11) and the definition of ϵ that the smallest diagonal element (and hence the smallest eigenvalue) of the 2×2-block matrix $\mathrm{Diag}(\widetilde{\Delta X}_{22}, \widetilde{\Delta S}_{11})$ is less than or equal to $-\epsilon^2\mu/2n^2$. (Straightforward proof by contradiction.)

2. The maximum diagonal element of $\widetilde{\Delta X}_{22}$ is less than $\mu/2n$. Since $\|\widetilde{\Delta X}_{22}\|_F \geq \mu$ there is an element of $\widetilde{\Delta X}_{22}$ of absolute value at least μ/n. If this is a diagonal element, it must be negative, and in particular, the most negative diagonal element is less than $-\epsilon^2\mu/2n^2$ as in Case 1 above. If it is an off-diagonal element, the associated 2 by 2 submatrix of $\widetilde{\Delta X}_{22}$ has diagonal elements $\leq \mu/2n$ and off-diagonal elements with absolute value $\geq \mu/n$. Straightforward calculations show that it therefore must have an eigenvalue $\leq -\mu/2n$. By the interlacing property, $\widetilde{\Delta X}_{22}$ has an eigenvalue less than or equal to $-\mu/2n$.

Thus, in both cases, the distance of $Z^{opt} + \lambda\Delta Z$ to $\boldsymbol{C}$ is at least $\lambda\epsilon^2\mu/2n^2$, and the function ϕ grows quadratically with λ.

The same argument holds when $\|\widetilde{\Delta S}_{11}\|_F \geq \mu$.

Step 4: Now assume that $\|\widetilde{\Delta X}_{22}\|_F < \mu$ and $\|\widetilde{\Delta S}_{11}\|_F < \mu$, so that Step 3 cannot be applied. In this case we show that $\hat{f}(Z^{opt} + \lambda\Delta Z)$ locally grows at least as $\lambda^2/2\|M^{-1}\|_2^2$.

From (10) follows

$$\|r\|^2 \leq \|\Lambda_{11}\widetilde{\Delta S}_{11}\|_F^2 + \|\widetilde{\Delta X}_{12}\Sigma_{22} + \Lambda_{11}\widetilde{\Delta S}_{12}\|_F^2 + \|\widetilde{\Delta X}_{22}\Sigma_{22}\|_F^2. \tag{12}$$

The inequality $\|r\|_2 \geq 1/\|M^{-1}\|_2$, relation (12), and the definition of ϵ imply

$$\|\widetilde{\Delta X}_{12}\Sigma_{22} + \Lambda_{11}\widetilde{\Delta S}_{12}\|_F^2 \geq 1/\|M^{-1}\|_2^2 - 2\mu^2/\epsilon^2 \geq 1/2\|M^{-1}\|_2^2. \tag{13}$$

Observe that

$$\begin{aligned} \hat{f}(Z^{opt} + \lambda\Delta Z) &= \|(\Lambda + \lambda\widetilde{\Delta X})(\Sigma + \lambda\widetilde{\Delta S}) - (\Sigma + \lambda\widetilde{\Delta S})(\Lambda + \lambda\widetilde{\Delta X})\|_F^2 \\ &= \lambda^2(\|\Lambda\widetilde{\Delta S} + \widetilde{\Delta X}\Sigma - \Sigma\widetilde{\Delta X} - \widetilde{\Delta S}\Lambda\|_F^2) + O(\lambda^3) \\ &\geq 2\lambda^2\|\Lambda_{11}\widetilde{\Delta S}_{12} + \widetilde{\Delta X}_{12}\Sigma_{22}\|_F^2 + O(\lambda^3). \end{aligned}$$

For small λ, the third-order term is dominated, and

$$\hat{f}(Z^{opt} + \lambda\Delta Z) \geq \lambda^2\|\Lambda_{11}\widetilde{\Delta S}_{12} + \widetilde{\Delta X}_{12}\Sigma_{22}\|_F^2 \geq \lambda^2/2\|M^{-1}\|_2^2,$$

which completes the proof. ∎

6. Cheap computation of the projection onto $\boldsymbol{A}$. First note that a projection onto an $n-1$-dimensional affine subspace of $\mathbb{R}^{2n}$ can (after an initial factorization of the projection matrix) generally be done in order n^2 operations. To make our algorithm practical, we show that it can be done in a cheaper way for the particular sets $\boldsymbol{A}$ arising in linear programming. (Of course, the same reasoning applies to semidefinite programming replacing A^T with the adjoint $\mathcal{A}^*$.)

The computation of the projection below is closely related to rank-one update formulae for inverse matrices. There are two differences: We update a projection rather than an inverse matrix, and the matrix defining the projection is never explicitly formed. (The matrix defining the projection may be nonsparse while A and the Cholesky factor of AA^T used below may be sparse.)

We assume that $\mathcal{L}+b = \{x \mid Ax = Ab\} \subset \mathbb{R}^n$ where A has full row rank. Let a point $x \in \mathbb{R}^n$ be given. Then it is easy to verify that

$$x - \Pi_{\mathcal{L}+b}(x) = A^T(AA^T)^{-1}A(x-b) \quad \text{and} \quad \|x-\Pi_{\mathcal{L}+b}(x)\|_2^2 = (x-b)^T A^T(AA^T)^{-1}A(x-b).$$

Likewise, for $s \in \mathbb{R}^n$ we have

$$s - \Pi_{\mathcal{L}^\perp+c}(s) = (I - A^T(AA^T)^{-1}A)(s-c)$$

and

$$\|s - \Pi_{\mathcal{L}^\perp+c}(s)\|_2^2 = (s-c)^T(I - A^T(AA^T)^{-1}A)(s-c).$$

The factorization of AA^T can be computed once in a preprocessing stage at the beginning of Algorithm 2 and can then be used without modification throughout. It is the same matrix that is usually factored in interior-point methods. For semidefinite programs, however, the factorization of $\mathcal{A}\mathcal{A}^*$ may be substantially cheaper than the systems factored at each iteration of an interior point algorithm; prime examples are semidefinite programs arising from the semidefinite relaxation [12] of the max clique problem [10] that results in the factorization of dense matrices in interior point methods while $\mathcal{A}\mathcal{A}^*$ is a *diagonal* matrix.

After this preprocessing the projection of a point $z = (x; s)$ onto

$$\boldsymbol{A_1} := (\mathcal{L}+b) \times (\mathcal{L}^\perp + c)$$

can be computed (separately for x and s) in order mn operations, namely two backsolves using the factorization of AA^T and two matrix vector products of the form Ax, two matrix products of the form A^Ty, as well as some order-n-operations.

Let

$$\boldsymbol{A_2} := \{(x; s) \mid \ \langle c, x\rangle + \langle b, s\rangle = \langle b, c\rangle\},$$

so that $\boldsymbol{A} = \boldsymbol{A_1} \cap \boldsymbol{A_2} \neq \emptyset$. We now compute the projection onto $\boldsymbol{A}$ given the projection onto $\boldsymbol{A_1}$. To this end first observe that for any affine space $\boldsymbol{A_1} = \boldsymbol{L_1} + \boldsymbol{b_1}$ and any hyperplane $\boldsymbol{A_2}$ written in the form $\boldsymbol{A_2} = \{z \mid \langle a, z\rangle = \alpha\}$ we have

$$\begin{aligned}
\boldsymbol{A_1} \cap \boldsymbol{A_2} &= \boldsymbol{L_1} + \boldsymbol{b_1} \cap \{z \mid \langle a, z\rangle \qquad = \alpha\} \\
&= (\boldsymbol{L_1} \qquad \cap \{z \mid \langle a, z\rangle \qquad = \alpha - \langle a, \boldsymbol{b_1}\rangle\}) + \boldsymbol{b_1} \\
&= (\boldsymbol{L_1} \qquad \cap \{z \mid \langle a, \Pi_{\boldsymbol{L_1}} z\rangle = \alpha - \langle a, \boldsymbol{b_1}\rangle\}) + \boldsymbol{b_1} \\
&= (\boldsymbol{L_1} \qquad \cap \{z \mid \langle \Pi_{\boldsymbol{L_1}} a, z\rangle = \alpha - \langle a, \boldsymbol{b_1}\rangle\}) + \boldsymbol{b_1} \\
&= \boldsymbol{L_1} + \boldsymbol{b_1} \cap \{z \mid \langle \Pi_{\boldsymbol{L_1}} a, z\rangle = \alpha - \langle a - \Pi_{\boldsymbol{L_1}} a, \boldsymbol{b_1}\rangle\},
\end{aligned}$$

where the third line is trivially true for $z \in \boldsymbol{L_1}$ and the fourth line holds since $\Pi_{\boldsymbol{L_1}}$ is symmetric. Hence, in a first step we may project a onto $\boldsymbol{L_1}$ and update α. In our case, we project b onto $\mathcal{L}^\perp$ and c onto $\mathcal{L}$ and replace $\langle b, c\rangle$ with 0. Just like the factorization of AA^T, this form of preprocessing is done only once for the algorithm; see section 7.1.

We therefore assume without loss of generality that $\boldsymbol{A_1} \perp \boldsymbol{A_2}$. In this case, the projection onto $\boldsymbol{A_1} \cap \boldsymbol{A_2}$ is particularly simple; it is the projection onto $\boldsymbol{A_1}$ followed by the projection onto $\boldsymbol{A_2}$. (The order in which the projections are applied plays no role.) We recall that the projection onto a hyperplane $\boldsymbol{A_2}$ can be carried out in order n operations,

$$z \mapsto z + \frac{\alpha - \langle a, z\rangle}{\langle a, a\rangle} a.$$

As we have just seen, the projection onto $\boldsymbol{A}$ at each iteration of Algorithm 2 takes two back-solves with the factorization of AA^T, two matrix vector multiplications Ax, and two matrix vector multiplications $A^T y$.

Next, we show by adding slack variables that the intersection of two cones can be decomposed into the Cartesian product of two cones at the expense of additional linear equations. These additional linear equations merely double the computational effort of the projection onto the affine manifold. (This is in contrast to interior point methods where additional inequalities for the Lovász–Schrijver relaxation of the max-clique problem imply a very substantial increase in computation time at each iteration.)

6.1. The intersection of two cones. A very rewarding application targeted by the apd approach is the Lovász relaxation of the max-clique problem for which the matrix $\mathcal{A}\mathcal{A}^*$ is a diagonal matrix, while interior point methods factor a full matrix of the same size at each iteration. The Lovász–Schrijver relaxation is a sharper relaxation for which the semidefinite cone is replaced with the intersection of the semidefinite cone and the cone of matrices with nonnegative entries. Unfortunately, while the projection onto either of the two cones is straightforward, the projection onto their intersection is less trivial. We therefore present an approach that allows the application to problems of the form

$$(\hat{P}) \qquad \text{minimize } \langle c, x\rangle \ \text{ s.t. } \ x \in K \cap \hat{K} \cap (\mathcal{L} + b),$$

where K and $\hat{K}$ are both pointed closed convex cones such that the interior of $K \cap \hat{K}$ is nonempty. Again, we assume that $\mathcal{L} + b$ is given by a set of linear equations $Ax = \bar{b}$ for which a factorization of AA^T is computed once, and that projections onto K and $\hat{K}$ are easy to compute.

Problem $(\hat{P})$ is equivalent to

$$\text{minimize } \left\langle \begin{pmatrix} c \\ 0 \end{pmatrix}, \begin{pmatrix} x \\ \hat{x} \end{pmatrix} \right\rangle \ \text{ s.t. } \ \begin{pmatrix} x \\ \hat{x} \end{pmatrix} \in (K \times \hat{K}) \cap \left(\hat{\mathcal{L}} + \begin{pmatrix} b \\ b \end{pmatrix} \right),$$

where

$$\hat{\mathcal{L}} + \begin{pmatrix} b \\ b \end{pmatrix} =: \left\{ \begin{pmatrix} x \\ \hat{x} \end{pmatrix} \ \middle| \ Ax = \bar{b}, \ \ x = \hat{x} \right\}.$$

This is a problem of the form (P). By our assumption, projections onto $K \times \hat{K}$—and hence also projections onto its dual—are easy to compute. Thus, in order to apply

the apd algorithm it suffices to verify that projections onto $\hat{\mathcal{L}}$ are easily computable given a factorization of AA^T.

This, however, is readily seen as

$$\begin{pmatrix} A & 0 \\ I & -I \end{pmatrix} \begin{pmatrix} A^T & I \\ 0 & -I \end{pmatrix} = \begin{pmatrix} AA^T & A \\ A^T & 2I \end{pmatrix} = \begin{pmatrix} 2(AA^T)^{-1} & -(AA^T)^{-1}A \\ -A^T(AA^T)^{-1} & \frac{1}{2}(I + A^T(AA^T)^{-1}A) \end{pmatrix}^{-1}$$

provides the desired factorization.

7. Numerical results. Algorithm 2 has been implemented in Matlab and tested on some randomly generated linear semidefinite programs (SDP) as well as on some SDP coming from combinatorial optimization. Initially, the algorithm can be applied to minimize the function $\tilde{\phi}$ for $z \in \boldsymbol{A}$, and when this minimization slows down, a Phase 2 is started, where a "regularizing" function $\hat{f}$ is added to $\tilde{\phi}$.

Under standard assumptions, Theorem 1 guarantees that the term $\|XS - SX\|_F^2$ may serve as a regularizing function. Note that also the term $\|XS + SX\|_F^2$ is minimized at the optimal solution of the problem (P). Thus, at the point Z^{opt} it has nonnegative curvature as well and hence, Theorem 1 also applies to the function

$$\hat{f}(X,S) := \|XS\|_F^2 = \frac{1}{4}(\|XS - SX\|_F^2 + \|XS + SX\|_F^2).$$

This term yielded the best numerical results in our examples, and the results listed below refer to this regularizing term – while Theorem 1 is proved under slightly weaker conditions (namely just the term $\|XS - SX\|_F^2$).

7.1. Rescaling. We emphasize that Algorithm 2 is essentially a first-order method, and hence, it is sensitive to scaling of the data. Even for data that "looks nice" (all data integers of absolute value less than 10), the following rescaling may turn out to be crucial:

First, replace b with $\Pi_{\mathcal{L}^\perp} b$. (The set $\mathcal{L} + b$ remains invariant with this change!) Likewise, replace c with $\Pi_{\mathcal{L}} c$. Then set $b = b/\|b\|_2$, $c = c/\|c\|_2$, and rescale x, s accordingly. Note that by this normalization, the duality simplifies to $\langle b, s\rangle + \langle c, x\rangle = 0$, and in particular, the set $\boldsymbol{A_2}$ now is a linear subspace perpendicular to $\boldsymbol{A_1}$.

Moreover, the origin $x = 0$ has distance exactly 1 from $\mathcal{L} + b$, and likewise $s = 0$ has distance exactly 1 from $\mathcal{L}^\perp + c$. For a semidefinite program, the point $x^{(0)} = s^{(0)} = I/\sqrt{n}$ is a canonical starting point: Its duality gap satisfies $\langle x^{(0)}, s^{(0)}\rangle = 1$, and the distance of $x^{(0)}$ from $\mathcal{L} + b$ is bounded by 2, same as the distance of $s^{(0)}$ from $\mathcal{L}^\perp + c$.

While the above rescaling of b and c appears to be natural, it is certainly far from optimal. When convergence slows down, it may be possible to identify a more suitable scaling based on the current iterate. The numerical results below simply refer to the above scaling.

7.2. Preconditioning. We point out that the above rescaling may be generalized slightly. Indeed, let M be a nonsingular matrix, then the preconditioning $X \to MXM^T$, $B \to MBM^T$, $\mathcal{L} \to M\mathcal{L}M^T$ and $S \to M^{-T}SM^{-1}$, $C \to M^{-T}CM^{-1}$ results in an equivalent semidefinite program, and the solution of either program can easily be recovered from the solution of the other. Of course, the functions $\tilde{\phi}$ and $\hat{f}$ change when replacing X, S with $MXM^T, M^{-T}SM^{-1}$, and thus the performance of Algorithm 2 will vary. It is still an open question how to determine suitable scalings that accelerate Algorithm 2. When M is a diagonal matrix, the projections onto

TABLE 1

Randomly generated SDP. The column labeled apd contains the function value after 50 *iterations of our augmented primal-dual method. The normalized primal and dual errors, multiplied by* 1000 *are given in columns* 4 *and* 5*, and the column labeled mprw provides the value computed by the boundary point method from* [13].

n	m	seed	$10^3 \cdot err_P$	$10^3 \cdot err_D$	apd	mprw
400	30000	4003030	−0.11	−0.12	1072.06	1072.14
500	30000	5003030	−0.09	−0.27	1108.21	1107.63
600	40000	6004030	−0.08	−0.23	307.45	306.62
700	50000	7005030	−0.14	−0.33	315.48	313.20
800	70000	8007030	−0.08	−0.22	2332.98	2331.39
900	100000	90010030	−0.06	−0.15	955.33	954.22
1000	100000	100010030	−0.11	−0.23	3099.51	3096.36

$M\mathcal{L}M^T$ and its orthogonal complement can be performed just as cheaply as for $\mathcal{L}$ and $\mathcal{L}^\perp$.

Likewise, one may look at preconditionings of the form $\|XS\|_F^2 \to \|MXS\tilde{M}\|_F^2$ for some nonsingular $M, \tilde{M}$. Here, the function $\tilde{\phi}$ is not changed, and here as well, the selection of suitable preconditionings is subject to further research.

7.3. Preliminary computational results. To give some impressions of the practical behavior of our approach, we provide some computational results both on randomly generated SDP used in [13], and on instances from the DIMACS collection [9], related to the Lovász theta number in graphs.

We first give a short description of the random instances from [13]. The linear constraints are generated to have a sparse Cholesky factor of AA^T. To achieve this, each A_i is generated to have nonzero support only on a submatrix of small order. Then a positive definite matrix X is generated, defining $b := A(X)$. On the dual side, the selection of y and $S \succ 0$ gives $C = A^T(y) + S$. This insures strong duality. The generator is written in MATLAB and is available at http://www.math.uni-klu.ac.at/or/Software.

In Table 1 we provide some preliminary computational results. The parameters n and m indicate the size of the problem as defined before. The parameter "seed" is used to initialize the random number generator and makes the instances reproducible using MATLAB. The column "mprw" contains the optimal value of the SDP, computed with a relative error tolerance of 10^{-7}. These values can therefore be considered as reasonably accurate. We provide the objective value of our approach (in column "apd") after 50 function evaluations. We have used a "quick-and-dirty" implementation of our approach, without any parameter tuning. Therefore we stop after a preset number of iterations to give a first impression of the potential of our approach. We reach optimality, once the smallest eigenvalues of X and S are nonnegative. Therefore we also provide the following (relative) error measures in the table.

$$err_P := \frac{\lambda_{\min}(X)}{1 + \|X\|}, \quad err_D := \frac{\lambda_{\min}(S)}{1 + \|S\|}.$$

We note that in all these instances, the most negative normalized eigenvalue (which keeps us away from optimality) is of order 10^{-4}.

The second table contains results on computing the Lovász theta number of a graph G, given through its edge list. The problem is of the form:

$$\max \sum_{ij} x_{ij} \text{ such that } x_{ij} = 0 \forall i \neq j, [ij] \notin E(G), \ \operatorname{trace}(X) = 1, \ X \succeq 0.$$

TABLE 2
The theta number for some DIMACS graphs. The first column gives the name of the graph.

name	n	m	$10^3 \cdot err_P$	$10^3 \cdot err_D$	apd	mprw
brock400-1	400	20078	−0.30	−0.57	39.67	39.70
p-hat500-1	500	93182	−0.36	−0.68	13.13	13.07
keller5	776	74711	−0.96	−1.66	30.69	31.00
brock800-1	800	112096	−0.26	−0.33	42.19	42.22
p-hat1000-3	1000	127755	−2.15	−8.61	77.83	84.80

Here $E(G)$ denotes the edge set of G. In this SDP, one therefore asks to have a zero in each position of X outside the main diagonal corresponding to non-edges. In addition, the trace of X should be one. These constraints have the nice feature that AA^T is in fact diagonal. In Table 2 we include as before the total number of equations in the column labeled m. We note that again after 50 iterations, we get rather good approximations of the theta number in all instances except keller5 and p-hat1000-3. For these two instances, the most negative eigenvalues are much bigger (in absolute value), as in the other cases. This would indicate that further iterations are necessary to come closer to optimality. Indeed, running the keller5 instance for 150 iterations, we obtain a value of 30.98 with relative errors -0.49e-3 and -0.58e-3. The theta number of these instances was only recently computed in [13]; the resulting SDPs seem to be beyond the capabilities of standard SDP solvers.

A final word on computation times and number of iterations is in order. First, the number of iterations of "mprw" to reach the required accuracy level is typically around 200 for the instances in the first table. For the instances in Table 2, the number of iterations varies greatly, depending on the instance. It took about 500 iterations for keller5 and more than 1000 for the last instance. So it should come as no surprise that our approach also has a harder time on these instances. We should also mention that the computational effort for a single iteration of our method is at least twice the effort of the boundary point method from [13], because we project X and S individually, while in the boundary point method, only one projection is necessary.

All computations were done on a Pentium IV (2.1 Ghz, 2G memory) using Matlab. It took about 45 minutes for the largest instance, and a few minutes for the smallest one. Since this is a preliminary implementation, we expect that there is quite a bit of room for improvement. The present paper sets the theoretical stage for the new approach. A competitive implementation is beyond the scope of the current paper and will be presented in a separate study.

8. Concluding remarks. This paper proposes a reformulation of a linear program over a convex cone into the problem of minimizing a differentiable convex apd-function in a certain primal-dual space. The apd-function is related to the augmented Lagrangian function, but is slightly less data dependent. For large classes of conic programs including linear, semidefinite, and SOC problems, its function and gradient evaluations are rather cheap. For the case of a semidefinite program, a certain regularization of the apd-function is analyzed. Numerical examples minimizing the function with a conjugate gradient method illustrate the potential of the approach. Extensions for the case that Slater's condition is not satisfied and to other cones are the subject of future research.

Acknowledgments. The authors thank the anonymous referees for their comments that helped to improve the presentation of the paper.

REFERENCES

[1] H.H. Bauschke, P.L. Combettes, and S.G. Kruk, *Extrapolation algorithm for affine-convex feasibility problems*, Numer. Algorithms, 41 (2006), pp. 239–274.

[2] S. Burer and R.D.C Monteiro, *A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization*, Math. Program. (Series B), 95 (2003), pp. 329–357.

[3] W. Burmeister, *Die Konvergenzordnung des Fletcher-Powell algorithmus*, Z. Angew. Math. Mech., 53 (1973), pp. 693–699.

[4] A. Cohen, *Rate of convergence of several conjugate gradient algorithms*, SIAM J. Numer. Anal., 20 (1983), pp. 187–209.

[5] A. Fiacco, *Introduction to Sensitivity and Stability Analysis*, in Nonlinear Programming (Mathematics in Science and Engineering), Academic Press Inc., 1983.

[6] R.W. Freund and F. Jarre, *A sensitivity result for semidefinite programs*, Oper. Res. Lett. 32 (2004), pp. 126–132.

[7] K. Hauk and F. Jarre, *Linear programs and implicit functions*, Pac. J. Optim. 3 (2007), pp. 53–72.

[8] C. Helmberg and F. Rendl, *A spectral bundle method for semidefinite programming*, SIAM J. Optim., 10 (2000), pp. 673–696.

[9] D.S. Johnson and M. Trick (eds.), *Cliques, Colorings and Satisfiability: Second DIMACS implementation challenge*, American Mathematical Society, Providence, RI, 1996.

[10] R. Karp, *Reducibility Among Combinatorial Problems*, in Proceedings of a Symposium on the Complexity of Computer Computations, Plenum Press, New York,1972.

[11] M. Kocvara and M. Stingl, *On the solution of large-scale sdp problems by the modified barrier method using iterative solvers*, Math. Program., 109 (2007), pp. 413–444.

[12] L. Lovász and A. Schrijver, *Matrix cones, projection representations, and stable set polyhedra*, in Polyhedral Combinatorics, DIMACS Series in Discrete Mathematics and Theoretical Computer Science I, 1990, pp. 1–17.

[13] J. Malick, J. Povh, F. Rendl, and A. Wiegele, *Regularization methods for semidefinite programming*, Technical report, University of Klagenfurt, 2007; available online from http://www.optimization-online.org/DB_HTML/2007/10/1800.html.

[14] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia, 1994.

[15] J. Nocedal and S.J. Wright, *Numerical Optimization*, Springer Series in Operations Research, Springer, New York, (1999).

[16] E. Polak and G. Ribière, *Note sur la convergence de méthodes directions conjugées*, Rev. Francaise d'Informatique et de Recherche Opérationelle, 16 (1969), pp. 35–43.

[17] M.J.D. Powell, *Nonconvex minimization calculations and the conjugate gradient method*, In Numerical Analysis, Lecture Notes in Mathematics 1066, Griffiths, ed., Springer-Verlag, Berlin, 1984, pp. 122–141.

[18] J. Povh, F. Rendl, and A. Wiegele, *Boundary point method to solve semidefinite programs*, Computing, 78 (2006), pp. 277–286.

[19] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Program. 58 (1993), pp. 353–367.

[20] J. Sun and D. Sun, *Semismooth matrix-valued functions*, Math. Oper. Res., 27 (2002), pp. 150–169.

[21] D. Sun and J. Sun, *Strong semismoothness of eigenvalues of symmetric matrices and its application to inverse eigenvalue problems*, SIAM J. Numer. Anal., 40 (2002), pp. 2352–2367.

[22] K.C. Toh, *Solving large scale semidefinite programs via an iterative solver on the augmented systems*, SIAM J. Optim., 14 (2004), pp. 670–698.

SIAM J. OPTIM.
Vol. 19, No. 2, pp. 824–835

EXISTENCE AND APPROXIMATION OF FIXED POINTS OF FIRMLY NONEXPANSIVE-TYPE MAPPINGS IN BANACH SPACES*

FUMIAKI KOHSAKA† AND WATARU TAKAHASHI‡

Abstract. A class of nonlinear operators in Banach spaces is proposed. We call each operator in this class a firmly nonexpansive-type mapping. This class contains the classes of firmly nonexpansive mappings in Hilbert spaces and resolvents of maximal monotone operators in Banach spaces. We study the existence and approximation of fixed points of firmly nonexpansive-type mappings in Banach spaces.

Key words. Banach space, convex optimization, firmly nonexpansive mapping, firmly nonexpansive-type mapping, fixed point theorem, monotone operator, nonexpansive mapping, proximal point algorithm

AMS subject classifications. 47H10, 47H05, 47J25

DOI. 10.1137/070688717

1. Introduction.

Let C be a nonempty closed convex subset of a (real) Hilbert space H. Then a mapping $T : C \to C$ is said to be *nonexpansive* if $\|Tx-Ty\| \le \|x-y\|$ for all $x, y \in C$. The mapping T is also said to be *firmly nonexpansive* if

$$\|Tx - Ty\|^2 \le \langle x - y, Tx - Ty\rangle \tag{1.1}$$

for all $x, y \in C$; see Bruck and Reich [7], Goebel and Kirk [14], and Goebel and Reich [15]. It is known that a mapping $T : C \to C$ is firmly nonexpansive if and only if

$$\|Tx - Ty\|^2 + \|(I - T)x - (I - T)y\|^2 \le \|x - y\|^2 \tag{1.2}$$

for all $x, y \in C$, where I is the identity operator on H. Martinet [23] showed that if C is a nonempty bounded closed convex subset of H and $T : C \to C$ is a firmly nonexpansive mapping, then for all $x \in C$, $\{T^n x\}$ converges weakly to an element of $F(T)$, where $F(T)$ is the set of fixed points of T. Since every firmly nonexpansive mapping T is *asymptotically regular*, that is, $\|T^{n+1}x - T^n x\| \to 0$ for all $x \in C$, the result of Martinet is a corollary of Opial's theorem [28]; see Goebel and Kirk [14] and Takahashi [41] on fixed point theory for nonexpansive mappings.

Fixed point theory for nonexpansive mappings can be applied to the problem of finding a point $u \in H$ satisfying

$$0 \in Au, \tag{1.3}$$

where $A : H \to 2^H$ is a maximal monotone operator defined in a Hilbert space H. This problem is related to convex optimization problems, minimax problems, variational inequality problems, equilibrium problems, and so on. We always identify

*Received by the editors April 19, 2007; accepted for publication (in revised form) February 15, 2008; published electronically August 1, 2008.
http://www.siam.org/journals/siopt/19-2/68871.html
†Department of Information Environment, Tokyo Denki University, Muzai Gakuendai, Inzai, Chiba, 270-1382, Japan (kohsaka@sie.dendai.ac.jp).
‡Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, Ohokayama, Meguro-ku, Tokyo, 152-8552, Japan (wataru@is.titech.ac.jp).

the set-valued mapping A with its graph $G(A) = \{(x, x^*) : x^* \in Ax\}$. Thus we write $A : H \to 2^H$ as follows: $A \subset H \times H$. By Browder [6] and Rockafellar [37], the maximal monotonicity of A implies that $R(I + rA) = H$ for all $r > 0$. Thus, for all $r > 0$, we can define the resolvent J_r of A by

$$J_r x = \{z \in H : x \in z + rAz\} = (I + rA)^{-1}(x) \tag{1.4}$$

for all $x \in H$. It is well known that $J_r : H \to H$ is a single-valued firmly nonexpansive mapping and (1.3) is equivalent to $J_r u = u$; see Rockafellar [38] and Takahashi [40, 41]. Therefore the problem (1.3) is reduced to a fixed point problem for a firmly nonexpansive mapping in the setting of Hilbert spaces.

In contrast to the case of Hilbert spaces, the resolvent of a maximal monotone operator is not generally a nonexpansive mapping in the case of Banach spaces. Recently, Matsushita and Takahashi [24, 25] introduced the class of *relatively nonexpansive mappings* in Banach spaces. They proved that if $A \subset E \times E^*$ is a maximal monotone operator defined in a strictly convex and uniformly smooth Banach space E such that $A^{-1}0$ is nonempty, $J : E \to E^*$ is the normalized duality mapping, $r > 0$, and $J_r = (J + rA)^{-1}J$, then J_r is a relatively nonexpansive mapping from E onto the domain $D(A)$ of A. They obtained weak and strong convergence theorems for a single relatively nonexpansive mapping in Banach spaces. See also Butnariu, Reich, and Zaslavski [9, 10], Censor and Reich [12], and Reich [33] for similar classes of nonlinear operators in Banach spaces.

The purpose of the present paper is to study the existence and approximation of fixed points of *firmly nonexpansive-type mappings* in Banach spaces. The class of firmly nonexpansive-type mappings contains the classes of firmly nonexpansive mappings in Hilbert spaces and resolvents of maximal monotone operators in Banach spaces. As we see below, the class of firmly nonexpansive-type mappings that have fixed points is contained in the class of strongly relatively nonexpansive mappings. Let E be a smooth Banach space, C be a nonempty closed convex subset of E, and J be the normalized duality mapping from E into E^*. Then we say that T is of *firmly nonexpansive type* if

$$\langle Tx - Ty, JTx - JTy\rangle \le \langle Tx - Ty, Jx - Jy\rangle \tag{1.5}$$

for all $x, y \in C$. Note that we do not assume the existence of fixed points of T. If E is a Hilbert space, then J is the identity operator on H, and hence (1.5) is reduced to (1.1). Since J is a monotone operator, every firmly nonexpansive-type mapping satisfies $\langle Tx - Ty, Jx - Jy\rangle \ge 0$ for all $x, y \in C$; that is, it is d-accretive in the sense of Alber and Reich [2].

Our paper is organized as follows: Section 2 is devoted to preliminaries. In section 3, we obtain a fixed point theorem for firmly nonexpansive-type mappings in Banach spaces (Theorem 3.2). In section 4, we prove some lemmas needed in section 5. In section 5, we show that every firmly nonexpansive-type mapping which has a fixed point is strongly relatively nonexpansive (Theorem 5.2) and then obtain a weak convergence theorem (Theorem 5.3). In section 6, we apply the obtained results to the proximal point algorithm for a monotone operator satisfying the range condition in Banach spaces.

2. Preliminaries. Throughout the present paper, every Banach space is real. Let $\mathbb{N}$ and $\mathbb{R}$ denote the sets of positive integers and real numbers, respectively. For a sequence $\{x_n\}$ in a Banach space E, the strong convergence and the weak convergence

of $\{x_n\}$ to $x \in E$ are denoted by $x_n \to x$ and $x_n \rightharpoonup x$, respectively. For a sequence $\{x_n^*\}$ of the dual space E^* of E, the weak* convergence of $\{x_n^*\}$ to $x^* \in E^*$ is also denoted by $x_n^* \overset{*}{\rightharpoonup} x^*$.

Let E be a Banach space and let $J : E \to E^*$ be the *normalized duality mapping* defined by

$$Jx = \{x^* \in E^* : \langle x, x^* \rangle = \|x\|^2 = \|x^*\|^2\} \tag{2.1}$$

for all $x \in E$. Let $S(E)$ be the unit sphere centered at the origin of E. Then the space E is said to be *smooth* if the limit

$$\lim_{t \to 0} \frac{\|x + ty\| - \|x\|}{t} \tag{2.2}$$

exists for all $x, y \in S(E)$. The space E is said to be *uniformly smooth* if the limit (2.2) converges uniformly in $x, y \in S(E)$. The norm of E is also said to be *uniformly Gâteaux differentiable* if for all $y \in S(E)$, the limit (2.2) converges uniformly in $x \in S(E)$. A Banach space E is said to be *strictly convex* if $\|(x+y)/2\| < 1$ whenever $x, y \in S(E)$ and $x \neq y$. The space E is also said to be *uniformly convex* if for all $\varepsilon \in (0, 2]$, there exists $\delta > 0$ such that $x, y \in S(E)$ and $\|x - y\| \geq \varepsilon$ imply $\|(x + y)/2\| \leq 1 - \delta$. The duality mapping J from a smooth Banach space E into E^* is said to be *weakly sequentially continuous* if $Jx_n \overset{*}{\rightharpoonup} Jx$ whenever $x_n \rightharpoonup x$. We know the following; see Cioranescu [13], Reich [32], and Takahashi [41] on geometry of Banach spaces:

1. If E is smooth, then J is single-valued;
2. if E is reflexive, then J is onto;
3. if E is strictly convex, then J is one-to-one; that is, $Jx \cap Jy \neq \emptyset$ implies that $x = y$;
4. if E is strictly convex, then J is strictly monotone, that is, if (x, x^*), $(y, y^*) \in J$ and $\langle x - y, x^* - y^* \rangle = 0$, then $x = y$.

Let E be a smooth Banach space. Following Alber [1] and Kamimura and Takahashi [18], let $\phi : E \times E \to \mathbb{R}$ be the mapping defined by

$$\phi(x, y) = \|x\|^2 - 2\langle x, Jy \rangle + \|y\|^2 \tag{2.3}$$

for all $x, y \in E$. Note that ϕ is the *Bregman distance* corresponding to $\|\cdot\|^2$; see Bregman [4], Butnariu and Iusem [8], and Censor and Lent [11]. If E is a Hilbert space, then we have $\phi(x, y) = \|x - y\|^2$ for all $x, y \in E$. We know that

$$(\|x\| - \|y\|)^2 \leq \phi(x, y) \leq (\|x\| + \|y\|)^2 \tag{2.4}$$

for all $x, y \in E$. If E is strictly convex, then

$$\phi(x, y) = 0 \Longleftrightarrow x = y. \tag{2.5}$$

It is also well known that

$$\phi(x, y) = \phi(x, z) + \phi(z, y) + 2\langle x - z, Jz - Jy \rangle \tag{2.6}$$

for all $x, y, z \in E$. If E is a Hilbert space, then this equality is reduced to

$$\|x - y\|^2 = \|x - z\|^2 + \|z - y\|^2 + 2\langle x - z, z - y \rangle \tag{2.7}$$

for all $x, y, z \in E$. It is easy to see from (2.6) that

$$\langle x - y, Jz - Jw \rangle = \frac{1}{2} \{\phi(x, w) + \phi(y, z) - \phi(x, z) - \phi(y, w)\} \tag{2.8}$$

for all $x, y, z, w \in E$. It is also easy to see that if $\{x_n\}$ and $\{y_n\}$ are bounded sequences of a smooth Banach space E, then $\|x_n - y_n\| \to 0$ implies that $\phi(x_n, y_n) \to 0$. The converse is also true if E is additionally assumed to be uniformly convex.

LEMMA 2.1 (Kamimura and Takahashi [18]). *Let E be a smooth and uniformly convex Banach space and let $\{x_n\}$ and $\{y_n\}$ be sequences of E such that $\{x_n\}$ or $\{y_n\}$ is bounded. If $\phi(x_n, y_n) \to 0$, then $\|x_n - y_n\| \to 0$.*

Let E be a smooth, strictly convex and reflexive Banach space. A set-valued mapping $A \subset E \times E^*$ with domain $D(A) = \{x \in E : Ax \neq \emptyset\}$ and range $R(A) = \bigcup\{Ax : x \in D(A)\}$ is said to be *monotone* if $\langle x - y, x^* - y^* \rangle \geq 0$ whenever $(x, x^*), (y, y^*) \in A$. A monotone operator $A \subset E \times E^*$ is also said to be *maximal monotone* if $A = B$ whenever $B \subset E \times E^*$ is a monotone operator such that $A \subset B$. The following theorem is well known:

THEOREM 2.2 (Browder [6] and Rockafellar [37]; see also Barbu [3] and Takahashi [40]). *Let E be a smooth, strictly convex and reflexive Banach space and let $A \subset E \times E^*$ be a monotone operator. A is maximal monotone if and only if $R(J + rA) = E^*$ for all $r > 0$.*

Let E be a smooth, strictly convex and reflexive Banach space, let C be a nonempty closed convex subset of E, and let $A \subset E \times E^*$ be a monotone operator satisfying

$$D(A) \subset C \subset J^{-1}R(J + rA) \tag{2.9}$$

for all $r > 0$. In view of Theorem 2.2, if A is maximal monotone, then (2.9) holds for $C = \overline{D(A)}$. Note that the result due to Rockafellar [36] ensures that $\overline{D(A)}$ is closed and convex; see also Barbu [3], Reich [30] and Takahashi [41]. If A satisfies (2.9), then for all $r > 0$, we can define the resolvent $J_r : C \to D(A)$ of A by

$$J_r x = \{z \in E : Jx \in Jz + rAz\} \tag{2.10}$$

for all $x \in C$. In other words, $J_r x = (J + rA)^{-1} Jx$ for all $x \in C$. For all $r > 0$, the *Yosida approximation* $A_r : C \to E^*$ of A is also defined by $A_r x = (Jx - JJ_r x)/r$ for all $x \in C$. We also denote by $A^{-1}0$ the set $\{z \in D(A) : 0 \in Az\}$. We know the following; see, for instance, Butnariu and Iusem [8], Kamimura [16], Kohsaka and Takahashi [19] and Matsushita and Takahashi [24]:

1. J_r is a single-valued mapping from C into $D(A)$ such that

   $$\phi(u, J_r x) + \phi(J_r x, x) \leq \phi(u, x) \tag{2.11}$$

   for all $(x, u) \in C \times A^{-1}0$;
2. $(J_r x, A_r x) \in A$ for all $x \in C$;
3. $F(J_r) = A^{-1}0$, where $F(J_r)$ is the set of fixed points of J_r.

Let C be a nonempty closed convex subset of a smooth Banach space E and let $T : C \to C$ be a mapping. The set of fixed points of T is denoted by $F(T)$. The mapping T is said to be *nonexpansive* if $\|Tx - Ty\| \leq \|x - y\|$ for all $x, y \in C$. A point $u \in C$ is said to be an *asymptotic fixed point* of T if C contains a sequence $\{x_n\}$ such that $x_n \rightharpoonup u$ and $\|x_n - Tx_n\| \to 0$; see [33]. The set of asymptotic fixed points of T is denoted by $\widehat{F}(T)$. Following [24, 25], we say that a mapping $T : C \to C$ is *relatively nonexpansive* if the following conditions are satisfied:

1. $F(T)$ is nonempty;
2. $\phi(u, Tx) \leq \phi(u, x)$ for all $(x, u) \in C \times F(T)$;
3. $\widehat{F}(T) = F(T)$.

If C is a nonempty closed convex subset of a Hilbert space and $T : C \to C$ is a nonexpansive mapping such that $F(T)$ is nonempty, then T is relatively nonexpansive. Following Reich [33], we say that a mapping $T : C \to C$ is *strongly relatively nonexpansive* if the following conditions are satisfied:

1. T is relatively nonexpansive;
2. if $\{x_n\}$ is a bounded sequence of C such that

$$\phi(u, x_n) - \phi(u, Tx_n) \to 0 \tag{2.12}$$

for some $u \in F(T)$, then $\phi(Tx_n, x_n) \to 0$.

Examples of relatively or strongly relatively nonexpansive mappings can be found in Kohsaka and Takahashi [20, 21], Matsushita and Takahashi [24, 25] and Reich [33].

Let C be a nonempty closed convex subset of a smooth Banach space E. Then a mapping $T : C \to C$ is said to be of *firmly nonexpansive type* if (1.5) is satisfied.

LEMMA 2.3. *Let E be a smooth, strictly convex and reflexive Banach space, let C be a nonempty closed convex subset of E, and let $A \subset E \times E^*$ be a monotone operator satisfying* (2.9). *Let r be a positive real number and let $J_r x = (J + rA)^{-1} Jx$ for all $x \in C$. Then $J_r : C \to D(A)$ is of firmly nonexpansive type.*

Proof. Let $x, y \in C$ be given. Then we have $(J_r x, A_r x), (J_r y, A_r y) \in A$. Since A is monotone, we have

$$\frac{1}{r}\langle J_r x - J_r y, Jx - JJ_r x - (Jy - JJ_r y)\rangle \geq 0. \tag{2.13}$$

Thus we have

$$\langle J_r x - J_r y, JJ_r x - JJ_r y\rangle \leq \langle J_r x - J_r y, Jx - Jy\rangle. \tag{2.14}$$

This shows that J_r is of firmly nonexpansive type. □

If C is a nonempty closed convex subset of a smooth, strictly convex and reflexive Banach space E, then for all $x \in E$, there exists a unique $z \in C$ (denoted by $\Pi_C(x)$) such that

$$\phi(z, x) = \min_{y \in C} \phi(y, x). \tag{2.15}$$

The mapping Π_C is called the *generalized projection* from E onto C; see Alber [1] and Alber and Reich [2]; see also Kamimura and Takahashi [18].

LEMMA 2.4. *Let E be a smooth, strictly convex and reflexive Banach space, let C be a nonempty closed convex subset of E, and let Π_C be the generalized projection from E onto C. Then $\Pi_C : E \to C$ is of firmly nonexpansive type.*

Proof. Let i_C be the indicator function for C; that is, $i_C(x) = 0$ if $x \in C$ and ∞ otherwise. Then $i_C : E \to (-\infty, \infty]$ is a proper lower semicontinuous convex function. Rockafellar's maximal monotonicity theorem [34, 35] ensures that the subdifferential $\partial i_C \subset E \times E^*$ of i_C is maximal monotone. In this case, it is known that ∂i_C is reduced to the normality operator N_C for C; that is,

$$N_C(x) = \{x^* \in E^* : \langle y - x, x^*\rangle \leq 0 \quad (\forall y \in C)\} \tag{2.16}$$

if $x \in C$ and $\emptyset$ if $x \in E \setminus C$. We also know that Π_C is the resolvent of N_C. In fact, we have

$$\begin{aligned} z = \Pi_C(x) &\Longleftrightarrow z = \arg\min_{y \in C} \phi(y, x) \\ &\Longleftrightarrow z = \arg\min_{y \in E} \{\phi(y, x) + i_C(x)\} \\ &\Longleftrightarrow 0 \in \partial(\phi(\cdot, x) + i_C)(z) \\ &\Longleftrightarrow 0 \in 2Jz - 2Jx + N_C(z) \\ &\Longleftrightarrow Jx \in Jz + 2^{-1} N_C(z) \\ &\Longleftrightarrow z = (J + 2^{-1} N_C)^{-1} Jx. \end{aligned} \tag{2.17}$$

Thus, Lemma 2.3 implies that Π_C is of firmly nonexpansive type. □

3. Fixed point theorem for firmly nonexpansive-type mappings. In this section, we study the existence of fixed points for firmly nonexpansive-type mappings in Banach spaces (Theorem 3.2). We need the following lemma:

LEMMA 3.1. *Let E be a smooth Banach space, let C be a nonempty closed convex subset of E, and let $T : C \to C$ be a mapping. Then T is of firmly nonexpansive type if and only if*

$$\phi(Tx, Ty) + \phi(Ty, Tx) + \phi(Tx, x) + \phi(Ty, y) \leq \phi(Tx, y) + \phi(Ty, x) \tag{3.1}$$

for all $x, y \in C$.

Proof. Let $x, y \in C$ be given. Then it follows from (2.8) that

$$\langle Tx - Ty, JTx - JTy \rangle \leq \langle Tx - Ty, Jx - Jy \rangle \tag{3.2}$$

is equivalent to

$$\begin{aligned} &\frac{1}{2} \{\phi(Tx, Ty) + \phi(Ty, Tx) - \phi(Tx, Tx) - \phi(Ty, Ty)\} \\ &\leq \frac{1}{2} \{\phi(Tx, y) + \phi(Ty, x) - \phi(Tx, x) - \phi(Ty, y)\}. \end{aligned} \tag{3.3}$$

This is also equivalent to (3.1). This completes the proof. □

Using the technique developed by Takahashi [39], we can prove the following fixed point theorem for firmly nonexpansive-type mappings in Banach spaces:

THEOREM 3.2. *Let E be a smooth, strictly convex and reflexive Banach space, let C be a nonempty closed convex subset of E, and let $T : C \to C$ be a firmly nonexpansive-type mapping. Then the following are equivalent:*

1. *There exists $x \in C$ such that $\{T^n x\}$ is bounded;*
2. *$F(T)$ is nonempty.*

Proof. It is obvious that (2) implies (1). We show that (1) implies (2). Suppose that there exists $x \in C$ such that $\{T^n x\}$ is bounded and let

$$S_n(z) = \frac{1}{n} \sum_{k=0}^{n-1} T^k z \tag{3.4}$$

for all $z \in C$ and $n \in \mathbb{N}$. Let $y \in C$ and $k \in \mathbb{N} \cup \{0\}$ be given. Then it follows from Lemma 3.1 and (2.6) that

$$\begin{aligned} &\phi(T^{k+1}x, Ty) + \phi(Ty, T^{k+1}x) + \phi(T^{k+1}x, T^k x) + \phi(Ty, y) \\ &\leq \phi(T^{k+1}x, y) + \phi(Ty, T^k x) \\ &= \phi(T^{k+1}x, Ty) + \phi(Ty, y) + 2\langle T^{k+1}x - Ty, JTy - Jy \rangle + \phi(Ty, T^k x). \end{aligned} \tag{3.5}$$

This implies that

$$\begin{aligned} 0 &\le 2\langle T^{k+1}x - Ty, JTy - Jy\rangle \\ &\quad + \phi(Ty, T^k x) - \phi(Ty, T^{k+1}x) - \phi(T^{k+1}x, T^k x) \\ &\le 2\langle T^{k+1}x - Ty, JTy - Jy\rangle + \phi(Ty, T^k x) - \phi(Ty, T^{k+1}x). \end{aligned} \tag{3.6}$$

Summing these inequalities with respect to $k = 0, 1, \ldots, n-1$, we have

$$0 \le 2\left\langle \sum_{k=0}^{n-1} T^{k+1}x - nTy, JTy - Jy \right\rangle + \phi(Ty, x) - \phi(Ty, T^n x). \tag{3.7}$$

Dividing (3.7) by n, we obtain

$$0 \le 2\langle S_n(Tx) - Ty, JTy - Jy\rangle + \frac{1}{n}\left\{\phi(Ty, x) - \phi(Ty, T^n x)\right\}. \tag{3.8}$$

Since $\{S_n(Tx)\}$ is bounded, we have a subsequence $\{S_{n_i}(Tx)\}$ such that $S_{n_i}(Tx) \rightharpoonup u \in C$. Tending $n_i \to \infty$ in (3.8), we have $0 \le 2\langle u - Ty, JTy - Jy\rangle$. Thus we have

$$0 \le \langle u - Ty, JTy - Jy\rangle \tag{3.9}$$

for all $y \in C$. Putting $y = u$ in (3.9), we have

$$0 \le \langle u - Tu, JTu - Ju\rangle = -\langle u - Tu, Ju - JTu\rangle. \tag{3.10}$$

Hence we have $\langle u - Tu, Ju - JTu\rangle \le 0$. On the other hand, since J is monotone, we have $\langle u - Tu, Ju - JTu\rangle \ge 0$. Therefore $\langle u - Tu, Ju - JTu\rangle = 0$. Since E is strictly convex, we obtain $Tu = u$. Thus T has a fixed point and the proof is completed. □

As direct consequences of Theorem 3.2, we obtain the following corollaries. Note that Corollary 3.4 actually holds for nonexpansive mappings in Hilbert spaces; see Reich [29] and Takahashi [41]:

COROLLARY 3.3. *Let E be a smooth, strictly convex and reflexive Banach space, let C be a nonempty bounded closed convex subset of E and let $T : C \to C$ be a firmly nonexpansive-type mapping. Then $F(T)$ is nonempty.*

COROLLARY 3.4. *Let H be a Hilbert space, let C be a nonempty closed convex subset of a Hilbert space H and let $T : C \to C$ be a firmly nonexpansive mapping. Then there exists $x \in C$ such that $\{T^n x\}$ is bounded if and only if $F(T)$ is nonempty.*

4. Lemmas. In this section, we show some properties for relatively nonexpansive mappings needed in section 5. For a bounded sequence $\{x_n\}$ of a reflexive Banach space E, let $\omega_{\mathrm{w}}(\{x_n\})$ be the set defined by

$$\omega_{\mathrm{w}}(\{x_n\}) = \left\{z \in E : \exists\{x_{n_i}\} \subset \{x_n\} \text{ s.t. } x_{n_i} \rightharpoonup z\right\}. \tag{4.1}$$

We first show the following lemmas:

LEMMA 4.1. *Let E be a smooth, strictly convex and reflexive Banach space and let $\{x_n\}$ be a bounded sequence of E such that $\lim_n \phi(u, x_n)$ exists for all $u \in \omega_{\mathrm{w}}(\{x_n\})$. If J is weakly sequentially continuous, then $\{x_n\}$ converges weakly.*

Proof. It suffices to show that $A = \omega_{\mathrm{w}}(\{x_n\})$ consists of one point. Since E is reflexive and $\{x_n\}$ is bounded, A is nonempty. Let $u, v \in A$. Then, by assumption,

$$\lim_{n\to\infty}\left\{\phi(u, x_n) - \phi(v, x_n)\right\} \tag{4.2}$$

exists. This implies that $\lim_n \langle v-u, Jx_n\rangle$ exists. Let $x_{n_i} \rightharpoonup u$ and $x_{m_j} \rightharpoonup v$. Since J is weakly sequentially continuous, we have $Jx_{n_i} \overset{*}{\rightharpoonup} Ju$ and $Jx_{m_j} \overset{*}{\rightharpoonup} Jv$. Then we have

$$\begin{aligned}\langle v-u, Ju\rangle &= \lim_{i\to\infty}\langle v-u, Jx_{n_i}\rangle \\ &= \lim_{n\to\infty}\langle v-u, Jx_n\rangle \\ &= \lim_{j\to\infty}\langle v-u, Jx_{m_j}\rangle = \langle v-u, Jv\rangle.\end{aligned} \tag{4.3}$$

Thus we obtain $\langle u-v, Ju-Jv\rangle = 0$. Since E is strictly convex, we have $u=v$. This completes the proof. ∎

LEMMA 4.2. *Let E be a smooth and uniformly convex Banach space, let C be a nonempty closed convex subset of E, and let $T : C \to C$ be a strongly relatively nonexpansive mapping. Then T is asymptotically regular; that is,*

$$\|T^{n+1}x - T^n x\| \to 0 \tag{4.4}$$

for all $x \in C$.

Proof. Let $x \in C$ and $u \in F(T)$ be given. By the relative nonexpansiveness of T, we have

$$\phi(u, T^{n+1}x) \le \phi(u, T^n x) \tag{4.5}$$

for all $n \in \mathbb{N}$. Thus $\lim_n \phi(u, T^n x)$ exists. By $(\|u\| - \|T^n x\|)^2 \le \phi(u, T^n x)$ for all $n \in \mathbb{N}$, $\{T^n x\}$ is bounded. On the other hand, we have

$$\lim_{n\to\infty}\{\phi(u, T^n x) - \phi(u, T^{n+1}x)\} = 0. \tag{4.6}$$

Since T is strongly relatively nonexpansive, we have $\lim_n \phi(T^{n+1}x, T^n x) = 0$. By Lemma 2.1, we obtain the desired conclusion. ∎

Using Lemma 4.1, we can prove the following lemma, which is an analogous result of Opial [28] for relatively nonexpansive mappings in Banach spaces:

LEMMA 4.3. *Let E be a smooth and uniformly convex Banach space, let C be a nonempty closed convex subset of E, and let $T : C \to C$ be a relatively nonexpansive mapping which is asymptotically regular. Then for all $x \in C$, the following hold:*

1. *$\{T^n x\}$ is bounded and $\omega_{\mathrm{w}}(\{T^n x\}) \subset F(T)$;*
2. *if J is weakly sequentially continuous, then $\{T^n x\}$ converges weakly to an element of $F(T)$.*

Proof. We first show the part (1). As in the proof of Lemma 4.2, it follows that $\{T^n x\}$ is bounded for all $x \in C$. Since T is asymptotically regular, we have $\omega_{\mathrm{w}}(\{T^n x\}) \subset \widehat{F}(T)$. Since T is relatively nonexpansive, we have $\widehat{F}(T) = F(T)$. Thus we have the conclusion.

We next show part (2). Assume that J is weakly sequentially continuous and let $u \in \omega_{\mathrm{w}}(\{T^n x\})$ be given. By the part (1), we have $u \in F(T)$. As in the proof of Lemma 4.2, we can show that $\lim_n \phi(u, T^n x)$ exists. Thus $\lim_n \phi(u, T^n x)$ exists for all $u \in \omega_{\mathrm{w}}(\{T^n x\})$. By Lemma 4.1, $\{T^n x\}$ converges weakly. Consequently, it converges weakly to an element of $F(T)$. ∎

5. Convergence theorem for firmly nonexpansive-type mappings. In this section, we prove a weak convergence theorem for firmly nonexpansive-type mappings in Banach spaces (Theorem 5.3). Before proving it, we show that every firmly

nonexpansive-type mapping with a fixed point is strongly relatively nonexpansive (Theorem 5.2).

LEMMA 5.1. *Let E be a strictly convex Banach space whose norm is uniformly Gâteaux differentiable, let C be a nonempty closed convex subset of E, and let $T : C \to C$ be a firmly nonexpansive-type mapping. Then $\widehat{F}(T) = F(T)$.*

Proof. Since $\widehat{F}(T) \supset F(T)$ is obvious, we prove $\widehat{F}(T) \subset F(T)$. Let $u \in \widehat{F}(T)$ be given. By the definition of $\widehat{F}(T)$, we have a sequence $\{x_n\}$ of C such that $x_n \rightharpoonup u$ and $\|x_n - Tx_n\| \to 0$. Then we have $Tx_n \rightharpoonup u$. Since the norm of E is uniformly Gâteaux differentiable, J is uniformly norm-to-weak* continuous on each bounded subset of E; see Reich [31] and Takahashi [41]. Hence it follows from $\|x_n - Tx_n\| \to 0$ that

$$\langle y, JTx_n - Jx_n \rangle \to 0 \tag{5.1}$$

for all $y \in E$.

On the other hand, since T is of firmly nonexpansive type, it follows from Lemma 3.1 that

$$\phi(Tx_n, Tu) + \phi(Tu, Tx_n) + \phi(Tx_n, x_n) + \phi(Tu, u) \le \phi(Tx_n, u) + \phi(Tu, x_n) \tag{5.2}$$

for all $n \in \mathbb{N}$. This implies that

$$\begin{aligned} &\phi(Tu, u) \\ &\le \phi(Tx_n, u) - \phi(Tx_n, Tu) + \phi(Tu, x_n) - \phi(Tu, Tx_n) - \phi(Tx_n, x_n) \\ &= 2\langle Tx_n, JTu - Ju \rangle + \|u\|^2 - \|Tu\|^2 \\ &\quad + 2\langle Tu, JTx_n - Jx_n \rangle + \|x_n\|^2 - \|Tx_n\|^2 - \phi(Tx_n, x_n) \\ &\le 2\langle Tx_n, JTu - Ju \rangle + \|u\|^2 - \|Tu\|^2 \\ &\quad + 2\langle Tu, JTx_n - Jx_n \rangle + (\|x_n\| + \|Tx_n\|)\|x_n - Tx_n\| \end{aligned} \tag{5.3}$$

for all $n \in \mathbb{N}$. Tending $n \to \infty$ in (5.3), it follows from (5.1) that

$$\begin{aligned} \phi(Tu, u) &\le 2\langle u, JTu - Ju \rangle + \|u\|^2 - \|Tu\|^2 \\ &= \phi(u, u) - \phi(u, Tu) = -\phi(u, Tu). \end{aligned} \tag{5.4}$$

Hence we have

$$\phi(Tu, u) + \phi(u, Tu) \le 0, \tag{5.5}$$

and hence $\phi(Tu, u) = \phi(u, Tu) = 0$. By the strict convexity of E, we obtain $Tu = u$. Therefore, $\widehat{F}(T) \subset F(T)$. This completes the proof. □

Using Lemma 5.1, we can show the following theorem:

THEOREM 5.2. *Let E be a strictly convex Banach space whose norm is uniformly Gâteaux differentiable, let C be a nonempty closed convex subset of E, and let $T : C \to C$ be a firmly nonexpansive-type mapping such that $F(T)$ is nonempty. Then T is strongly relatively nonexpansive.*

Proof. Let $(x, u) \in C \times F(T)$ be given. Then it follows from Lemma 3.1 that

$$\phi(Tx, Tu) + \phi(Tu, Tx) + \phi(Tx, x) + \phi(Tu, u) \le \phi(Tx, u) + \phi(Tu, x). \tag{5.6}$$

This implies that

$$\phi(u, Tx) + \phi(Tx, x) \le \phi(u, x). \tag{5.7}$$

So, we have $\phi(u,Tx) \leq \phi(u,x)$. On the other hand, it also follows from Lemma 5.1 that $\widehat{F}(T) = F(T)$. By assumption, $F(T)$ is nonempty. Thus T is relatively nonexpansive.

We finally show that T is strongly relatively nonexpansive. Let $\{x_n\}$ be a bounded sequence of C such that

$$\phi(u,x_n) - \phi(u,Tx_n) \to 0 \tag{5.8}$$

for some $u \in F(T)$. By (5.7), we have

$$\phi(Tx_n,x_n) \leq \phi(u,x_n) - \phi(u,Tx_n) \tag{5.9}$$

for all $n \in \mathbb{N}$. Then it follows from (5.8) and (5.9) that $\lim_n \phi(Tx_n,x_n) = 0$. Thus T is strongly relatively nonexpansive and the proof is completed. □

By Lemmas 4.2 and 4.3 and Theorem 5.2, we obtain the following weak convergence theorem for firmly nonexpansive-type mappings in Banach spaces:

THEOREM 5.3. *Let E be a uniformly convex Banach space whose norm is uniformly Gâteaux differentiable, let C be a nonempty closed convex subset of E, and let $T : C \to C$ be a firmly nonexpansive-type mapping such that $F(T)$ is nonempty. Then for all $x \in C$, the following hold:*

1. *$\{T^n x\}$ is bounded and $\omega_{\mathrm{w}}(\{T^n x\}) \subset F(T)$;*
2. *if J is weakly sequentially continuous, then $\{T^n x\}$ converges weakly to an element of $F(T)$.*

As a direct consequence of Theorem 5.3, we have the following result due to Martinet [23]; see also Bruck and Reich [7]:

COROLLARY 5.4 (Martinet [23]). *Let C be a nonempty closed convex subset of a Hilbert space H and let T be a firmly nonexpansive mapping from C into itself such that $F(T)$ is nonempty. Then for all $x \in C$, $\{T^n x\}$ converges weakly to an element of $F(T)$.*

6. Applications to the proximal point algorithm. In the final section, we study the proximal point algorithm for monotone operators in Banach spaces first introduced by Martinet [22] and generally studied by Rockafellar [38] in Hilbert spaces; see also Brézis and Lions [5], Bruck and Reich [7], and Nevanlinna and Reich [27]. This is an iterative procedure, which generates a sequence $\{x_n\}$ by the rule $x_1 = x \in H$ and

$$x_{n+1} = J_{r_n} x_n \quad (n = 1,2,\ldots), \tag{6.1}$$

where $A \subset H \times H$ is a maximal monotone operator defined in a Hilbert space H, $J_r = (I + rA)^{-1}$ for all $r > 0$, and $\{r_n\}$ is a sequence of positive real numbers. Rockafellar's theorem [38] ensures that if $A^{-1}0$ is nonempty and $\liminf_n r_n > 0$, then $\{x_n\}$ converges weakly to an element of $A^{-1}0$.

Using Lemma 2.3 and Theorem 3.2, we can study the existence of zeros of monotone operators in Banach spaces. This result is closely related to the result due to Matsushita and Takahashi [26].

THEOREM 6.1. *Let E be a smooth, strictly convex and reflexive Banach space and let $A \subset E \times E^*$ be a monotone operator and C be a nonempty closed convex subset of E satisfying* (2.9). *Let r be a positive real number and $J_r x = (J + rA)^{-1} Jx$ for all $x \in C$. Then $A^{-1}0$ is nonempty if and only if there exists $x \in C$ such that $\{J_r^n x\}$ is bounded. In particular, if $D(A)$ is bounded, then $A^{-1}0$ is nonempty.*

Proof. Let r be a positive real number. From Lemma 2.3, we know that $J_r : C \to D(A)$ is a firmly nonexpansive-type mapping. Since $D(A) \subset C$ from (2.9), we obtain the desired result from Theorem 3.2. If $D(A)$ is bounded, then $\{J_r^n x\}$ is bounded for all $x \in C$. So, $A^{-1}0$ is nonempty. □

Using Lemma 2.3 and Theorem 5.3, we finally study the convergence of the proximal point algorithm for monotone operators in Banach spaces; see Kamimura [16] and Kamimura, Kohsaka and Takahashi [17] for similar results on this subject.

THEOREM 6.2. *Let E be a uniformly convex Banach space whose norm is uniformly Gâteaux differentiable, let $A \subset E \times E^*$ be a monotone operator such that $A^{-1}0$ is nonempty and let C be a nonempty closed convex subset of E satisfying* (2.9). *Let r be a positive real number and let $J_r x = (J + rA)^{-1} Jx$ for all $x \in C$. If J is weakly sequentially continuous, then for all $x \in C$, $\{J_r^n x\}$ converges weakly to an element of $A^{-1}0$.*

Proof. Let r be a positive real number. Since E is uniformly convex, E is reflexive and strictly convex. So, from Lemma 2.3, we have that $J_r : C \to D(A)$ is a firmly nonexpansive-type mapping. From (2.9) and the part (2) of Theorem 5.3, we have the desired result. □

Acknowledgment. The authors would like to express their sincere appreciation to the anonymous referee for valuable comments on the original version of the manuscript.

REFERENCES

[1] Y. I. ALBER, *Metric and generalized projection operators in Banach spaces: Properties and applications*, Theory and Applications of Nonlinear Operators of Accretive and Monotone Type, A. G. Kartsatos, ed., Lecture Notes in Pure and Appl. Math. 178, Dekker, NY, 1996, pp. 15–50.

[2] Y. I. ALBER AND S. REICH, *An iterative method for solving a class of nonlinear operator equations in Banach spaces*, Panamer. Math. J., 4 (1994), pp. 39–54.

[3] V. BARBU, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Editura Academiei Republicii Socialiste România, Bucharest, 1976.

[4] L. M. BREGMAN, *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming*, USSR Comput. Math. and Math. Phys., 7 (1967), pp. 200–217.

[5] H. BRÉZIS AND P.-L. LIONS, *Produits infinis de résolvantes*, Israel J. Math., 29 (1978), pp. 329–345.

[6] F. E. BROWDER, *Nonlinear maximal monotone operators in Banach space*, Math. Ann., 175 (1968), pp. 89–113.

[7] R. E. BRUCK AND S. REICH, *Nonexpansive projections and resolvents of accretive operators in Banach spaces*, Houston J. Math., 3 (1977), pp. 459–470.

[8] D. BUTNARIU AND A. N. IUSEM, *Totally Convex Functions for Fixed Points Computation and Infinite Dimensional Optimization*, Kluwer Academic, Dordrecht, 2000.

[9] D. BUTNARIU, S. REICH, AND A. J. ZASLAVSKI, *Asymptotic behavior of relatively nonexpansive operators in Banach spaces*, J. Appl. Anal., 7 (2001), pp. 151–174.

[10] D. BUTNARIU, S. REICH, AND A. J. ZASLAVSKI, *Weak convergence of orbits of nonlinear operators in reflexive Banach spaces*, Numer. Funct. Anal. Optim., 24 (2003), pp. 489–508.

[11] Y. CENSOR AND A. LENT, *An iterative row-action method for interval convex programming*, J. Optim. Theory Appl., 34 (1981), pp. 321–353.

[12] Y. CENSOR AND S. REICH, *Iterations of paracontractions and firmly nonexpansive operators with applications to feasibility and optimization*, Optimization, 37 (1996), pp. 323–339.

[13] I. CIORANESCU, *Geometry of Banach Spaces, Duality Mappings and Nonlinear Problems*, Kluwer Academic, Dordrecht, 1990.

[14] K. GOEBEL AND W. A. KIRK, *Topics in Metric Fixed Point Theory*, Cambridge Studies in Advanced Mathematics 28, Cambridge University Press, Cambridge, 1990.

[15] K. GOEBEL AND S. REICH, *Uniform Convexity, Hyperbolic Geometry, and Nonexpansive Mappings*, Monographs and Textbooks in Pure and Applied Mathematics 83, Marcel Dekker Inc., NY, 1984.

[16] S. Kamimura, *The proximal point algorithm in a Banach space*, Proceedings of the Third International Conference on Nonlinear Analysis and Convex Analysis, W. Takahashi and T. Tanaka. eds., Yokohama Publishers, Yokohama, 2004, pp. 143–148.
[17] S. Kamimura, F. Kohsaka, and W. Takahashi, *Weak and strong convergence theorems for maximal monotone operators in a Banach space*, Set-Valued Anal., 12 (2004), pp. 417–429.
[18] S. Kamimura and W. Takahashi, *Strong convergence of a proximal-type algorithm in a Banach space*, SIAM J. Optim., 13 (2002), pp. 938–945.
[19] F. Kohsaka and W. Takahashi, *Strong convergence of an iterative sequence for maximal monotone operators in a Banach space*, Abstr. Appl. Anal., 2004 (2004), pp. 239–249.
[20] F. Kohsaka and W. Takahashi, *Block iterative methods for a finite family of relatively nonexpansive mappings in Banach spaces*, Fixed Point Theory Appl., 2007 (2007), Article ID 21972, pp. 1–18.
[21] F. Kohsaka and W. Takahashi, *Approximating common fixed points of countable families of strongly nonexpansive mappings*, Nonlinear Stud., 14 (2007), pp. 219–234.
[22] B. Martinet, *Régularisation d'inéquations variationnelles par approximations successives*, Rev. Française Informat. Recherche Opérationnelle, 4 (1970), pp. 154–158.
[23] B. Martinet, *Détermination approchée d'un point fixe d'une application pseudo-contractante. Cas de l'application prox*, C. R. Acad. Sci. Paris Sér. A, 274 (1972), pp. 163–165.
[24] S. Matsushita and W. Takahashi, *Weak and strong convergence theorems for relatively nonexpansive mappings in Banach spaces*, Fixed Point Theory Appl., 2004 (2004), pp. 37–47.
[25] S. Matsushita and W. Takahashi, *A strong convergence theorem for relatively nonexpansive mappings in a Banach space*, J. Approx. Theory, 134 (2005), pp. 257–266.
[26] S. Matsushita and W. Takahashi, *The existence of zeros of monotone operators concerning optimization problems*, Sūrikaisekikenkyūsho Kōkyūroku, Kyoto, 1461 (2005), pp. 40–46 (in Japanese).
[27] O. Nevanlinna and S. Reich, *Strong convergence of contraction semigroups and of iterative methods for accretive operators in Banach spaces*, Israel J. Math., 32 (1979), pp. 44–58.
[28] Z. Opial, *Weak convergence of the sequence of successive approximations for nonexpansive mappings*, Bull. Amer. Math. Soc., 73 (1967), pp. 591–597.
[29] S. Reich, *Fixed point iterations of nonexpansive mappings*, Pacific J. Math., 60 (1975), pp. 195–198.
[30] S. Reich, *The range of sums of accretive and monotone operators*, J. Math. Anal. Appl., 68 (1979), pp. 310–317.
[31] S. Reich, *On the asymptotic behavior of nonlinear semigroups and the range of accretive operators*, J. Math. Anal. Appl., 79 (1981), pp. 113–126.
[32] S. Reich, *Book review: Geometry of Banach Spaces, Duality Mappings and Nonlinear Problems by I. Cioranescu*, Kluwer Academic, Dordrecht (1990), Bull. Amer. Math. Soc., 26 (1992), pp. 367–370.
[33] S. Reich, *A weak convergence theorem for the alternating method with Bregman distances*, Theory and Applications of Nonlinear Operators of Accretive and Monotone Type, A. G. Kartsatos, ed., Lecture Notes in Pure and Appl. Math. 178, Dekker, NY, 1996, pp. 313–318.
[34] R. T. Rockafellar, *Characterization of the subdifferentials of convex functions*, Pacific J. Math., 17 (1966), pp. 497–510.
[35] R. T. Rockafellar, *On the maximal monotonicity of subdifferential mappings*, Pacific J. Math., 33 (1970), pp. 209–216.
[36] R. T. Rockafellar, *On the virtual convexity of the domain and range of a nonlinear maximal monotone operator*, Math. Ann., 185 (1970), pp. 81–90.
[37] R. T. Rockafellar, *On the maximality of sums of nonlinear monotone operators*, Trans. Amer. Math. Soc., 149 (1970), pp. 75–88.
[38] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.
[39] W. Takahashi, *A nonlinear ergodic theorem for an amenable semigroup of nonexpansive mappings in a Hilbert space*, Proc. Amer. Math. Soc., 81 (1981), pp. 253–256.
[40] W. Takahashi, *Convex Analysis and Approximation of Fixed Points*, Yokohama Publishers, Yokohama, 2000 (in Japanese).
[41] W. Takahashi, *Nonlinear Functional Analysis. Fixed Point Theory and its Applications*, Yokohama Publishers, Yokohama, 2000.

SIAM J. OPTIM.
Vol. 19, No. 2, pp. 836–845

GEOMETRIC DUALITY IN MULTIPLE OBJECTIVE LINEAR PROGRAMMING*

FRANK HEYDE† AND ANDREAS LÖHNE†

Abstract. We develop in this article a geometric approach to duality in multiple objective linear programming. This approach is based on a very old idea, the duality of polytopes, which can be traced back to the old Greeks. We show that there is an inclusion-reversing one-to-one map between the minimal faces of the image of the primal objective and the maximal faces of the image of the dual objective map.

Key words. duality theory, multiobjective optimization, linear programming, dual polytopes

AMS subject classifications. 90C29, 90C05, 90C46

DOI. 10.1137/060674831

1. Introduction.

Duality for multiple objective linear programs seems to have its origin in the 1970s; see, e.g., Kornbluth [13], Rödder [17], Isermann [9, 10], and Brumelle [2]. More recent expositions are Jahn [11, 12], Luc [15], and Göpfert and Nehse [4], where nonlinear problems are also considered.

As noticed in [4, p. 64], the practical relevance of vectorial duality theory is quite low in comparison with the relevance of duality in scalar optimization. Moreover, in the linear case there occurs some difficulties, such as a duality gap in the case $b = 0$ (where b is the right-hand side of the inequality constraints). In [6], this duality gap could be closed by using a set-valued approach. In [14, 7, 8], this set-valued approach is revisited from a lattice theoretic point of view. The aim of these papers is to work in an appropriate complete lattice in order to have a duality theory which can be formulated along the lines of the scalar duality theory. In particular, the infimum and supremum can be used to define solutions. Another goal (especially in [8]) is to have a "simple" dual problem. This means that the dual problem should not be more complicated than the primal problem.

Nevertheless, in all the mentioned references there is a basic difference from the present article. Instead of speaking about strong duality if the optimal values of a pair of dual optimization problems are equal, we deal with a duality relation between the polyhedral image set of the primal problem and the polyhedral image of the dual problem, which is similar to duality of polytopes (see Figure 1.1).

It is well known from the theory of convex polytopes (see, e.g., [5]) that two polytopes $\mathcal{P}$ and $\mathcal{P}^*$ in $\mathbb{R}^q$ are said to be dual to each other provided that there exists a one-to-one mapping Ψ between the set of all faces of $\mathcal{P}$ and the set of all faces of $\mathcal{P}^*$ such that Ψ is inclusion-reversing; i.e., faces $\mathcal{F}_1$ and $\mathcal{F}_2$ of $\mathcal{P}$ satisfy $\mathcal{F}_1 \subseteq \mathcal{F}_2$ if and only if the faces $\Psi(\mathcal{F}_1)$ and $\Psi(\mathcal{F}_2)$ satisfy $\Psi(\mathcal{F}_1) \supseteq \Psi(\mathcal{F}_2)$.

Denoting by $\mathcal{P}$ and $\mathcal{D}$ the images of the objective functions of our given problem (P) and its dual problem (D), respectively, we show that there is an inclusion reversing one-to-one map Ψ between the set of all K-maximal proper faces of $\mathcal{D}$ and the set of all weakly C-minimal proper faces of $\mathcal{P}$, where K and C are appropriate ordering

*Received by the editors November 13, 2006; accepted for publication (in revised form) February 24, 2008; published electronically August 1, 2008.
http://www.siam.org/journals/siopt/19-2/67483.html
†Institute of Mathematics, MLU Halle–Wittenberg, 06099 Halle (Saale), Germany (frank.heyde@mathematik.uni-halle.de, andreas.loehne@mathematik.uni-halle.de).

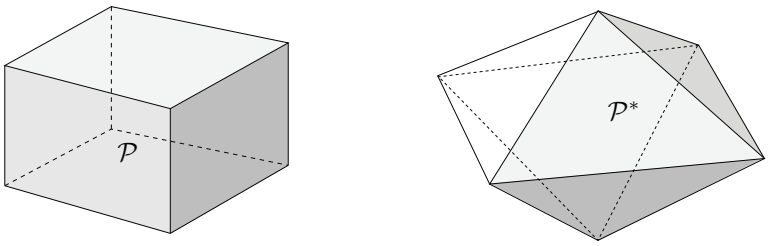


FIG. 1.1. *Example of a pair of dual polytopes in* $\mathbb{R}^3$.

cones. With the aid of such a map Ψ we can compute the weakly C-minimal faces of $\mathcal{P}$ whenever we know the K-maximal faces of $\mathcal{D}$ and vice versa. In particular, we are given by Ψ a one-to-one correspondence between weakly C-minimal vertices (facets) of $\mathcal{P}$ and K-maximal facets (vertices) of $\mathcal{D}$. It is worth mentioning that there is a connection between the lattice theoretic duality in [8] and the geometric duality in the present article. This is briefly discussed at the end of section 3.

In a forthcoming paper [3] we give an application of geometric duality, a dual variant of Benson's outer approximation algorithm [1].

2. Preliminaries. Let $\mathcal{A} \subseteq \mathbb{R}^q$, and let $\mathcal{C} \subseteq \mathbb{R}^q$ be a closed convex cone. Denoting by $\operatorname{ri}\mathcal{C}$ the relative interior of $\mathcal{C}$, we set

$$\operatorname{Min}_{\mathcal{C}}\mathcal{A} := \{y \in \mathcal{A} \mid (\{y\} - \operatorname{ri}\mathcal{C}) \cap \mathcal{A} = \emptyset\} \quad \text{and} \quad \operatorname{Max}_{\mathcal{C}}\mathcal{A} := \operatorname{Min}_{(-\mathcal{C})}\mathcal{A}.$$

In the following we consider two special ordering cones, namely,

$$C := \mathbb{R}^q_+ \quad \text{and} \quad K := \mathbb{R}_+ \cdot (0, 0, \ldots, 0, 1)^T = \{y \in \mathbb{R}^q \mid y_1 = \cdots = y_{q-1} = 0, y_q \geq 0\}.$$

Note that throughout this article elements of the vector space $\mathbb{R}^q$ (or $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively) are considered to be column vectors.

For the choice $\mathcal{C} = C = \mathbb{R}^q_+$ we have $\operatorname{ri}\mathcal{C} = \operatorname{int} C$; hence

$$\operatorname{Min}_{C}\mathcal{A} = \{y \in \mathcal{A} \mid (\{y\} - \operatorname{int} C) \cap \mathcal{A} = \emptyset\}$$

coincides with the set of weakly C-minimal elements of $\mathcal{A}$. In case of $\mathcal{C} = K$ we have $\operatorname{ri} K = K \setminus \{0\}$; hence

$$\operatorname{Max}_{K}\mathcal{A} = \{y \in \mathcal{A} \mid (\{y\} + K \setminus \{0\}) \cap \mathcal{A} = \emptyset\}$$

coincides with the set of K-maximal elements of $\mathcal{A}$.

For the convenience of the reader, we recall some facts concerning the facial structure of polyhedral sets [18, section 3.2]. A polyhedral set is defined to be the intersection of a finite collection of closed half-spaces. Clearly, polyhedral sets are closed and convex. Let $\mathcal{A} \subseteq \mathbb{R}^q$ be a convex set. A convex subset $\mathcal{F} \subseteq \mathcal{A}$ is called a *face* of $\mathcal{A}$ if

$$\big(y^1, y^2 \in \mathcal{A}, \quad \lambda \in (0,1), \quad \lambda y^1 + (1-\lambda)y^2 \in \mathcal{F}\big) \quad \Rightarrow \quad y^1, y^2 \in \mathcal{F}.$$

A face $\mathcal{F}$ of $\mathcal{A}$ is called *proper* if $\emptyset \neq \mathcal{F} \neq \mathcal{A}$. A set $\mathcal{E} \subseteq \mathcal{A}$ is called an *exposed face* of $\mathcal{A}$ if there are $c \in \mathbb{R}^q$ and $\gamma \in \mathbb{R}$ such that $\mathcal{A} \subseteq \{y \in \mathbb{R}^q \mid c^T y \geq \gamma\}$ and $\mathcal{E} = \{y \in \mathbb{R}^q \mid c^T y = \gamma\} \cap \mathcal{A}$. The proper $(r-1)$-dimensional faces of an r-dimensional

polyhedral convex set $\mathcal{A}$ are called *facets* of $\mathcal{A}$. A point $y \in \mathcal{A}$ is called a vertex of $\mathcal{A}$ if $\{y\}$ is a face of $\mathcal{A}$. Let $\mathcal{A}$ be a polyhedral set in $\mathbb{R}^q$. Then $\mathcal{A}$ has a finite number of faces, each of which is exposed, and a polyhedral set. Every proper face of $\mathcal{A}$ is the intersection of those facets of $\mathcal{A}$ that contain it, and the relative boundary of $\mathcal{A}$ is the union of all the facets of $\mathcal{A}$. If $\mathcal{A}$ has a nonempty face of dimension s, then $\mathcal{A}$ has faces of all dimensions from s to $\dim \mathcal{A}$ (see [18, Theorem 3.2.2]).

If $\operatorname{int} \mathcal{A} \neq \emptyset$, then $\mathcal{A}$ is a q-dimensional polyhedral set; hence the facets of $\mathcal{A}$ are the $(q-1)$-dimensional faces of $\mathcal{A}$, i.e., the maximal (w.r.t. inclusion) proper faces. A subset $\mathcal{F} \subseteq \mathcal{A}$ is a proper face if and only if it is a proper exposed face, i.e., there is a supporting hyperplane $\mathcal{H}$ to $\mathcal{A}$ such that $\mathcal{F} = \mathcal{H} \cap \mathcal{A}$. We call a hyperplane $\mathcal{H} := \left\{ y \in \mathbb{R}^q \mid c^T y = \gamma \right\}$ (i.e., $c \neq 0$) supporting to $\mathcal{A}$ if

$$\forall y \in \mathcal{A} : c^T y \geq \gamma \qquad \wedge \qquad \exists y^0 \in \mathcal{A} : c^T y^0 = \gamma.$$

3. Main result. Throughout the article, let $m, n, q \in \mathbb{N}$ and $A \in \mathbb{R}^{m \times n}$, $M \in \mathbb{R}^{q \times n}$, $b \in \mathbb{R}^m$ be given, and let the ordering cones C and K be defined as above. Further we set $k = (1, \ldots, 1)^T \in \mathbb{R}^q$. We consider the following vector optimization problem:

$$\text{(P)} \qquad \operatorname{Min}_C M[\mathcal{X}], \quad \mathcal{X} := \{x \in \mathbb{R}^n \mid Ax \geq b\}.$$

We define a dual linear objective function by $D : \mathbb{R}^m \times \mathbb{R}^q \to \mathbb{R}^q$, $D(u,c) := \left(c_1, \ldots, c_{q-1}, b^T u\right)^T$ and consider the following dual vector optimization problem:

$$\text{(D)} \qquad \operatorname{Max}_K D[\mathcal{U}], \quad \mathcal{U} := \left\{(u,c) \in \mathbb{R}^m \times \mathbb{R}^q \mid (u,c) \geq 0,\ A^T u = M^T c,\ k^T c = 1\right\}.$$

Remark. The fact that the primal and dual problems are not symmetric is a feature shared by all attempts at duality for vector optimization problems. The special choice of the cone K for the dual problem reflects a parametric character of the dual problem. In fact, a point $D(\bar{u}, \bar{c})$ is a K-maximal point of $D[\mathcal{U}]$ if and only if, for $\bar{c}$ fixed, $\bar{u}$ maximizes $b^T u$ over the set $\{u \in \mathbb{R}^m \mid (u, \bar{c}) \in \mathcal{U}\}$.

It is our goal to show a duality relation between the sets

$$\begin{aligned} \mathcal{P} &:= M[\mathcal{X}] + C = \{y \in \mathbb{R}^q \mid \quad \exists x \in \mathcal{X} : \quad y \in \{Mx\} + C \quad \} \qquad \text{and} \\ \mathcal{D} &:= D[\mathcal{U}] - K = \{y \in \mathbb{R}^q \mid \exists (u,c) \in \mathcal{U} : \quad y \in \{D(u,c)\} - K \ \}. \end{aligned}$$

To this end we construct an inclusion-reversing one-to-one map Ψ between the K-maximal proper faces of $\mathcal{D}$ and the weakly C-minimal proper faces of $\mathcal{P}$.

Consider the coupling function $\varphi : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}$, defined by

$$\varphi(y,v) := \sum_{i=1}^{q-1} y_i v_i + y_q \left(1 - \sum_{i=1}^{q-1} v_i\right) - v_q.$$

Note that $\varphi(\cdot, v)$ and $\varphi(y, \cdot)$ are affine. Choosing the values of the primal and dual objective functions as arguments, we just get

$$\varphi(Mx, D(u,c)) = c^T Mx - b^T u. \tag{3.1}$$

The coupling function φ is used to define the following two set-valued maps:

$$\mathcal{H} : \mathbb{R}^q \rightrightarrows \mathbb{R}^q, \quad \mathcal{H}(v) := \{y \in \mathbb{R}^q \mid \varphi(y,v) = 0\},$$

$$\mathcal{H}^* : \mathbb{R}^q \rightrightarrows \mathbb{R}^q, \quad \mathcal{H}^*(y) := \{v \in \mathbb{R}^q \mid \varphi(y,v) = 0\}.$$

Of course, $\mathcal{H}(v)$ and $\mathcal{H}^*(y)$ are hyperplanes in $\mathbb{R}^q$ for all $v, y \in \mathbb{R}^q$. Using the notation

$$c(v) := \left(v_1, \ldots, v_{q-1}, 1 - \sum_{i=1}^{q-1} v_i \right)^T \qquad \text{and} \qquad c^*(y) := \left(y_1 - y_q, \ldots, y_{q-1} - y_q, -1 \right)^T,$$

it is easy to see that

$$\mathcal{H}(v) = \left\{ y \in \mathbb{R}^q \mid c(v)^T y = v_q \right\} \qquad \text{and} \qquad \mathcal{H}^*(y) = \left\{ v \in \mathbb{R}^q \mid c^*(y)^T v = -y_q \right\}.$$

Obviously, the set-valued maps $\mathcal{H}$ and $\mathcal{H}^*$ are injective. The map $\mathcal{H}$ is now used to define the function $\Psi : 2^{\mathbb{R}^q} \to 2^{\mathbb{R}^q}$:

$$\Psi(\mathcal{F}^*) := \bigcap_{v \in \mathcal{F}^*} \mathcal{H}(v) \cap \mathcal{P}.$$

Now we have the main result, which shows that Ψ is a duality map between $\mathcal{P}$ and $\mathcal{D}$.

THEOREM 3.1. *Ψ is an inclusion-reversing one-to-one map between the set of all K-maximal proper faces of $\mathcal{D}$ and the set of all weakly C-minimal proper faces of $\mathcal{P}$, and the inverse map is given by*

$$\Psi^{-1}(\mathcal{F}) = \bigcap_{y \in \mathcal{F}} \mathcal{H}^*(y) \cap \mathcal{D}. \tag{3.2}$$

Moreover, for every K-maximal proper face $\mathcal{F}^*$ of $\mathcal{D}$ it holds that $\dim \mathcal{F}^* + \dim \Psi(\mathcal{F}^*) = q - 1$.

The proof of this theorem is given in the last section.

Let us consider an important special case. Vertices as well as facets are actually the most important faces from the point of view of applications. Therefore we extract some corresponding conclusions from the above theorem.

COROLLARY 3.2. *The following statements are equivalent.*

(i) *v is a K-maximal vertex of $\mathcal{D}$.*

(ii) *$\mathcal{H}(v) \cap \mathcal{P}$ is a weakly C-minimal $(q-1)$-dimensional facet of $\mathcal{P}$.*

Moreover, if $\mathcal{F}$ is a weakly C-minimal $(q-1)$-dimensional facet of $\mathcal{P}$, there is some uniquely defined point $v \in \mathbb{R}^q$ such that $\mathcal{F} = \mathcal{H}(v) \cap \mathcal{P}$.

Proof. (i) $\Rightarrow$ (ii). Since $\mathcal{H}(v) \cap \mathcal{P} = \Psi(\{v\})$, Theorem 3.1 implies that $\mathcal{H}(v) \cap \mathcal{P}$ is a weakly C-minimal proper face of $\mathcal{P}$. Theorem 3.1 also implies that $\dim(\mathcal{H}(v) \cap \mathcal{P}) = q - 1 - \dim\{v\} = q - 1$.

(ii) $\Rightarrow$ (i). Let $\mathcal{H}(v) \cap \mathcal{P}$ be a weakly C-minimal $(q-1)$-dimensional facet of $\mathcal{P}$. By Theorem 3.1, $\Psi^{-1}(\mathcal{H}(v) \cap \mathcal{P})$ is a K-maximal vertex of $\mathcal{D}$, denoted by $\bar{v}$. It follows that $\Psi \circ \Psi^{-1}(\mathcal{H}(v) \cap \mathcal{P}) = \Psi(\{\bar{v}\})$ and hence $\mathcal{H}(v) \cap \mathcal{P} = \mathcal{H}(\bar{v}) \cap \mathcal{P}$, implying $\mathcal{H}(v) = \mathcal{H}(\bar{v})$ as $\dim(\mathcal{H}(v) \cap \mathcal{P}) = q - 1$. The mapping $\mathcal{H}$ being injective implies $v = \bar{v}$.

To show the last statement, let $\mathcal{F}$ be a weakly C-minimal $(q-1)$-dimensional facet of $\mathcal{P}$. Hence $\Psi^{-1}(\mathcal{F})$ is a K-maximal vertex of $\mathcal{D}$, denoted by v. It follows that $\mathcal{F} = \Psi \circ \Psi^{-1}(\mathcal{F}) = \Psi(\{v\}) = \mathcal{H}(v) \cap \mathcal{P}$. By $\dim(\mathcal{H}(v) \cap \mathcal{P}) = q - 1$ and $\mathcal{H}$ being injective, v is uniquely defined. ∎

COROLLARY 3.3. *The following statements are equivalent.*

(i) *y is a weakly C-minimal vertex of $\mathcal{P}$.*

(ii) *$\mathcal{H}^*(y) \cap \mathcal{D}$ is a K-maximal $(q-1)$-dimensional facet of $\mathcal{D}$.*

Moreover, if $\mathcal{F}^*$ is a K-maximal $(q-1)$-dimensional facet of $\mathcal{D}$, there is some uniquely defined point $y \in \mathbb{R}^q$ such that $\mathcal{F}^* = \mathcal{H}^*(y) \cap \mathcal{D}$.

Proof. (i) $\Rightarrow$ (ii). Let y be a weakly C-minimal vertex of $\mathcal{P}$. By Theorem 3.1, the set $\mathcal{F}^* := \Psi^{-1}(\{y\}) = \mathcal{H}^*(y) \cap \mathcal{D}$ is a K-maximal face of $\mathcal{D}$. From Theorem 3.1 we also conclude that $\dim \mathcal{F}^* = q - 1 - \dim \{y\} = q - 1$. Thus $\mathcal{F}^*$ is a facet of $\mathcal{D}$.

(ii) $\Rightarrow$ (i). Let $\mathcal{H}^*(y) \cap \mathcal{D}$ be a K-maximal $(q-1)$-dimensional facet of $\mathcal{D}$. By Theorem 3.1 $\Psi(\mathcal{H}^*(y) \cap \mathcal{D})$ is a weakly C-minimal vertex of $\mathcal{P}$, denoted by $\bar{y}$. It follows that $\Psi^{-1} \circ \Psi(\mathcal{H}^*(y) \cap \mathcal{D}) = \Psi^{-1}(\{\bar{y}\})$ and hence $\mathcal{H}^*(y) \cap \mathcal{D} = \mathcal{H}^*(\bar{y}) \cap \mathcal{D}$. Since $\dim(\mathcal{H}^*(y) \cap \mathcal{D}) = q - 1$ and $\mathcal{H}^*$ is injective, we get $y = \bar{y}$.

To show the last statement, let $\mathcal{F}^*$ be a K-maximal $(q-1)$-dimensional facet of $\mathcal{D}$. Hence $\Psi(\mathcal{F}^*)$ is a C-minimal vertex of $\mathcal{P}$, denoted by y. It follows that $\mathcal{F}^* = \Psi^{-1} \circ \Psi(\mathcal{F}^*) = \Psi^{-1}(\{y\}) = \mathcal{H}^*(y) \cap \mathcal{D}$. By $\dim(\mathcal{H}^*(y) \cap \mathcal{D}) = q - 1$ and $\mathcal{H}^*$ being injective, y is uniquely defined. ∎

Remark. In addition to the correspondence between the faces of $\mathcal{P}$ and $\mathcal{D}$, Theorem 3.1 provides a relationship between the optimal values of (P) and (D) in the following way. If $y \in \operatorname{Min}_C M[\mathcal{X}]$, then there exists some $v \in \operatorname{Max}_K D[\mathcal{U}]$ with $\varphi(y, v) = 0$ and if $v \in \operatorname{Max}_K D[\mathcal{U}]$, then there exists some $y \in \operatorname{Min}_C M[\mathcal{X}]$ with $\varphi(y, v) = 0$.

In [8] we developed a duality theory based on a lattice theoretic approach. The dual problem (D) in the present article is related to the set-valued dual problem in [8]. Indeed, both problems have the same constraints, given by $\mathcal{U}$. The set-valued objective map of the dual problem in [8] can be expressed by the objective function of (D) as $(u, c) \mapsto \mathcal{H}(D(u, c))$. Moreover, (u, c) being a weakly efficient solution for the dual problem (LD) in [8] is equivalent to $D(u, c)$ being a K-maximal point of $\mathcal{D}$.

4. Examples. The geometric duality is illustrated by the following two examples.

Example 1. Consider problem (P) with the following data:

$$M = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad A = \begin{pmatrix} 0 & -1 \\ 2 & 1 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}, \qquad b = \begin{pmatrix} -4 \\ 4 \\ 3 \\ 4 \end{pmatrix}.$$

The set $\mathcal{D}$ can be easily calculated as $\mathcal{D} = \operatorname{co}\left\{(\frac{1}{3}, \frac{4}{3})^T, (\frac{1}{2}, \frac{3}{2})^T, (\frac{2}{3}, \frac{4}{3})^T, (1, 0)^T\right\} - K$, where $\operatorname{co} \mathcal{A}$ denotes the convex hull of a set $\mathcal{A}$ (see Figure 4.1).

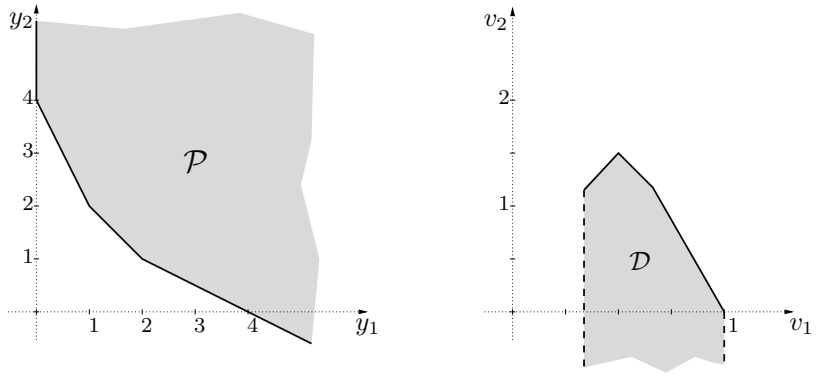


FIG. 4.1. *The three weakly C-minimal vertices of $\mathcal{P}$ correspond to the three K-maximal facets of $\mathcal{D}$, and the four weakly C-minimal facets of $\mathcal{P}$ correspond to the four K-maximal vertices of $\mathcal{D}$.*

Example 2. Consider problem (P) with the following data:

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

An easy computation shows that $\mathcal{D} = \operatorname{co}\left\{(0,0,0)^T, (1,0,0)^T, (0,1,0)^T, (\frac{1}{3},\frac{1}{3},\frac{1}{3})^T\right\} - K$ (see Figure 4.2).

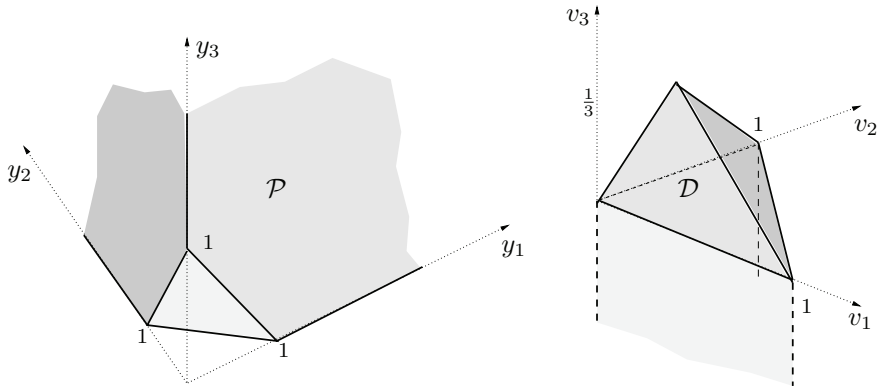


FIG. 4.2. *The three weakly C-minimal vertices of $\mathcal{P}$ correspond to the three K-maximal facets of $\mathcal{D}$, the six weakly C-minimal edges of $\mathcal{P}$ correspond to the six K-maximal edges of $\mathcal{D}$, and the four weakly C-minimal facets of $\mathcal{P}$ correspond to the four K-maximal vertices of $\mathcal{D}$.*

5. Proof of the main result. The proof of the main result is based on several auxiliary assertions, which are given below. There, the following pairs of dual scalar linear optimization problems, depending on parameters $v, y \in \mathbb{R}^q$, play an important role:

$$(\mathrm{P}_1(v)) \quad \min_{x\in\mathcal{X}} c(v)^T Mx \qquad \mathcal{X} := \{x \in \mathbb{R}^n \mid Ax \geq b\},$$

$$(\mathrm{D}_1(v)) \quad \max_{u\in\mathcal{T}(v)} b^T u \qquad \mathcal{T}(v) := \left\{u \in \mathbb{R}^m \mid u \geq 0, \;\; A^T u = M^T c(v)\right\},$$

$$(\mathrm{P}_2(y)) \quad \min_{(x,z)\in\mathcal{S}(y)} z \qquad \mathcal{S}(y) := \{(x,z) \in \mathbb{R}^n \times \mathbb{R} \mid Ax \geq b, \;\; Mx - kz \leq y\},$$

$$(\mathrm{D}_2(y)) \quad \max_{(u,c)\in\mathcal{U}} (b^T u - y^T c)$$

$$\mathcal{U} := \left\{(u,c) \in \mathbb{R}^m \times \mathbb{R}^q \mid (u,c) \geq 0, \;\; A^T u = M^T c, \;\; k^T c = 1\right\}.$$

Note that with the above notation it holds that

$$\mathcal{D} = \left\{v \in \mathbb{R}^q \mid c(v) \geq 0, \exists u \in \mathcal{T}(v) : b^T u \geq v_q\right\}. \tag{5.1}$$

We start with a characterization of weakly C-minimal points of $\mathcal{P}$.

LEMMA 5.1. *The following three statements are equivalent.*

(i) $y^0 \in \operatorname{Min}_C \mathcal{P}$.

(ii) *There is some $x^0 \in \mathbb{R}^n$ such that $(x^0, 0)$ solves $(\mathrm{P}_2(y^0))$.*

(iii) *There is some* $(u^0, c^0) \in \mathcal{U}$ *with* $b^T u^0 = {y^0}^T c^0$ *solving* $(\mathrm{D}_2(y^0))$.

Proof. (ii)⇒(i). If $(x^0, 0)$ solves $(\mathrm{P}_2(y^0))$, then $x^0 \in \mathcal{X}$ and $Mx^0 \leq y^0$; hence $y^0 \in \mathcal{P}$. Assume that there is some $y \in \mathcal{P}$ (i.e., there is some $x \in \mathcal{X}$ with $Mx \leq y$) with $y < y^0$; then there is some $z < 0$ such that $y \leq y^0 + kz$, whence $Mx - kz \leq y - kz \leq y^0$. Thus we have $(x, z) \in \mathcal{S}(y^0)$, where $z < 0$. This contradicts the optimality of $(x^0, 0)$.

(i)⇒(ii). If $y^0 \in \mathrm{Min}_C \mathcal{P}$, then there exists some $x^0 \in \mathcal{X}$ with $Mx^0 \leq y^0$; i.e., $(x^0, 0) \in \mathcal{S}(y^0)$. Assume that there is some $(x, z) \in \mathcal{S}(y^0)$ with $z < 0$. Let $y := y^0 + zk$. Then $y < y^0$ and $Mx \leq y^0 + kz = y$; i.e., $y \in \mathcal{P}$. This contradicts y^0 being weakly C-minimal.

(ii)⇔(iii). This holds by duality of $(\mathrm{P}_2(y^0))$ and $(\mathrm{D}_2(y^0))$. ∎

LEMMA 5.2. *Every K-maximal proper face of $\mathcal{D}$ contains a vertex.*

Proof. Let $\mathcal{F}^*$ be a K-maximal proper face of $\mathcal{D}$. It suffices to show that $\mathcal{F}^*$ contains no lines ([16, Cor. 18.5.3]). Assume on the contrary that $\mathcal{F}^*$ contains a line; i.e., there are $\bar{v} \in \mathcal{F}^*$ and $\psi \in \mathbb{R}^q \setminus \{0\}$ such that $\bar{v} + \lambda\psi \in \mathcal{F}^*$ for all $\lambda \in \mathbb{R}$. Since for every $v \in \mathcal{F}^* \subseteq \mathcal{D}$ it holds that $v_1 \geq 0, \ldots, v_{q-1} \geq 0$, we have $\psi_1 = \cdots = \psi_{q-1} = 0$. Thus, $\psi \neq 0$ implies $K \subseteq \{\lambda\psi \mid \lambda \in \mathbb{R}\}$. We get $\{\bar{v}\} + K \subseteq \mathcal{F}^*$, contradicting the K-maximality of $\mathcal{F}^*$. ∎

LEMMA 5.3. *Consider a hyperplane* $\mathcal{H}^* := \{v \in \mathbb{R}^q \mid {c^*}^T v = \gamma\}$. *Then the following statements are equivalent.*

(i) *$\mathcal{H}^*$ is a supporting hyperplane to $\mathcal{D}$ such that $\mathcal{H}^* \cap \mathcal{D}$ is K-maximal.*

(ii) *$\mathcal{H}^*$ is a supporting hyperplane to $D[\mathcal{U}]$, and $c_q^* < 0$.*

Proof. (i) ⇒ (ii). If $\mathcal{H}^*$ is a supporting hyperplane to $\mathcal{D}$, then there is some $v^0 \in \mathcal{D}$ with ${c^*}^T v^0 = \gamma$, and for $v \in \mathcal{D}$ it holds that ${c^*}^T v \geq \gamma$. By definition of $\mathcal{D}$ we have $\bar{v} := v^0 - e_q \in \mathcal{D}$ ($e_q = (0, \ldots, 0, 1)^T$), implying that $c_q^* \leq 0$. Since $c_q^* = 0$ would imply $\bar{v} \in \mathcal{H}^* \cap \mathcal{D}$ and $v^0 \in (\bar{v} + K \setminus \{0\}) \cap \mathcal{D}$, contradicting the maximality of $\mathcal{H}^* \cap \mathcal{D}$, we conclude that $c_q^* < 0$. As $v^0 \in \mathcal{D}$, there are $v^1 \in D[\mathcal{U}] \subseteq \mathcal{D}$ and $z \geq 0$ such that $v^0 = v^1 - e_q z$. Hence ${c^*}^T v^1 = {c^*}^T v^0 + c_q^* z \leq \gamma$. This implies ${c^*}^T v^1 = \gamma$. Therefore $\mathcal{H}^*$ is a supporting hyperplane to $D[\mathcal{U}]$.

(ii) ⇒ (i). If $\mathcal{H}^*$ is a supporting hyperplane to $D[\mathcal{U}]$, then there is some $v^0 \in D[\mathcal{U}]$ with ${c^*}^T v^0 = \gamma$ and for all $v \in D[\mathcal{U}]$ it holds that ${c^*}^T v \geq \gamma$. Since $c_q^* < 0$, it follows that ${c^*}^T v \geq \gamma$ for all $v \in D[\mathcal{U}] - K = \mathcal{D}$. By $v^0 \in \mathcal{D}$ and ${c^*}^T v^0 = \gamma$ we conclude that $\mathcal{H}^*$ is a supporting hyperplane to $\mathcal{D}$.

In order to show that $\mathcal{H}^* \cap \mathcal{D}$ is K-maximal, let $v^0 \in \mathcal{H}^* \cap \mathcal{D}$ be given. Hence, ${c^*}^T v^0 = \gamma$. For every $v \in v^0 + K \setminus \{0\}$ it holds that ${c^*}^T v < \gamma$, because of $c_q^* < 0$. Since ${c^*}^T v \geq \gamma$ for all $v \in \mathcal{D}$, we obtain $(v^0 + K \setminus \{0\}) \cap \mathcal{D} = \emptyset$. ∎

LEMMA 5.4. *Let $y \in \mathbb{R}^q$. The following statements are equivalent.*

(i) *y is a weakly C-minimal point of $\mathcal{P}$.*

(ii) *$\mathcal{H}^*(y) \cap \mathcal{D}$ is a K-maximal proper face of $\mathcal{D}$.*

Moreover, for every K-maximal proper face $\mathcal{F}^*$ of $\mathcal{D}$ there exists some $y \in \mathbb{R}^q$ such that $\mathcal{F}^* = \mathcal{H}^*(y) \cap \mathcal{D}$.

Proof. By Lemma 5.1, (i) is equivalent to the following:

(iii) There exists some $(u^0, c^0) \in \mathcal{U}$ with $y^T c^0 = b^T u^0$ solving $(\mathrm{D}_2(y))$.

Taking into account (3.1), we see that (iii) is equivalent to the following:

(iv) $\varphi(y, v) \geq 0$ for all $v \in D[\mathcal{U}]$, and there exists some $v^0 \in D[\mathcal{U}]$ with $\varphi(y, v^0) = 0$.

Statement (iv) is equivalent to the following:

(v) $\mathcal{H}^*(y)$ is a supporting hyperplane to $D[\mathcal{U}]$.

Regarding the fact that $\mathcal{H}^*(y) = \left\{v \in \mathbb{R}^q \mid c^*(y)^T v = -y_q\right\}$ with $c^*(y)_q = -1 < 0$, (v) is equivalent to (ii) by Lemma 5.3.

Let $\mathcal{F}^*$ be a K-maximal proper face of $\mathcal{D}$. Then there exists a supporting hyperplane $\mathcal{H}^* := \{v \in \mathbb{R}^q \mid {c^*}^T v = \gamma\}$ (i.e., $c^* \neq 0$) to $\mathcal{D}$ such that $\mathcal{F}^* = \mathcal{H}^* \cap \mathcal{D}$. By Lemma 5.3, we have $c_q^* < 0$. Setting

$$y := \left(\frac{\gamma - c_1^*}{c_q^*}, \ldots, \frac{\gamma - c_{q-1}^*}{c_q^*}, \frac{\gamma}{c_q^*} \right)^T,$$

we obtain $\mathcal{H}^* = \mathcal{H}^*(y)$. Hence $\mathcal{F}^* = \mathcal{H}^*(y) \cap \mathcal{D}$. □

LEMMA 5.5. *Consider a hyperplane* $\mathcal{H} := \{y \in \mathbb{R}^q \mid c^T y = \gamma\}$. *The following statements are equivalent.*

(i) $\mathcal{H}$ *is a supporting hyperplane to* $\mathcal{P}$.

(ii) $c \geq 0$, *and* $\mathcal{H}$ *is a supporting hyperplane to* $M[\mathcal{X}]$.

Proof. (i) $\Rightarrow$ (ii). If $\mathcal{H}$ is a supporting hyperplane to $\mathcal{P}$, then there is some $y^0 \in \mathcal{P}$ with $c^T y^0 = \gamma$ and for all $y \in \mathcal{P}$ it holds that $c^T y \geq \gamma$. By the definition of $\mathcal{P}$ we have $y^0 + w \in \mathcal{P}$ for all $w \in C = \mathbb{R}^q_+$; hence $c^T w \geq 0$ for all $w \in \mathbb{R}^q_+$. This implies $c \geq 0$. Since $y^0 \in \mathcal{P}$, there is $y^1 \in M[\mathcal{X}] \subseteq \mathcal{P}$ and $w \in C$ such that $y^0 = y^1 + w$. Hence $c^T y^1 = c^T y^0 - c^T w \leq \gamma$. This implies $c^T y^1 = \gamma$. Therefore $\mathcal{H}$ is a supporting hyperplane to $M[\mathcal{X}]$.

(ii) $\Rightarrow$ (i). If $\mathcal{H}$ is a supporting hyperplane to $M[\mathcal{X}]$, then there is some $y^0 \in M[\mathcal{X}]$ with $c^T y^0 = \gamma$, and for all $y \in M[\mathcal{X}]$ it holds that $c^T y \geq \gamma$. Since $c \geq 0$, it follows that $c^T y \geq \gamma$ for all $y \in M[\mathcal{X}] + \mathbb{R}^q_+$. By $y^0 \in \mathcal{P}$ and $c^T y^0 = \gamma$ we conclude that $\mathcal{H}$ is a supporting hyperplane to $\mathcal{P}$. □

LEMMA 5.6. *Every proper face of* $\mathcal{P}$ *is weakly* C*-minimal.*

Proof. Let $\mathcal{F}$ be a proper face of $\mathcal{P}$. There is a supporting hyperplane $\mathcal{H} := \{y \in \mathbb{R}^q \mid c^T y = \gamma\}$ (i.e., $c \neq 0$) to $\mathcal{P}$ such that $\mathcal{F} = \mathcal{H} \cap \mathcal{P}$. By Lemma 5.5 we have $c \geq 0$. Let $y \in \mathcal{F}$; then $y \in \mathcal{P}$ implying the existence of $x^0 \in \mathcal{X}$ such that $Mx^0 \leq y$; i.e., $(x^0, 0) \in \mathcal{S}(y)$ and $c^T y = \gamma$. Suppose that there are $x \in \mathcal{X}$ and $z < 0$ such that $Mx - kz \leq y$, i.e., $Mx < y$. Since $\mathcal{H} = \{y \in \mathbb{R}^q \mid c^T y = \gamma\}$ is a supporting hyperplane to $\mathcal{P}$ and $Mx \in \mathcal{P}$, we have $\gamma \leq c^T Mx < c^T y = \gamma$, a contradiction. Hence $(x, 0)$ solves $(\mathrm{P}_2(y))$. By Lemma 5.1 this implies that $y \in \mathrm{Min}_C \mathcal{P}$. □

LEMMA 5.7. *Let* $v \in \mathbb{R}^q$. *The following statements are equivalent.*

(i) v *is a* K*-maximal point of* $\mathcal{D}$.

(ii) $\mathcal{H}(v) \cap \mathcal{P}$ *is a weakly* C*-minimal proper face of* $\mathcal{P}$.

Moreover, for every proper face $\mathcal{F}$ *of* $\mathcal{P}$ *there exists some* $v \in \mathbb{R}^q$ *such that* $\mathcal{F} = \mathcal{H}(v) \cap \mathcal{P}$.

Proof. Taking into account (5.1), we conclude that (i) is equivalent to the following:

(iii) $c(v) \geq 0$, and there exists some $u^0 \in \mathbb{R}^m$ solving $(\mathrm{D}_1(v))$ such that $v_q = b^T u^0$.

By duality between $(\mathrm{P}_1(v))$ and $(\mathrm{D}_1(v))$, (iii) is equivalent to the following:

(iv) $c(v) \geq 0$, and there exists some $x^0 \in \mathbb{R}^n$ solving $(\mathrm{P}_1(v))$ such that $v_q = c(v)^T Mx^0$.

Statement (iv) is equivalent to the following:

(v) $c(v) \geq 0$, and $\mathcal{H}(v)$ is a supporting hyperplane to $M[\mathcal{X}]$.

By Lemmas 5.5 and 5.6, (v) is equivalent to (ii).

To show the last conclusion, let $\mathcal{F}$ be a proper face of $\mathcal{P}$. Hence there exists some supporting hyperplane $\mathcal{H} := \{y \in \mathbb{R}^q \mid c^T y = \gamma\}$ (i.e., $c \neq 0$) to $\mathcal{P}$ such that $\mathcal{F} = \mathcal{H} \cap \mathcal{P}$. By Lemma 5.5, we have $c \geq 0$. Without loss of generality we can assume that $k^T c = 1$ ($k = (1, \ldots, 1)^T$). Setting $v_i := c_i$ for $i = 1, \ldots, q-1$ and $v_q := \gamma$, we have $\mathcal{H} = \mathcal{H}(v)$. Hence $\mathcal{F} = \mathcal{H}(v) \cap \mathcal{P}$. □

Now we are able to give the proof of our main result.

Proof of Theorem 3.1. (a) We show that if $\mathcal{F}^*$ is a K-maximal proper face of $\mathcal{D}$, then $\Psi(\mathcal{F}^*)$ is a weakly C-minimal proper face of $\mathcal{P}$. By Lemma 5.7, $\mathcal{H}(v) \cap \mathcal{P}$ is a weakly C-minimal proper face of $\mathcal{P}$ for each $v \in \mathcal{F}^*$; hence $\Psi(\mathcal{F}^*)$ is a weakly C-minimal face of $\mathcal{P}$. It remains to show that $\Psi(\mathcal{F}^*)$ is nonempty. By Lemma 5.4 there is some $y^0 \in \operatorname{Min}_C \mathcal{P}$ such that $\mathcal{F}^* = \mathcal{H}^*(y^0) \cap \mathcal{D}$; hence $y^0 \in \Psi(\mathcal{F}^*)$.

(b) We prove that $\Psi^*(\mathcal{F}) := \bigcap_{y \in \mathcal{F}} \mathcal{H}^*(y) \cap \mathcal{D}$ is a K-maximal proper face of $\mathcal{D}$ if $\mathcal{F}$ is a weakly C-minimal proper face of $\mathcal{P}$. By Lemma 5.4, $\mathcal{H}^*(y) \cap \mathcal{D}$ is a K-maximal proper face of $\mathcal{D}$ for each $y \in \mathcal{F}$. Hence $\Psi^*(\mathcal{F})$ is a K-maximal proper face of $\mathcal{D}$ if this set is nonempty. Indeed, by Lemma 5.7, there is some $v^0 \in \operatorname{Max}_K \mathcal{D}$ such that $\mathcal{F} = \mathcal{H}(v^0) \cap \mathcal{P}$ implying $v^0 \in \Psi^*(\mathcal{F})$.

(c) In order to show that Ψ is a bijection and that $\Psi^{-1}(\mathcal{F}) = \bigcap_{y \in \mathcal{F}} \mathcal{H}^*(y) \cap \mathcal{D} =: \Psi^*(\mathcal{F})$, we have to show the following two statements: (c_1) $\Psi^*(\Psi(\mathcal{F}^*)) = \mathcal{F}^*$ for all K-maximal proper faces $\mathcal{F}^*$ of $\mathcal{D}$ and (c_2) $\Psi(\Psi^*(\mathcal{F})) = \mathcal{F}$ for all weakly C-minimal proper faces $\mathcal{F}$ of $\mathcal{P}$.

(c_1) First we show that $\mathcal{F}^* \subseteq \Psi^*(\Psi(\mathcal{F}^*))$. Assume the contrary; i.e., there is some $v^0 \in \mathcal{F}^*$ such that $v^0 \notin \Psi^*(\Psi(\mathcal{F}^*))$. Hence there exists some $y^0 \in \Psi(\mathcal{F}^*)$ such that $v^0 \notin \mathcal{H}^*(y^0) \cap \mathcal{D}$. This implies $v^0 \notin \mathcal{H}^*(y^0)$ since $v^0 \in \mathcal{D}$. It follows that $y^0 \notin \mathcal{H}(v^0)$, whence $y^0 \notin \Psi(\mathcal{F}^*)$, a contradiction. To show the opposite inclusion, let $y^0 \in \operatorname{Min}_C \mathcal{P}$ such that $\mathcal{F}^* = \mathcal{H}^*(y^0) \cap \mathcal{D}$. The existence of such a point y^0 is ensured by Lemma 5.4. It follows that $y^0 \in \Psi(\mathcal{F}^*)$. Hence $\Psi^*(\Psi(\mathcal{F}^*)) \subseteq \mathcal{H}^*(y^0) \cap \mathcal{D} = \mathcal{F}^*$.

(c_2) The proof works analogously using Lemma 5.7 instead of Lemma 5.4.

(d) Obviously, Ψ is inclusion-reversing.

(e) It remains to prove that $\dim \mathcal{F}^* + \dim \Psi(\mathcal{F}^*) = q-1$ for all K-maximal proper faces $\mathcal{F}^*$ of $\mathcal{D}$. Consider some fixed $\mathcal{F}^*$, and set $r := \dim \mathcal{F}^*$ and $s := \dim \Psi(\mathcal{F}^*)$. By the first part of the proof, $\mathcal{F} := \Psi(\mathcal{F}^*)$ is a weakly C-minimal face of $\mathcal{P}$. Hence there exist proper faces $\mathcal{F} \subsetneq \mathcal{F}_1 \subsetneq \mathcal{F}_2 \subsetneq \cdots \subsetneq \mathcal{F}_{q-1-s}$ (all of them being weakly C-minimal by Lemma 5.6) such that $\dim \mathcal{F}_{q-1-s} = q-1$. From the properties of Ψ, we conclude that $0 \le \dim \Psi^{-1}(\mathcal{F}_{q-1-s}) \le r - (q-1-s)$. Hence $r + s \ge q-1$. Since every K-maximal face of $\mathcal{D}$ has a vertex (Lemma 5.2), there are K-maximal faces $\mathcal{F}^* \supsetneq \mathcal{F}_1^* \supsetneq \mathcal{F}_2^* \supsetneq \cdots \supsetneq \mathcal{F}_r^*$ such that $\dim \mathcal{F}_r^* = 0$. It follows that $s + r \le \dim \Psi(\mathcal{F}_r^*) \le q-1$. Together we have $s + r = q - 1$. □

REFERENCES

[1] H. P. Benson, *An outer approximation algorithm for generating all efficient extreme points in the outcome set of a multiple objective linear programming problem*, J. Global Optim., 13 (1998), pp. 1–24.

[2] S. Brumelle, *Duality for multiple objective convex programs*, Math. Oper. Res., 6 (1981), pp. 159–172.

[3] M. Ehrgott, A. Löhne, and L. Shao, *A dual variant of Benson's outer approximation algorithm*, J. Global Optim., submitted.

[4] A. Göpfert and R. Nehse, *Vektoroptimierung*, Teubner, Leipzig, Germany, 1990.

[5] B. Grünbaum, *Convex Polytopes*, 2nd ed., Grad. Texts in Math. 221, V. Kaibel, V. Klee, and G. M. Ziegler, eds., Springer-Verlag, New York, 2003.

[6] A. Hamel, F. Heyde, A. Löhne, C. Tammer, and K. Winkler, *Closing the duality gap in linear vector optimization*, J. Convex Anal., 11 (2004), pp. 163–178.

[7] F. Heyde, A. Löhne, and C. Tammer, *The attainment of the solution of the dual program in vertices for vectorial linear programs*, Proceedings of the 7th International Conference on Multi-Objective Programming and Goal Programming, Tours, France, 2006, accepted.

[8] F. Heyde, A. Löhne, and C. Tammer, *Set-valued duality theory for multiple objective linear programs and application to mathematical finance*, Math. Methods Oper. Res., submitted.

[9] H. Isermann, *On some relations between a dual pair of multiple objective linear programs*, Z. Operations Res. Ser. A-B, 22 (1978), pp. A33–A41.

[10] H. Isermann, *Duality in multiple objective linear programming*, in Multiple Criteria Problem Solving (Buffalo, NY, 1977), Lect. Notes Econ. Math. Syst. 155, Springer-Verlag, New York, 1978, pp. 274–285.
[11] J. Jahn, *Mathematical Vector Optimization in Partially Ordered Linear Spaces*, Verlag Peter Lang, Frankfurt am Main, Bern, New York, 1986.
[12] J. Jahn, *Vector Optimization. Theory, Applications, and Extensions*, Springer-Verlag, Berlin, 2004.
[13] J. S. H. Kornbluth, *Duality, indifference and sensitivity analysis in multiple objective linear programming*, Operational Res. Quart., 25 (1974), pp. 599–614.
[14] A. Löhne and C. Tammer, *A new approach to duality in linear vector optimization*, Optimization, 56 (2007), pp. 221–239.
[15] D. T. Luc, *Theory of Vector Optimization*, Lect. Notes Econ. Math. Sci. 319, Springer-Verlag, Berlin, 1988.
[16] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1972.
[17] W. Rödder, *A generalized saddlepoint theory; its application to duality theory for linear vector optimum problems*, European J. Oper. Res., 1 (1977), pp. 55–59.
[18] R. Webster, *Convexity*, Oxford Sci. Publ., Oxford University Press, New York, 1994.

SIAM J. OPTIM.
Vol. 19, No. 2, pp. 846–862

CONVERGENCE RATE OF AN OPTIMIZATION ALGORITHM FOR MINIMIZING QUADRATIC FUNCTIONS WITH SEPARABLE CONVEX CONSTRAINTS*

RADEK KUČERA†

Abstract. A new active set algorithm for minimizing quadratic functions with separable convex constraints is proposed by combining the conjugate gradient method with the projected gradient. It generalizes recently developed algorithms of quadratic programming constrained by simple bounds. A linear convergence rate in terms of the Hessian spectral condition number is proven. Numerical experiments, including the frictional three-dimensional (3D) contact problems of linear elasticity, illustrate the computational performance.

Key words. quadratic function, separable convex constraints, active set, conjugate gradient method, projected gradient, convergence rate

AMS subject classifications. 65K05, 90C25

DOI. 10.1137/060670456

1. Introduction.

We shall be concerned with solving

$$\min_{x\in\Omega} f(x), \tag{1.1}$$

where $f(x) = \frac{1}{2}\, x^\top Ax - x^\top b$, $A \in \mathbb{R}^{n\times n}$ is symmetric positive definite, $b \in \mathbb{R}^n$, $\Omega = \Omega_1 \times \cdots \times \Omega_m$, and $\Omega_i = \{\boldsymbol{x}_i \in \mathbb{R}^{n_i} : f_i(\boldsymbol{x}_i) \le 0\}$ are defined by continuously differentiable convex functions $f_i : \mathbb{R}^{n_i} \to \mathbb{R}$ so that $n_i \ge 1$, $\sum_{i=1}^m n_i = n$. Let us note that the feasible set Ω is *separable* in the sense that each part $\boldsymbol{x}_i$ of $x = (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_m^\top)^\top$ is subject to one constraint $\boldsymbol{x}_i \in \Omega_i$.

This problem includes several independently investigated subproblems originating, for instance, in duality-based methods for the solution of contact problems of linear elasticity:

- If $n_i = 1$ and $f_i(x_i) \equiv l_i - x_i$ with l_i given, we obtain the *simple bound*

$$l_i \le x_i \tag{1.2}$$

  arising from two-dimensional contact problems [4].
- If $n_i = 2$, $\boldsymbol{x}_i = (x_{2i-1}, x_{2i})^\top$ and $f_i(\boldsymbol{x}_i) \equiv x_{2i-1}^2 + x_{2i}^2 - r_i^2$ with r_i given, we arrive at

$$x_{2i-1}^2 + x_{2i}^2 \le r_i^2 \tag{1.3}$$

  that can be interpreted as the *circular constraint*. A source of such constraints is an isotropic friction law for three-dimensional (3D) contact problems [8].
- If an anisotropic friction law is considered and if a finite element tearing and interconnecting (FETI) domain decomposition method is used for solving 3D

*Received by the editors September 22, 2006; accepted for publication (in revised form) April 3, 2008; published electronically August 1, 2008. This work was supported by grant 101/08/0574 of the Grant Agency of the Czech Republic and by the Research Project MSM6198910027 of the Czech Ministry of Education.

http://www.siam.org/journals/siopt/19-2/67045.html

†Department of Mathematics and Descriptive Geometry, VŠB-Technical University of Ostrava, 17. listopadu 15, CZ-70833 Ostrava-Poruba, Czech Republic (radek.kucera@vsb.cz).

contact problems [11], then (1.3) is replaced by the *ellipsoidal constraint*,

$$(\boldsymbol{x}_i - \boldsymbol{s}_i)^\top \boldsymbol{F}_i (\boldsymbol{x}_i - \boldsymbol{s}_i) \le r_i^2,$$

with $\boldsymbol{F}_i \in \mathbb{R}^{2\times 2}$ positive definite and $\boldsymbol{s}_i = (s_{2i-1}, s_{2i})^\top$ given.

– Finally let us note that the unconstrained case may be described by $f_i(\boldsymbol{x}_i) = -1$.

Through the whole paper, we shall have in mind large-scale problems in which the Hessian matrix A is not formed explicitly. In this case, an iterative method is a suitable tool for solving (1.1) since it is the only action of A which is needed. A class of efficient algorithms appropriate for our research is based on the active set strategy that dates back at least to Polyak [13]. His algorithm solves quadratic programming problems constrained by simple bounds (1.2) using a restarted conjugate gradient method. After each start, a subset of variables is fixed at bounds (the active set) and the conjugate gradient method minimizes f with respect to remaining variables. The "inner" minimization is terminated if either the minimum is reached or an infeasible iteration is generated. In the first case, some of the fixed variables are released while, in the second case, new variables are added to the active set. In both cases, the value of f decreases so that the active set can never reappear, and therefore the algorithm converges in a finite number of steps.

As the Polyak algorithm suffers from several drawbacks, it was modified in order to exclude doubts about its efficiency; see discussions in [1, 5]. Here, we mention two improvements relevant for our more general constraints. Firstly, the exact solution of auxiliary "inner" problems can be replaced by an inexact one. We shall use a theoretically supported strategy of an adaptive precision control presented by Friedlander and Martínez [6] and by Dostál [2]. The basic idea is to control the "inner" precision by a ratio of the norms of a violation of the Karush–Kuhn–Tucker (KKT) conditions at fixed and free variables. Secondly, the qualitative progress has been achieved by Dostál and Schöberl [5] using the projected gradient for expanding the active set. It permits rapid changes in the active set without the necessity to perform computationally expensive steps (e.g., backtracking). Moreover, the algorithm has a linear convergence rate in terms of the spectral condition number of the Hessian matrix A.

The scheme of the later algorithm was successfully extended for solving 3D contact problems with an isotropic friction in [11]. As the constraints are circular (1.3), the algorithm does search not only for the active set corresponding to the solution, but also for the positions of the pairs $(x_{2i-1}, x_{2i})^\top$ lying on the boundaries of the circles $x_{2i-1}^2 + x_{2i}^2 = r_i^2$. It is easily seen that the finite terminating property cannot be expected in such cases, and therefore the convergence was proven in [10] by different arguments. On the other hand, it is relatively surprising that a linear convergence rate may be derived like for the simple bound case. This proof is the main goal of the paper.

Let us briefly outline the structure of the paper. After introducing notations in section 2, we prove that the KKT optimality conditions are equivalent to the zero projected gradient. Our algorithm for solving (1.1) is proposed in section 3 in a form suitable for theoretical analysis. Section 4 summarizes auxiliary statements on the projected gradient, while section 5 gives the main result of the paper concerning a linear convergence rate. A practical implementation of the algorithm for simple bounds (1.2) and circular constraints (1.3) is discussed in section 6, and finally, section 7 presents the results of numerical experiments.

2. Preliminaries and notations. We shall always assume that the feasible set Ω in (1.1) is nonempty. As f is the strictly convex function and Ω is the convex set,

the existence of a unique solution to (1.1) is guaranteed. We shall denote it by x^*. It is well known that x^* is fully determined by the KKT conditions [12]. Before giving their appropriate form, we shall introduce notations.

Recall that a continuously differentiable function $F : \Omega \to \mathbb{R}$ is convex iff

$$F(y) - F(x) \geq (y - x)^\top \nabla F(x) \quad \forall x, y \in \Omega, \tag{2.1}$$

where ∇F denotes the gradient of F. The gradient of the objective function f at x shall be denoted by

$$g = g(x) = Ax - b.$$

Let $\mathcal{M}$ denote the set of indices so that

$$\mathcal{M} = \{1, \ldots, m\}.$$

We shall use the following convention: if $x \in \mathbb{R}^n$ is a vector, then $x_i \in \mathbb{R}$ is its ith entry, $1 \leq i \leq n$, and $\boldsymbol{x}_i \in \mathbb{R}^{n_i}$ is its ith segment, $i \in \mathcal{M}$, so that $x = (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_m^\top)^\top$. We shall denote by $\|\boldsymbol{x}_i\|$ the Euclidean norm of $\boldsymbol{x}_i$.

In order to exclude pathological situations, we shall assume without loss of generality that f_i are not identically zero in Ω_i. In that case Ω_i has a nonempty interior, $\operatorname{int}\Omega_i = \{\boldsymbol{x}_i \in \mathbb{R}^{n_i} : f_i(\boldsymbol{x}_i) < 0\}$, and a possibly nonempty boundary $\partial\Omega_i = \{\boldsymbol{x}_i \in \mathbb{R}^{n_i} : f_i(\boldsymbol{x}_i) = 0\}$. The convexity of f_i implies that if $\partial\Omega_i \neq \emptyset$ and $\boldsymbol{x}_i \in \partial\Omega_i$, then $\nabla f_i(\boldsymbol{x}_i)$ is the outward normal vector to $\partial\Omega_i$ at $\boldsymbol{x}_i$; see Figure 2.1.a.

It is well known that the solution x^* to (1.1) is characterized by the existence of Lagrange multipliers λ_i^*, $i \in \mathcal{M}$, such that [12]

$$\boldsymbol{g}_i^* + \lambda_i^* \nabla f_i(\boldsymbol{x}_i^*) = \boldsymbol{0},\ f_i(\boldsymbol{x}_i^*) \leq 0,\ \lambda_i^* \geq 0,\ \lambda_i^* f_i(\boldsymbol{x}_i^*) = 0, \quad i \in \mathcal{M},$$

where $\boldsymbol{g}_i^*$ denotes the ith segment of $g^* = g(x^*)$. After eliminating λ_i^*, we obtain the following theorem.

THEOREM 2.1. *The vector $x^* \in \Omega$ is the solution to* (1.1) *iff for $i \in \mathcal{M}$:*

$$f_i(\boldsymbol{x}_i^*) < 0 \quad \textit{implies} \quad \boldsymbol{g}_i^* = \boldsymbol{0}, \tag{2.2}$$

$$f_i(\boldsymbol{x}_i^*) = 0 \quad \textit{implies} \quad \boldsymbol{g}_i^* + \frac{\|\boldsymbol{g}_i^*\|}{\|\nabla f_i(\boldsymbol{x}_i^*)\|} \nabla f_i(\boldsymbol{x}_i^*) = \boldsymbol{0}. \tag{2.3}$$

Conditions (2.2) and (2.3) are called the *inner KKT conditions* and the *boundary KKT conditions*, respectively.

As the feasible set Ω is separable, the projection $P_\Omega : \mathbb{R}^n \mapsto \Omega$ can be put together by the projections $P_{\Omega_i} : \mathbb{R}^{n_i} \mapsto \Omega_i$, $i \in \mathcal{M}$. Thus we define $P_\Omega(x)$ for any $x \in \mathbb{R}^n$ by

$$P_\Omega(x) = \begin{pmatrix} P_{\Omega_1}(\boldsymbol{x}_1) \\ \vdots \\ P_{\Omega_m}(\boldsymbol{x}_m) \end{pmatrix}, \tag{2.4}$$

where $P_{\Omega_i}(\boldsymbol{x}_i)$ are defined by

$$P_{\Omega_i}(\boldsymbol{x}_i) = \begin{cases} \boldsymbol{x}_i & \text{for } \boldsymbol{x}_i \in \Omega_i, \\ \boldsymbol{z}_i & \text{for } \boldsymbol{x}_i \notin \Omega_i, \end{cases}$$

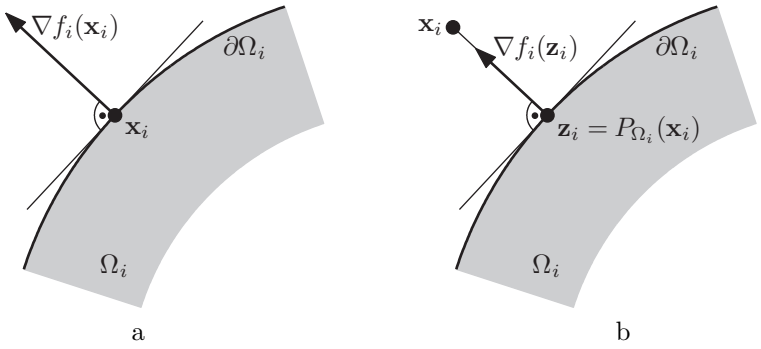


FIG. 2.1. (a) *The outward normal vector to* $\partial\Omega_i$*;* (b) *The projection to* $\partial\Omega_i$*.*

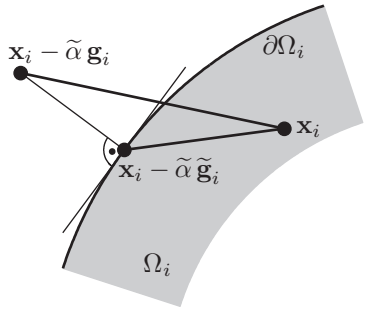


FIG. 2.2. *The projected gradient on* Ω_i*.*

with (see Figure 2.1.b)

$$f_i(\boldsymbol{z}_i) = 0, \tag{2.5}$$

$$\boldsymbol{z}_i + \frac{\|\boldsymbol{x}_i - \boldsymbol{z}_i\|}{\|\nabla f_i(\boldsymbol{z}_i)\|}\nabla f_i(\boldsymbol{z}_i) = \boldsymbol{x}_i. \tag{2.6}$$

Let us note that $\boldsymbol{z}_i$ is well defined by (2.5) and (2.6) because Ω_i is convex. The equality (2.6) yields orthogonality of P_{Ω_i} and, consequently, of P_Ω that is equivalent to the variational inequality

$$(x - P_\Omega(x))^\top (y - P_\Omega(x)) \le 0 \quad \forall x \in \mathbb{R}^n \ \forall y \in \Omega.$$

Let us define the *projected gradient* $\widetilde{g} = \widetilde{g}(x)$ at $x \in \Omega$ for fixed $\widetilde{\alpha} > 0$ by

$$\widetilde{g}(x) = \frac{1}{\widetilde{\alpha}}\left(x - P_\Omega\left(x - \widetilde{\alpha} g(x)\right)\right). \tag{2.7}$$

This definition enables us to describe the projection of the gradient step without P_Ω so that

$$P_\Omega(x - \widetilde{\alpha} g(x)) = x - \widetilde{\alpha}\widetilde{g}(x);$$

see Figure 2.2. In the next theorem, we prove that the zero project gradient represents an alternative optimality criterion to KKT conditions.

THEOREM 2.2. *The vector* $x^* \in \Omega$ *solves* (1.1) *iff* $\widetilde{g}(x^*) = 0$*.*

Proof. The equality $\widetilde{g}(x^*) = 0$ is equivalent to

$$\boldsymbol{x}_i^* = P_{\Omega_i}(\boldsymbol{x}_i^* - \widetilde{\alpha}\boldsymbol{g}_i^*), \quad i \in \mathcal{M}. \tag{2.8}$$

We shall prove that the KKT conditions (2.2) and (2.3) are equivalently satisfied. We distinguish two cases. (i) Let $f_i(\boldsymbol{x}_i^*) < 0$. Then $\boldsymbol{x}_i^* \in \operatorname{int}\Omega_i$, which is equivalent by (2.8) to $\boldsymbol{x}_i^* - \widetilde{\alpha}\boldsymbol{g}_i^* \in \operatorname{int}\Omega_i$. Therefore

$$\boldsymbol{x}_i^* = P_{\Omega_i}(\boldsymbol{x}_i^* - \widetilde{\alpha}\boldsymbol{g}_i^*) = \boldsymbol{x}_i^* - \widetilde{\alpha}\boldsymbol{g}_i^*$$

so that $\boldsymbol{g}_i^* = \boldsymbol{0}$, and therefore (2.2) holds. (ii) Let $f_i(\boldsymbol{x}_i^*) = 0$. Using (2.6) with $\boldsymbol{x}_i$ replaced by $\boldsymbol{x}_i^* - \widetilde{\alpha}\boldsymbol{g}_i^*$ and with $\boldsymbol{z}_i$ replaced by $\boldsymbol{x}_i^*$, we obtain after simple manipulation

$$\boldsymbol{g}_i^* + \frac{\|\boldsymbol{g}_i^*\|}{\|\nabla f_i(\boldsymbol{x}_i^*)\|}\nabla f_i(\boldsymbol{x}_i^*) = \boldsymbol{0}$$

so that (2.3) holds. ∎

We shall decompose $\mathcal{M}$ at $x \in \Omega$ on the *free set* $\mathcal{F}(x)$ and the *active set* $\mathcal{A}(x)$ as

$$\mathcal{F}(x) = \{i \in \mathcal{M} : f_i(\boldsymbol{x}_i) < 0\},$$
$$\mathcal{A}(x) = \{i \in \mathcal{M} : f_i(\boldsymbol{x}_i) = 0\}.$$

Analogously, we can decompose $\widetilde{g}$ on the *projected free gradient* $\widetilde{\varphi} = \widetilde{\varphi}(x)$ and the *projected boundary gradient* $\widetilde{\beta} = \widetilde{\beta}(x)$ so that

$$\widetilde{\boldsymbol{\varphi}}_i = \widetilde{\boldsymbol{g}}_i \quad \text{for } i \in \mathcal{F}(x), \quad \widetilde{\boldsymbol{\varphi}}_i = \boldsymbol{0} \quad \text{for } i \in \mathcal{A}(x), \tag{2.9}$$

$$\widetilde{\boldsymbol{\beta}}_i = \boldsymbol{0} \quad \text{for } i \in \mathcal{F}(x), \quad \widetilde{\boldsymbol{\beta}}_i = \widetilde{\boldsymbol{g}}_i \quad \text{for } i \in \mathcal{A}(x). \tag{2.10}$$

Thus, the inner KKT conditions (2.2) are satisfied iff $\widetilde{\varphi}(x) = 0$, and the boundary KKT conditions (2.3) are satisfied iff $\widetilde{\beta}(x) = 0$. Moreover, we define the *free gradient* $\varphi = \varphi(x)$ so that

$$\widetilde{\boldsymbol{\varphi}}_i = \widetilde{\boldsymbol{g}}_i \quad \text{for } i \in \mathcal{F}(x), \quad \widetilde{\boldsymbol{\varphi}}_i = \boldsymbol{0} \quad \text{for } i \in \mathcal{A}(x). \tag{2.11}$$

Finally let us denote the A-energy norm of $x \in \mathbb{R}^n$ by $\|x\|_A$. Thus $\|x\|_A = (x^\top Ax)^{1/2}$, and $\|x\| = \|x\|_I = (x^\top x)^{1/2}$ is the Euclidean norm. The analogous notation will be used for the induced matrix norm so that

$$\kappa(A) = \|A\|\|A^{-1}\|$$

is the spectral condition number of A.

3. Algorithm. In this section we present our algorithm for solving (1.1) in a form convenient for the analysis, while technical details are postponed to section 6. The algorithm exploits a given constant $\Gamma > 0$ to decide on interrupting conjugate gradient iterations and a fixed steplength $\widetilde{\alpha} \in (0, \|A\|^{-1}]$ defining the projected gradient.

We shall combine three steps to generate a sequence of iterates $\{x^k\}$ that approximates the solution to (1.1):

– the *expansion step*, which may add indices to the active set, is defined by

$$x^{k+1} = x^k - \widetilde{\alpha}\widetilde{\varphi}(x^k); \tag{3.1}$$

– the *proportioning step*, which may release indices from the active set, reads as

$$x^{k+1} = x^k - \widetilde{\alpha}\widetilde{\beta}(x^k); \tag{3.2}$$

– the *conjugate gradient step* is given by

$$x^{k+1} = x^k - \alpha_{cg}^k p^k, \quad \alpha_{cg}^k = \frac{g(x^k)^\top p^k}{(p^k)^\top A p^k}, \tag{3.3}$$

where the conjugate gradient directions p^k are constructed recurrently [7].

The conjugate gradient steps are used to carry out a minimization of the objective function f efficiently on the interior of the *face*

$$W_{\mathcal{A}(x^s)} = \{x \in \Omega : \boldsymbol{x}_i = \boldsymbol{x}_i^s \ \text{ for } \ i \in \mathcal{A}(x^s)\},$$

where x^s is determined by the expansion step or by the proportioning step. It requires that the parts of p^k corresponding to the indices of $\mathcal{A}(x^s)$ vanish, i.e., $\boldsymbol{p}_i^k = \mathbf{0}$ for $i \in \mathcal{A}(x^s)$. We can easily fulfill this requirement by adapting the classical recurrence generating p^k. We start (or restart) from $p^s = \varphi(x^s)$ and use

$$p^k = \varphi(x^k) - \gamma^k p^{k-1}, \quad \gamma^k = \frac{\varphi(x^k)^\top A p^{k-1}}{(p^{k-1})^\top A p^{k-1}}, \quad k > s. \tag{3.4}$$

The formulae (3.4) is used while the sequence of the conjugate gradient steps is unbroken. After changing the active set, we must always restart.

Later on, we shall need the following lemma.

LEMMA 3.1. *Let x^{k+1} be generated by the conjugate gradient step. Then*

$$f(x^{k+1}) \le f(x^k - \alpha\varphi(x^k)) \quad \forall \alpha \in \mathbb{R}.$$

Proof. It is a well-known property of the conjugate gradient method [7] that

$$f(x^{k+1}) = \min_{x \in x^s + \mathrm{Span}\{p^s, \dots, p^k\}} f(x). \tag{3.5}$$

As (3.4) implies $\varphi(x^k) \in \mathrm{Span}\{p^s, \dots, p^k\}$ and $x^k \in x^s + \mathrm{Span}\{p^s, \dots, p^{k-1}\}$, we obtain $x^k - \alpha\varphi(x^k) \in x^s + \mathrm{Span}\{p^s, \dots, p^k\}$. The lemma follows using (3.5). ∎

The last ingredient of our algorithm is the *releasing criterion*:

$$\widetilde{\beta}(x^k)^\top g(x^k) \le \Gamma^2 \widetilde{\varphi}(x^k)^\top g(x^k). \tag{3.6}$$

If this inequality holds, we call the iterate x^k *strictly proportional*. The criterion (3.6) is used to decide which of the steps will be performed.

ALGORITHM 3.1. *Let $x^0 \in \Omega$, $\Gamma > 0$, and $\widetilde{\alpha} \in (0, \|A\|^{-1}]$ be given. For $k \ge 0$ and x^k known, choose x^{k+1} by the following rules:*

(i) *If $\widetilde{g}(x^k) = 0$, set $x^{k+1} = x^k$.*

(ii) *If x^k is strictly proportional and $\widetilde{g}(x^k) \ne 0$, try to generate x^{k+1} by the conjugate gradient step. If $x^{k+1} \in \Omega$ and if it does not change the active set, then accept it, else generate x^{k+1} by the expansion step.*

(iii) *If x^k is not strictly proportional, then define x^{k+1} by the proportioning step.*

4. Properties of the projected gradient. In this section we summarize results necessary for the next analysis.

LEMMA 4.1. *It holds that*

$$\|\widetilde{\varphi}(x)\|^2 \le \widetilde{\varphi}(x)^\top g(x), \tag{4.1}$$

$$\|\widetilde{\beta}(x)\|^2 \le \widetilde{\beta}(x)^\top g(x), \tag{4.2}$$

$$\widetilde{\varphi}(x)^\top g(x) \le \varphi(x)^\top g(x). \tag{4.3}$$

Proof. The definition of the projected gradient (2.7) implies that

$$P_{\Omega_i}(\boldsymbol{x}_i - \widetilde{\alpha}\boldsymbol{g}_i) = \boldsymbol{x}_i - \widetilde{\alpha}\widetilde{\boldsymbol{g}}_i.$$

As P_{Ω_i} is orthogonal, we obtain

$$0 \ge (\boldsymbol{x}_i - \widetilde{\alpha}\boldsymbol{g}_i - P_{\Omega_i}(\boldsymbol{x}_i - \widetilde{\alpha}\boldsymbol{g}_i))^\top (\boldsymbol{x}_i - P_{\Omega_i}(\boldsymbol{x}_i - \widetilde{\alpha}\boldsymbol{g}_i)) = \widetilde{\alpha}^2(\widetilde{\boldsymbol{g}}_i - \boldsymbol{g}_i)^\top \widetilde{\boldsymbol{g}}_i$$

so that

$$\|\widetilde{\boldsymbol{g}}_i\|^2 \le \widetilde{\boldsymbol{g}}_i^\top \boldsymbol{g}_i. \tag{4.4}$$

Summing (4.4) over the indices of $\mathcal{F}(x)$ or $\mathcal{A}(x)$, we obtain (4.1) or (4.2), respectively. Using the well-known Cauchy inequality in (4.4), we get $\|\widetilde{\boldsymbol{g}}_i\| \le \|\boldsymbol{g}_i\|$, and therefore

$$\widetilde{\varphi}(x)^\top g(x) = \sum_{i\in\mathcal{F}(x)} \widetilde{\boldsymbol{g}}_i^\top \boldsymbol{g}_i \le \sum_{i\in\mathcal{F}(x)} \|\widetilde{\boldsymbol{g}}_i\|\|\boldsymbol{g}_i\| \le \sum_{i\in\mathcal{F}(x)} \boldsymbol{g}_i^\top \boldsymbol{g}_i = \varphi(x)^\top g(x). \qquad \square$$

LEMMA 4.2. *Let $d \in \mathbb{R}^n$, $d \ne 0$, and $x \in \Omega$. Then*

$$f(x) - f(x - \alpha d) \le \alpha d^\top g(x) \quad \forall \alpha \in \mathbb{R}. \tag{4.5}$$

Moreover, let $d^\top g(x) \ge 0$, and $\alpha_d = d^\top g(x)/d^\top A d$. Then

$$f(x) - f(x - \alpha d) \ge \frac{1}{2}\alpha d^\top g(x) \quad \forall \alpha \in [0, \alpha_d]. \tag{4.6}$$

Proof. The assertion (4.5) is equivalent to the convexity of f (compare with (2.1)). For $\alpha \in [0, \alpha_d]$, we derive

$$\begin{aligned} f(x - \alpha d) &= f(x) - \alpha d^\top g(x) + \frac{1}{2}\alpha^2 d^\top A d \\ &\le f(x) - \alpha d^\top g(x) + \frac{1}{2}\alpha\alpha_d d^\top A d \\ &= f(x) - \frac{1}{2}\alpha d^\top g(x). \qquad \square \end{aligned}$$

COROLLARY 4.3. *If d is replaced in (4.6) by φ, $\widetilde{\varphi}$, and $\widetilde{\beta}$, then the corresponding three inequalities hold for all $\alpha \in [0, \|A\|^{-1}]$.*

Proof. For $d = \varphi(x)$, we have

$$\alpha_d = \frac{\varphi(x)^\top g(x)}{\varphi(x)^\top A \varphi(x)} \ge \frac{\varphi(x)^\top \varphi(x)}{\|A\|\|\varphi(x)\|^2} = \|A\|^{-1}.$$

The same follows using Lemma 4.1 for $d = \widetilde{\varphi}(x)$ and $d = \widetilde{\beta}(x)$. $\square$

LEMMA 4.4. *Let $x \in \Omega$. Then*

$$f(x) - f(x - \alpha\widetilde{g}(x)) \leq \alpha(\widetilde{\varphi}(x)^\top g(x) + \widetilde{\beta}(x)^\top g(x)) \quad \forall \alpha \in \mathbb{R}.$$

Proof. Using (4.5), we obtain

$$f(x) - f(x - \alpha\widetilde{g}(x)) \leq \alpha\widetilde{g}(x)^\top g(x) = \alpha \left(\sum_{i \in \mathcal{F}(x)} \widetilde{\boldsymbol{g}}_i^\top \boldsymbol{g}_i + \sum_{i \in \mathcal{A}(x)} \widetilde{\boldsymbol{g}}_i^\top \boldsymbol{g}_i \right),$$

and the definitions (2.9) and (2.10) complete the proof. ∎

LEMMA 4.5. *Let $x^* \in \Omega$ denote the solution to* (1.1)*, λ_1 denote the smallest eigenvalue of A, and $x \in \Omega$. Then*

$$f(x) - f(x - \widetilde{\alpha}\widetilde{g}(x)) \geq \widetilde{\alpha}\lambda_1(f(x) - f(x^*)) \quad \forall \widetilde{\alpha} \in [0, \|A\|^{-1}].$$

Proof. For $\widetilde{\alpha} \in (0, \|A\|^{-1}]$, we define

$$F(y) = \widetilde{\alpha} f(y) + \frac{1}{2}(y - x)^\top (I - \widetilde{\alpha}A)(y - x).$$

Notice that $I - \widetilde{\alpha}A$ is positive semidefinite. Therefore

$$F(y) \geq \widetilde{\alpha} f(y),$$

$$\nabla F(y) = \widetilde{\alpha} g(y) + (I - \widetilde{\alpha}A)(y - x) = y - x + \widetilde{\alpha} g(x), \tag{4.7}$$

and $F(x) = \widetilde{\alpha} f(x)$, $\nabla F(x) = \widetilde{\alpha} g(x)$. Let us note that the Hessian of F is the identity I. Let $y \in \Omega$ be arbitrary. The convexity of F, (4.7), and the orthogonality of P_Ω yields

$$\begin{aligned} F(y) - F(P_\Omega(x - \widetilde{\alpha}g(x))) &\geq (y - P_\Omega(x - \widetilde{\alpha}g(x)))^\top \nabla F(P_\Omega(x - \widetilde{\alpha}g(x))) \\ &= -(y - P_\Omega(x - \widetilde{\alpha}g(x)))^\top (x - \widetilde{\alpha}g(x) - P_\Omega(x - \widetilde{\alpha}g(x))) \\ &\geq 0 \end{aligned}$$

so that

$$F(y) \geq F(P_\Omega(x - \widetilde{\alpha}g(x))) \quad \forall y \in \Omega.$$

Using $\widetilde{\alpha}\lambda_1 \leq \|A\|^{-1}\lambda_1 \leq 1$ and denoting $d = x^* - x$, we derive

$$\begin{aligned} f(x - \widetilde{\alpha}\widetilde{g}(x)) &= f(P_\Omega(x - \widetilde{\alpha}g(x))) \\ &\leq \widetilde{\alpha}^{-1} F(P_\Omega(x - \widetilde{\alpha}g(x))) \;\leq\; \widetilde{\alpha}^{-1} \min_{y \in \Omega} F(y) \;\leq\; \widetilde{\alpha}^{-1} \min_{t \in [0,1]} F(x + td) \\ &\leq \widetilde{\alpha}^{-1} F(x + \widetilde{\alpha}\lambda_1 d) \;=\; f(x) + \widetilde{\alpha}\lambda_1 d^\top g(x) + \frac{1}{2}\widetilde{\alpha}\lambda_1^2 d^\top d \\ &\leq (1 - \widetilde{\alpha}\lambda_1) f(x) + \widetilde{\alpha}\lambda_1 f(x) + \widetilde{\alpha}\lambda_1 d^\top g(x) + \frac{1}{2}\widetilde{\alpha}\lambda_1 d^\top A d \\ &= (1 - \widetilde{\alpha}\lambda_1) f(x) + \widetilde{\alpha}\lambda_1 f(x^*). \qquad \blacksquare \end{aligned}$$

Lemma 4.4 and Lemma 4.5 immediately imply the following result.

COROLLARY 4.6. *Let $x^* \in \Omega$ denote the solution to* (1.1)*, λ_1 denote the smallest eigenvalue of A, and $x \in \Omega$. Then*

$$\lambda_1(f(x) - f(x^*)) \leq \widetilde{\varphi}(x)^\top g(x) + \widetilde{\beta}(x)^\top g(x)$$

for all $\widetilde{\alpha} \in (0, \|A\|^{-1}]$ defining $\widetilde{\varphi}$ and $\widetilde{\beta}$.

5. Rate of convergence.

THEOREM 5.1. *Let $x^0 \in \Omega$, $\Gamma > 0$, and $\widetilde{\alpha} \in (0, \|A\|^{-1}]$ be given. Let $x^* \in \Omega$ denote the solution to (1.1), λ_1 denote the smallest eigenvalue of A, and $\widehat{\Gamma} = \max\{\Gamma, \Gamma^{-1}\}$. Let $\{x^k\}$ be the sequence generated by Algorithm 3.1. Then*

$$f(x^{k+1}) - f(x^*) \leq \eta \left(f(x^k) - f(x^*) \right), \tag{5.1}$$

where

$$\eta = 1 - \frac{\widetilde{\alpha}\lambda_1}{2 + 2\widehat{\Gamma}^2} < 1. \tag{5.2}$$

The error in the A-energy norm is bounded by

$$\|x^k - x^*\|_A^2 \leq 2\eta^k \left(f(x^0) - f(x^*) \right). \tag{5.3}$$

Proof. We shall estimate separately all three possible steps of Algorithm 3.1. As $\widetilde{\alpha} \in (0, \|A\|^{-1}]$, we can use the bound (4.6) for each of the steps due to Corollary 4.3.

Let us first assume that x^{k+1} is generated by the expansion step (3.1). Using (4.6), we obtain

$$f(x^{k+1}) = f(x^k - \widetilde{\alpha}\widetilde{\varphi}(x^k)) \leq f(x^k) - \frac{1}{2}\,\widetilde{\alpha}\widetilde{\varphi}(x^k)^\top g(x^k). \tag{5.4}$$

If x^{k+1} is generated by the conjugate gradient step (3.3), we use Lemma 3.1, (4.6), and (4.3) so that

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k - \widetilde{\alpha}\varphi(x^k)) \\ &\leq f(x^k) - \frac{1}{2}\,\widetilde{\alpha}\varphi(x^k)^\top g(x^k) \leq f(x^k) - \frac{1}{2}\,\widetilde{\alpha}\widetilde{\varphi}(x^k)^\top g(x^k). \end{aligned} \tag{5.5}$$

Comparing (5.4) and (5.5), we may observe that we have the same estimates for both of the expansion and conjugate gradient steps. These steps are taken only when x^k is strictly proportional, i.e., when

$$\widetilde{\beta}(x^k)^\top g(x^k) \leq \Gamma^2 \widetilde{\varphi}(x^k)^\top g(x^k).$$

After using Corollary 4.6, we get

$$\widetilde{\varphi}(x^k)^\top g(x^k) \geq \frac{\lambda_1}{1 + \Gamma^2}(f(x^k) - f(x^*)). \tag{5.6}$$

The estimates (5.4), (5.5) combined with (5.6) imply that

$$\begin{aligned} f(x^{k+1}) - f(x^*) &\leq f(x^k) - f(x^*) - \frac{1}{2}\,\widetilde{\alpha}\widetilde{\varphi}(x^k)^\top g(x^k) \\ &\leq \left(1 - \frac{\widetilde{\alpha}\lambda_1}{2 + 2\Gamma^2}\right)(f(x^k) - f(x^*)). \end{aligned} \tag{5.7}$$

Let us finally assume that x^{k+1} is generated by the proportioning step (3.2). Again using (4.6), we obtain

$$f(x^{k+1}) = f(x^k - \widetilde{\alpha}\widetilde{\beta}(x^k)) \leq f(x^k) - \frac{1}{2}\,\widetilde{\alpha}\widetilde{\beta}(x^k)^\top g(x^k). \tag{5.8}$$

As the proportioning step is taken when

$$\widetilde{\beta}(x^k)^\top g(x^k) > \Gamma^2 \widetilde{\varphi}(x^k)^\top g(x^k),$$

Corollary 4.6 yields

$$\widetilde{\beta}(x^k)^\top g(x^k) \geq \frac{\lambda_1}{1+\Gamma^{-2}}(f(x^k) - f(x^*)). \tag{5.9}$$

The estimate (5.8) combined with (5.9) implies that

$$f(x^{k+1}) - f(x^*) \leq \left(1 - \frac{\widetilde{\alpha}\lambda_1}{2+2\Gamma^{-2}}\right)(f(x^k) - f(x^*)). \tag{5.10}$$

Comparing the inequalities (5.10) and (5.7), and taking into account that by definition $\Gamma \leq \widehat{\Gamma}$ and $\Gamma^{-1} \leq \widehat{\Gamma}$, we can see that (5.1) holds.

In order to prove the error bound (5.3), we use (5.1) and the fact that the solution x^* to (1.1) is characterized by the variational inequality $(x - x^*)^\top g(x^*) \geq 0$ for all $x \in \Omega$. We obtain

$$\begin{aligned} \|x^k - x^*\|_A^2 &= 2(f(x^k) - f(x^*) - (x^k - x^*)^\top g(x^*)) \\ &\leq 2(f(x^k) - f(x^*)) \ \leq \ 2\eta^k(f(x^0) - f(x^*)). \end{aligned}$$ □

Theorem 5.1 yields the best estimate for $\Gamma = \widehat{\Gamma} = 1$ and $\widetilde{\alpha} = \|A\|^{-1}$ when

$$\eta = 1 - \frac{1}{4}\kappa(A)^{-1}.$$

We shall see that this result is in agreement with numerical experiments.

6. Implementation. We shall give the details of the implementation of Algorithm 3.1. Firstly, we describe a general scheme of the algorithm independently on the type of constraints. Then we show how to compute projections on the feasible set in our numerical experiments.

6.1. A general algorithmic scheme. The presented implementation differs from Algorithm 3.1 in that it exploits the current conjugate gradient direction to generate an intermediate iteration $x^{k+1/2}$ before the expansion step that generates x^{k+1} from $x^{k+1/2}$. Such modification does not require any additional matrix-vector multiplication, and the estimate (5.1) remains valid [5].

We describe the algorithm in an easily understandable variant of the Matlab language, in which we do not distinguish a generation of variables by indices unless it is convenient for further references.

It should be noted that the presented implementation is not influenced by the type of constraints defining the feasible set Ω. The relation to constraints is hidden in the components $\widetilde{\beta}$, $\widetilde{\varphi}$ of the projected gradient $\widetilde{g}$, and in the feasible steplength α_f. The projected gradient can be directly evaluated by its definition (2.7) using P_Ω.

ALGORITHM 6.1. *Let $x^0 \in \Omega$, $\Gamma > 0$, $\widetilde{\alpha} \in (0, \|A\|^{-1}]$, and $\epsilon > 0$ be given.*

Set $k = 0$, $g = Ax^0 - b$, $p = \varphi(x^0)$. % *Initialization.*
while $\|\widetilde{g}(x^k)\| > \epsilon$
  if $\widetilde{\beta}(x^k)^\top g(x^k) \le \Gamma^2 \widetilde{\varphi}(x^k)^\top g(x^k)$
    $\alpha_{cg} = g^\top p / p^\top A p$ % *Conjugate gradient steplength.*
    $\alpha_f = \max\{\alpha : x^k - \alpha p \in \Omega\}$ % *Feasible steplength.*
    if $\alpha_{cg} < \alpha_f$ % *Conjugate gradient step.*
      $x^{k+1} = x^k - \alpha_{cg} p$, $g = g - \alpha_{cg} A p$, $\gamma = \varphi(x^{k+1})^\top A p / p^\top A p$, $p = \varphi(x^{k+1}) - \gamma p$
    else % *Expansion step.*
      $x^{k+1/2} = x^k - \alpha_f p$
      $x^{k+1} = x^{k+1/2} - \widetilde{\alpha}\widetilde{\varphi}(x^{k+1/2})$, $g = Ax^{k+1} - b$, $p = \varphi(x^{k+1})$
    end*if*
  else % *Proportioning step.*
    $x^{k+1} = x^k - \widetilde{\alpha}\widetilde{\beta}(x^k)$, $g = Ax^{k+1} - b$, $p = \varphi(x^{k+1})$
  end*if*
  $k = k + 1$
end*while*
$x = x^k$. % *Return step.*

6.2. Projections and the feasible steplength. In the previous sections, we have exploited the implicit definition of the projections given by (2.5) and (2.6). Now we shall show how to compute P_{Ω_i} for simple bounds (1.2) and circular constraints (1.3).

Let us consider the following variant of the problem (1.1):

$$\left\{\begin{array}{ll} \text{minimize} & f(x), \\ \text{subject to} & x_i \ge l_i,\ i = 1, \dots, m_1, \\ & x^2_{m_1+2i-1} + x^2_{m_1+2i} \le r_i^2,\ i = 1, \dots, m_2, \end{array}\right. \tag{6.1}$$

where $m_1 + 2m_2 < n$ and l_i, r_i are given. The feasible set Ω in (6.1) is described by Ω_i, $i = 1, \dots, m$, $m = m_1 + m_2 + 1$, of three different types. Recall that each of Ω_i is defined by $\Omega_i = \{\boldsymbol{x}_i \in \mathbb{R}^{n_i} : f_i(\boldsymbol{x}_i) \le 0\}$ where $f_i : \mathbb{R}^{n_i} \to \mathbb{R}$ is continuously differentiable and convex. Let us denote the feasible steplength with respect to the ith constraint by

$$\alpha_{f,i} = \max\{\alpha : \boldsymbol{x}_i - \alpha \boldsymbol{p}_i \in \Omega_i\}.$$

(1) Let $n_i = 1$, and $f_i(x_i) \equiv l_i - x_i$. Then $x_i \in \Omega_i$, $i = 1, \dots, m_1$ represent the simple bounds in (6.1). The set Ω_i describes the half-line so that the projection is given by

$$P_{\Omega_i}(x_i) = \left\{\begin{array}{ll} x_i & \text{for } l_i \le x_i, \\ l_i & \text{for } l_i > x_i \end{array}\right.$$

and the feasible steplength by

$$\alpha_{f,i} = \left\{\begin{array}{ll} (x_i - l_i)/p_i & \text{for } p_i > 0, \\ \infty & \text{for } p_i \le 0. \end{array}\right.$$

(2) Let $n_{m_1+i}=2$, $f_{m_1+i}(\boldsymbol{x}_{m_1+i}) \equiv \boldsymbol{x}_{m_1+i}^\top \boldsymbol{x}_{m_1+i} - r_i^2$, and $\boldsymbol{x}_{m_1+i} = (x_{m_1+2i-1}, x_{m_1+2i})^\top$. Then $\boldsymbol{x}_{m_1+i} \in \Omega_{m_1+i}$, $i=1,\dots,m_2$ represent the circular constraints in (6.1). The set Ω_{m_1+i} describes the circle with the center at the origin of $\mathbb{R}^2$ and with the radius r_i so that the projection is given by

$$P_{\Omega_{m_1+i}}(\boldsymbol{x}_{m_1+i}) = \begin{cases} \boldsymbol{x}_{m_1+i} & \text{for } \|\boldsymbol{x}_{m_1+i}\| \le r_i, \\ \dfrac{r_i}{\|\boldsymbol{x}_{m_1+i}\|}\,\boldsymbol{x}_{m_1+i} & \text{for } \|\boldsymbol{x}_{m_1+i}\| > r_i \end{cases}$$

and the feasible steplength by

$$\alpha_{f,m_1+i} = \begin{cases} (\boldsymbol{x}_i^\top \boldsymbol{p}_i + \sqrt{(\boldsymbol{x}_i^\top \boldsymbol{p}_i)^2 - (\|\boldsymbol{x}_i\|^2 - r_i^2)\|\boldsymbol{p}_i\|^2}\,)/\|\boldsymbol{p}_i\|^2 & \text{for } \boldsymbol{p}_i \neq 0, \\ \infty & \text{for } \boldsymbol{p}_i = 0. \end{cases}$$

(3) Let $n_m = n - m_1 - 2m_2$, $\boldsymbol{x}_m = (x_{m_1+2m_2+1}, \dots, x_n)^\top$, and $f_m(\boldsymbol{x}_m) = -1$. Then $\Omega_m = \mathbb{R}^{n_m}$, i.e., the components of $\boldsymbol{x}_m$ are unconstrained. For completeness, we define

$$P_{\Omega_m}(\boldsymbol{x}_m) = \boldsymbol{x}_m \quad \text{and} \quad \alpha_{f,m} = \infty.$$

Let us conclude that the projection P_Ω is put together by P_{Ω_i} as in (2.4) and the feasible steplength α_f is given by

$$\alpha_f = \min\{\alpha_{f,i} : i = 1, \dots, m\}.$$

7. Numerical tests. We shall assess the performance of the algorithm by three examples. The first one is the benchmark of [10], in which only circular constraints occur. The second example represents a one-dimensional obstacle problem comprising both simple bounds and circular constraints. A more realistic third example shows the solution of frictional 3D contact problems of linear elasticity by means of a sequence of problems (6.1).

Let us note that either we shall use Algorithm 6.1 with $x^0 = 0$, $\Gamma = 1$, $\widetilde{\alpha} = \|A\|^{-1}$, and $\epsilon = 10^{-5}\|b\|$, or we shall comment different choices.

Example 7.1. Let us consider the problem (6.1) for $n = 12$, $m_1 = 0$, and $m_2 = 6$ with the five-diagonal matrix A:

$$\begin{aligned} &A = (a_{ij}),\ a_{ii} = 4,\ a_{ii\pm1} = a_{ii\pm2} = -1, \\ &b = Ay, \\ &r = (2, 1, 0.5, 2, 10^{-3}, 154)^\top, \end{aligned}$$

and $y = (2, 1, 0.5, 0, 0, 11, 10^{-5}, -1, \sqrt{2}, -0.1, 4.1 \cdot 10^{-4}, 143)^\top$. The solution $x^* \in \mathbb{R}^{12}$ has three active constraints so that $\mathcal{A}(x^*) = \{2, 3, 5\}$. Here, we shall denote Algorithm 6.1 by QPC and the algorithm of [10] by QPQ. In Table 7.1 we compare QPQ and QPC by the number of matrix-vector multiplications for various Γ. The algorithm QPC has better performance in all cases. Moreover, the experiments are in agreement with the conclusions of Theorem 5.1.

In order to explain the progress, we mention the ideas of QPQ. This algorithm is constructed by the KKT conditions that are given here in Theorem 2.1. Introducing the *turned boundary gradient* $\beta = \beta(x)$ at $x \in \Omega$ as

$$\boldsymbol{\beta}_i = \boldsymbol{0} \quad \text{for } i \in \mathcal{F}(x), \quad \boldsymbol{\beta}_i = \boldsymbol{g}_i + \frac{\|\boldsymbol{g}_i\|}{\|\nabla f_i(\boldsymbol{x}_i)\|}\nabla f_i(\boldsymbol{x}_i) \quad \text{for } i \in \mathcal{A}(x),$$

TABLE 7.1
Comparisons of the algorithms QPQ and QPC.

Γ	100	10	5	1	0.5	0.4	0.2	0.1	0.05	0.01	0.001
QPQ	43	41	39	32	37	40	48	41	38	36	36
QPC	30	24	21	19	20	20	22	22	26	37	50

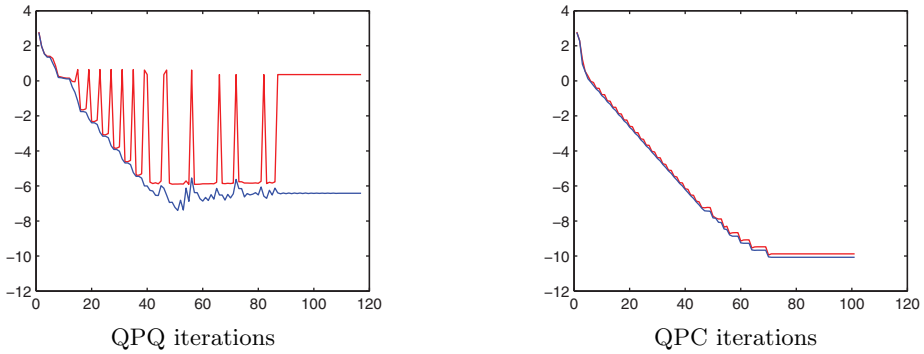


FIG. 7.1. *The iteration history of QPQ and QPC (logarithmic scale).*

we can define the *turned gradient* by $\nu = \nu(x) = \varphi(x) + \beta(x)$ where $\varphi(x)$ is the free gradient (2.9). The solution x^* to the problem (1.1) is then characterized by $\nu(x^*) = 0$. The proportioning step of QPC is replaced in QPQ by the step deactivation, in which β replaces $\tilde{\beta}$ (for a different steplength). It is easy to show that the deactivation step releases all indices from the current active set, for which the boundary KKT conditions (2.3) are not satisfied. After each release, the inner KKT conditions are violated usually on the same level. This observation is a practical consequence of the fact that the functions $\beta(x)$ and $\nu(x)$ are discontinuous (in contrast to $\tilde{\beta}$ and $\tilde{g}$). The typical situation is drawn in Figure 7.1 (left), where the norm of the projected gradient (solid) and the turned gradient (dotted) are depicted. In Figure 7.1 (right) we can see that oscillations arising in QPQ are eliminated in QPC. Comparing the stagnation levels, we can conclude that QPC is considerably more robust. Notice that the turned gradient did not curiously recognize the solution in QPQ due to round-off errors.

Example 7.2. In the second example, we solve

$$\text{minimize } \frac{1}{2}\int_0^1 \|\boldsymbol{x}'(t)\|^2\,dt - \int_0^1 \boldsymbol{x}(t)^\top \boldsymbol{f}(t)\,dt$$

subject to $\boldsymbol{x} = (X_1, X_2)^\top \in \mathcal{K}$, where

$$\mathcal{K} = \{\boldsymbol{x} \in (H_0^1(0,1))^2 :\ X_2(t) \geq l \text{ on } (0,0.5),\ \|\boldsymbol{x}(t)\| \leq r \text{ on } (0.5,1)\},$$

and $\boldsymbol{f}(t) = (36\pi^2 \sin 6\pi t, -4\pi^2 \sin 2\pi t)^\top$. This problem describes the loaded wire (see Figure 7.2) that is partially above the plan far off the distance l and partially inside the cylindrical tube of the radius r. A finite element discretization on a regular grid with n degrees of freedom leads to the problem (6.1) where $m_1 = m_2 = n/4$, $l_i = l$, and $r_i = r$. In tables below we summarize numbers of matrix-vector multiplications and information on active and free constraints as

$$n_{b,\mathcal{A}} : n_{b,\mathcal{F}} / n_{c,\mathcal{A}} : n_{c,\mathcal{F}},$$

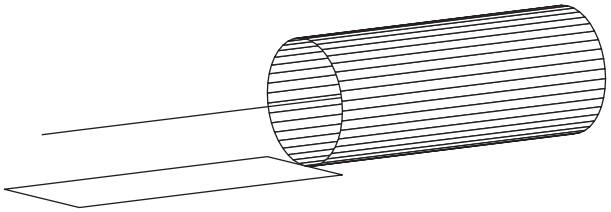


FIG. 7.2. *Geometry of the wire.*

TABLE 7.2
$r = 2$, only simple bounds are active in the solution.

n	$l = -1.5$	$l = -1$	$l = -0.8$	$l = -0.5$	$l = -0.1$	$l = 0$
	0:8/0:8	1:7/0:8	1:7/0:8	3:5/0:8	4:4/0:8	5:3/0:8
32	3	16	26	24	24	24
	0:16/0:16	1:15/0:16	1:15/0:16	4:12/0:16	6:10/0:16	9:7/0:16
64	3	25	57	57	72	62
	0:32/0:32	1:31/0:32	3:29/0:32	6:26/0:32	12:20//0:32	19:13/0:32
128	3	41	92	137	167	112
	0:64/0:64	1:63/0:64	4:60/0:64	10:54/0:64	23:41/0:64	37:27/0:64
256	3	68	262	307	369	269
	0:128/0:128	1:127/0:128	8:120/0:128	21:107/0:128	45:83/0:128	73:55/0:128
512	3	4	494	556	876	710
	0:256/0:256	1:255/0:256	16:240/0:256	40:216/0:256	89:167/0:256	146:110/0:256
1024	3	4	1305	1641	1966	1530

TABLE 7.3
$l = 0$, both simple bounds and circular constraints are active in the solution.

n	$r = 1.4$	$r = 1$	$r = 0.5$	$r = 0.3$	$r = 0.01$	$r = 0.001$
	5:3/2:6	6:2/4:4	7:1/5:3	7:1/5:3	8:0/8:0	8:0/8:0
32	89	80	54	42	17	19
	10:6/2:14	11:5/5:11	13:3/6:10	14:2/9:7	16:0/16:0	16:0/16:0
64	205	240	132	90	23	27
	20:12/4:28	22:10/5:27	26:6/10:22	29:3/16:16	32:0//31:1	32:0/32:0
128	608	620	324	239	96	41
	39:25/4:60	45:19/8:56	52:12/18:46	57:7/26:38	64:0/58:6	64:0/64:0
256	1677	1764	951	695	215	121
	77:51/4:124	89:39/12:116	104:24/33:95	114:14/49:79	127:1/111:17	128:0/126:2
512	4117	5248	3755	1960	646	309
	155:101/6:250	177:79/22:234	208:48/60:196	228:28/95:161	253:3/219:37	256:0/249:7
1024	12084	16851	10516	7065	2051	793

where $n_{b,\mathcal{A}}$, $n_{b,\mathcal{F}}$, $n_{c,\mathcal{A}}$, and $n_{c,\mathcal{F}}$ are numbers of active simple bounds, free simple bounds, active circular constraints, and free circular constraints, respectively.

(i) Let $r = 2$ so that no circular constraint is active in the solution; see Table 7.2. In this case, Algorithm 6.1 reduces to a variant of the algorithm of [5] with the finite terminating property. Let us note that the column labeled $l = -1.5$ corresponds to the unconstrained problem, in which our algorithm became the conjugate gradient method.

(ii) Let $l = 0$ so that both simple bounds and circular constraints may be active in the solution; see Table 7.3. It turns out that the algorithm is more efficient in situations when the constraints are tighter, i.e., when r and l are near zero so that the number of active constraints is considerably higher than the number of free constraints.

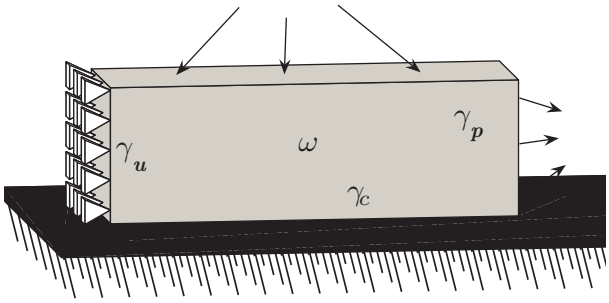


FIG. 7.3. *Geometry of the brick.*

Example 7.3. Let us consider the steel brick lying on a rigid foundation as it is shown in Figure 7.3. The brick occupies in the reference configuration the domain $\omega \subset \mathbb{R}^3$, whose boundary $\partial\omega$ is split into three nonempty disjoint parts γ_u, γ_p, and γ_c with different boundary conditions. The zero displacements are prescribed on γ_u, whereas the surface tractions act on γ_p. On γ_c, we consider the contact conditions, i.e., the nonpenetration and the effect of friction. The elastic behavior of the brick is described by Lamé equations that, after finite element discretization, lead to a symmetric positive definite stiffness matrix $K \in \mathbb{R}^{3n_c \times 3n_c}$ and to a load vector $f \in \mathbb{R}^{3n_c}$. Moreover, we introduce full rank matrices $N, T_1, T_2 \in \mathbb{R}^{m_c \times 3n_c}$ projecting displacements at contact nodes to normal and tangential directions, respectively, and we denote $B = \left(N^\top, T_1^\top, T_2^\top\right)^\top \in \mathbb{R}^{3m_c \times 3n_c}$. For more details about this model problem, we refer to [8]. Here, we shall use the dual formulation in terms of contact stresses.

We start with the contact problem with *Tresca friction* that reads as

$$\left\{\begin{array}{ll} \text{minimize} & \frac{1}{2}\lambda^\top Q\lambda - \lambda^\top h, \\ \text{subject to} & \lambda_{\nu,i} \geq 0,\ \lambda_{t_1,i}^2 + \lambda_{t_2,i}^2 \leq r_i^2,\ i = 1, \ldots, m_c, \\ & \lambda = (\lambda_\nu^\top, \lambda_{t_1}^\top, \lambda_{t_2}^\top)^\top,\ \lambda_\nu, \lambda_{t_1}, \lambda_{t_2} \in \mathbb{R}^{m_c}, \end{array}\right. \tag{7.1}$$

where $Q = BK^{-1}B^\top$, $h = BK^{-1}f$, and $r_i \geq 0$ are given slip bound values at contact nodes. Let us point out that λ_ν and $\lambda_{t_1}, \lambda_{t_2}$ represent normal and tangential contact stresses, respectively. It should be noted that the problem (7.1) can be solved by Algorithm 6.1 after rearranging the unknowns. In order to simplify notations, we denote $\mathbb{R}^{m_c}_+ = \{s \in \mathbb{R}^{m_c} : s_i \geq 0\}$.

The contact problem with *Coulomb friction* uses the friction law, in which the slip bound $r \in \mathbb{R}^{m_c}_+$ depends on the normal contact stress $\lambda_\nu \in \mathbb{R}^{m_c}_+$ by

$$r \equiv F\lambda_\nu,$$

where $F > 0$ is a coefficient of friction. As r is the input for (7.1) and λ_ν is the output, the problem with Tresca friction defines the mapping

$$\Psi : \mathbb{R}^{m_c}_+ \longmapsto \mathbb{R}^{m_c}_+ : r \longmapsto F\lambda_\nu.$$

It is easily seen that a fixed point of Ψ solves the problem with Coulomb friction, i.e., a point r such that $\Psi(r) = r$. Notice that Ψ is contractive for sufficiently small F,

TABLE 7.4
Contact problem with Coulomb friction.

dof		$F = 0.3$				$F = 0.6$			
$3n_c$	$3m_c$	*Time*	*Iter*	n_Q	n_Q/n	*Time*	*Iter*	n_Q	n_Q/n
900	180	4	5	**535**	2.97	6	7	**801**	4.45
2646	378	24	5	**638**	1.68	35	6	**906**	2.40
5832	648	104	5	**758**	1.17	136	6	**1001**	1.54
10890	990	317	5	**814**	0.82	443	6	**1145**	1.16
18252	1404	789	5	**854**	0.61	1122	6	**1232**	0.88
28350	1890	1833	5	**947**	0.50	2222	6	**1169**	0.62

and then there is a unique fixed point. Moreover, successive approximations can be used for its computation:

$$r^0 \in \mathbb{R}^{m_c}_+ \text{ given;} \quad \text{for } k = 1, 2, \dots \text{ set } r^k = \Psi(r^{k-1}).$$

As the evaluation of Ψ requires us to solve (7.1), we can repeatedly apply Algorithm 6.1. In order to perform these computations efficiently, we initialize each iteration by results from the previous one. Finally we terminate if

$$\|r^k - r^{k-1}\|/\|r^k\| \le 10^{-4}.$$

In our numerical experiments, we consider the steel brick $\omega = (0,3)\times(0,1)\times(0,1)$ partitioned into $3N \times N \times N$ cubes by trilinear finite elements for $N = 4, 6, 8, 10, 12,$ and 14. The size of problems solved by Algorithm 6.1 is $n = 3m_c = 9N(N+1)$. In Table 7.4, we report CPU time in seconds (*time*), the number of successive approximations (*iter*), the total complexity by matrix-vector multiplications (n_Q), and the relative complexity (n_Q/n). The computations are carried out in Matlab 7 on Pentium(R)4, 3GHz, 512MB. The obtained results are promising; especially, n_Q is only mildly dependent on the finite element discretization so that the relative complexity considerably decreases for finer grids.

8. Comments and conclusions. We have analyzed a new active set algorithm for minimizing strictly convex quadratic functions with separable convex constraints. It generalizes a recently developed algorithm of quadratic programming constrained by simple bounds [5]. It should be noted that we did not need any requirement on nondegeneracy of the problem so that our algorithm is globally convergent for both the nondegenerate as well as the degenerate case.

The main goal is the proof of a linear convergence rate, which is in an optimal case described by the factor $\eta = 1 - \frac{1}{4}\kappa(A)^{-1}$ in terms of the condition number of the Hessian matrix A. Notice that η is not influenced by constraints. The key assumption is the restricted steplength $\widetilde{\alpha}$ defining the projected gradient $\widetilde{g}$, i.e., $\widetilde{\alpha} \le \|A\|^{-1}$, which means that the components $\widetilde{\varphi}$ and $\widetilde{\beta}$ of $\widetilde{g}$ are descent directions. We use them for adding/releasing indices to/from the active set in the step expansion/proportioning. Unfortunately, our analysis requires us to replace the conjugate gradient steplengths by $\widetilde{\alpha}$, which seems to be too restrictive, and the obtained convergence rate may be a bit pessimistic. That is the case for simple bound problems. If the constraints are conic (e.g., circular), the algorithm usually performs few valuable conjugate gradient steps combined with expansion steps, and then it alternates many proportioning steps with short conjugate gradient steps. The convergence rate is more realistic in such situations.

The algorithm presented here is an important ingredient in the numerical solution of 3D contact problems. It was shown in [3] that FETI domain decomposition methods are scalable for the frictionless contact. Our paper enables us to extend this result for frictional problems. It will be published in forthcoming papers.

Another class of algorithms relevant for our research (contact problems) is based on a specific semismooth Newton method that is identical again with an active set strategy [9]. The fundamental discrepancy is the fact that it allows infeasible iterates. In this case, the convergence rate is superlinear but requires a sufficiently accurate initial approximation. In general, the computational performance is comparable.

REFERENCES

[1] A. Conn, N. Gould, and P. Toint, *Testing a class of algorithms for solving minimization problems with simple bounds on the variables*, Math. Comp., 50 (1988), pp. 399–430.

[2] Z. Dostál, *Box constrained quadratic programming with proportioning and projections*, SIAM J. Optim., 7 (1997), pp. 871–887.

[3] Z. Dostál, *An optimal algorithm for a class of equality constrained quadratic programming problems with bounded spectrum*, Comput. Optim. Appl., 38 (2007), pp. 47–59.

[4] Z. Dostál, J. Haslinger, and R. Kučera, *Implementation of the fixed point method in contact problems with Coulomb friction based on a dual splitting type technique*, J. Comput. Appl. Math., 140 (2002), pp. 245–256.

[5] Z. Dostál and J. Schöberl, *Minimizing quadratic functions over non-negative cone with the rate of convergence and finite termination*, Comput. Optim. Appl., 30 (2005), pp. 23–44.

[6] A. Friedlander and M. Martínez, *On the maximization of a concave quadratic function with box constraints*, SIAM J. Optim., 4 (1994), pp. 117–192.

[7] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.

[8] J. Haslinger, R. Kučera, and Z. Dostál, *An algorithm for the numerical realization of 3D contact problems with Coulomb friction*, J. Comput. Appl. Math., 164-165 (2004), pp. 387–408.

[9] S. Hüeber, G. Stadler, and B. I. Wohlmuth, *A primal-dual active set algorithm for three-dimensional contact problems with Coulomb friction*, SIAM J. Sci. Comput., 30 (2008), pp. 572–596.

[10] R. Kučera, *Minimizing quadratic functions with separable quadratic constraints*, Optim. Methods Softw., 22 (2007), pp. 453–467.

[11] R. Kučera, J. Haslinger, and Z. Dostál, *A new FETI-based algorithm for solving 3D contact problems with Coulomb friction*, Lect. Notes Comput. Sci. Eng. 55, Springer, Berlin, 2007, pp. 645–652.

[12] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, New York, 1999.

[13] B. T. Polyak, *The conjugate gradient method in extremal problems*, USSR Comput. Math. and Math. Phys., 9 (1969), pp. 94–112.

SIAM J. OPTIM.
Vol. 19, No. 2, pp. 863–882

SUBDIFFERENTIAL CALCULUS RULES IN CONVEX ANALYSIS: A UNIFYING APPROACH VIA POINTWISE SUPREMUM FUNCTIONS*

A. HANTOUTE†, M. A. LÓPEZ‡, AND C. ZĂLINESCU§

Abstract. We provide a rule to calculate the subdifferential set of the pointwise supremum of an arbitrary family of convex functions defined on a real locally convex topological vector space. Our formula is given exclusively in terms of the data functions and does not require any assumption either on the index set on which the supremum is taken or on the involved functions. Some other calculus rules, namely chain rule formulas of standard type, are obtained from our main result via new and direct proofs.

Key words. convex analysis, convex subdifferential, pointwise supremum function, calculus rules

AMS subject classifications. 52A41, 90C25, 15A39

DOI. 10.1137/070700413

1. Introduction.

Many operations with convex functions preserve convexity, and so it is natural to ask if the subdifferential of the resulting function can be written in terms of the data functions. Specific to convex analysis is the classical operation of taking the pointwise supremum of an arbitrarily indexed family of convex functions. It has no equivalence in the classical theory of differentiable analysis and constitutes a largely used tool in convex optimization, in theory as well as in practice (see, for instance, [1], [10], and the references therein). In [5] and [8] certain specific techniques relying on the supremum function were applied in the framework of semi-infinite linear optimization.

In this paper, we provide explicit characterizations for the subdifferential mapping of the supremum function of an arbitrarily indexed family of convex functions, exclusively in terms of the data functions. The main virtue of our approach is that the index set over which the supremum is taken is arbitrary, without any algebraic or topological structure. Also the convex functions we consider in this paper are general, defined on a separated locally convex space, and not necessarily lower semicontinuous (lsc) or even proper. Further, we do not assume regularity conditions such as the continuity of the supremum function, the continuity of the data functions, conditions on their domains, and the like.

Since many convex functions can be written as the supremum of continuous affine mappings, numerous operations dealing with such (convex) functions can be formulated as a pointwise supremum of other functions whose subdifferentials can easily be characterized. Specifically, we have proved that our formulas also lead to other

*Received by the editors August 17, 2007; accepted for publication (in revised form) March 26, 2008; published electronically August 13, 2008. Research supported by grants MTM2005-08572-C03 (01) from MEC (Spain) and FEDER (E.U.), ACOMP06/117 and ACOMP/2007/247-292 from Generalitat Valenciana (Spain), and ID-PCE-379 (Romania).

http://www.siam.org/journals/siopt/19-2/70041.html

†Operations Research Center, Miguel Hernández University of Elche, 03202 Elche (Alicante), Spain (hantoute@ua.es).

‡Department of Statistics and Operations Research, University of Alicante, 03071 Alicante, Spain (marco.antonio@ua.es).

§University "Al. I. Cuza" Iaşi, Faculty of Mathematics, Bd. Carol I, Nr. 11, 700506 Iaşi, Romania and Institute of Mathematics Octav Mayer, Iaşi, Romania (zalinesc@uaic.ro).

calculus rules for the subdifferentials of certain operations with convex functions, such as the sum and the composition with affine applications. In this way, our approach gives rise to a unifying view of many well-known calculus rules in convex analysis.

Deriving calculus rules for subdifferentials is one of the first issues raised in convex analysis. Consequently, many earlier contributions dealing with pointwise supremum functions can be found in the literature. See, for instance, [26] to trace out the historical origins of the issue, as well as [2], [3], [4], [12], [13], [15], [20], [21], and [27]. This is why we make a short historical review of some of these results.

Consider the pointwise supremum $f := \sup_{t\in T} f_t$ of a collection of convex functions $f_t : X \to \mathbb{R} \cup \{+\infty\}$, $t \in T \neq \emptyset$, defined on a separated locally convex space X, and let $z \in \operatorname{dom} f$. When T is finite and each f_t is continuous at z, a basic result due to Dubovitskij and Milyutin asserts that (see, e.g., [13])

$$\partial f(z) = \operatorname{co}\left(\bigcup_{t\in T(z)} \partial f_t(z)\right),$$

where

$$T(z) := \{t \in T \mid f_t(z) = f(z)\},$$

and co stands for the convex hull. When T is a separated compact topological space and the function $(t, x) \to f_t(x)$ is upper semicontinuous with respect to t for each x, then assuming that each f_t is continuous at z, the following formula can be found, for instance, in [32, Thm. 2.4.18]:

$$\partial f(z) = \operatorname{cl}\left(\operatorname{co}\bigcup_{t\in T(z)} \partial f_t(z)\right),$$

where the closure, cl, is taken in the topological dual space X^* with respect to the weak* topology $w^* = \sigma(X^*, X)$.

According to [26], the last result was first established by Levin [15] for a finite-valued convex function defined on $\mathbb{R}^n$. The continuity assumption on the data functions is weakened in [29] and [21, Thm. 4].

Even in simple situations dealing with finitely many functions, the problem is involved so that simple examples in the Euclidean space show that these nice formulae above do not hold in general. Nevertheless, in order to overcome this difficulty, Brøndsted [2] used the concept of ε-subdifferential to establish the following formula, which is valid when $T = \{1, 2, \ldots, k\}$ and all of the functions $f_i,\ i = 1, 2, \ldots, k$ agree at z :

$$\partial f(z) = \bigcap_{\varepsilon>0} \operatorname{cl}\left(\operatorname{co}\bigcup_{i=1}^{k} \partial_\varepsilon f_i(z)\right).$$

In the case of an infinite collection of convex functions (T infinite), and following [10, p. 405], the most elaborated results are due to Valadier in [27] where, in the context of normed vector spaces and assuming that the supremum function f is continuous at z, the subdifferential $\partial f(z)$ is expressed by considering not only z but all nearby points around it. More precisely, denoting by $\|\cdot\|$ the corresponding norm in X, the following formula is given in [27]:

$$\partial f(z) = \bigcap_{\varepsilon>0} \operatorname{cl}\left[\operatorname{co}\left(\bigcup\{\partial f_t(y) \mid y \in X,\ t \in T : \|y - z\| \leq \varepsilon,\ f_t(z) \geq f(z) - \varepsilon\}\right)\right].$$

By using the concept of ε-subdifferential, Volle [28] obtained another characterization of $\partial f(z)$ where only the nominal point z appears but in terms of approximate subgradients:

$$\partial f(z) = \bigcap_{\varepsilon>0} \operatorname{cl}\left[\operatorname{co}\left(\bigcup\{\partial_\varepsilon f_t(z) \mid t \in T : f_t(z) \geq f(z) - \varepsilon\}\right)\right].$$

It is worth noting that if either all of the functions f_t are affine or if the space X is Banach, then the last two formulas above are equivalent. The equivalence for affine functions is clear while in the Banach spaces setting this observation is partly due to Brøndsted–Rockafellar's theorem, expressing the ε-subdifferential by means of exact subdifferentials at nearby points. As it can be seen, the advantage of using such an enlargement of the subdifferential, namely, the ε-subdifferential, is to avoid qualifications type conditions. Such an idea is exploited in the survey paper [11] (see also references therein) to provide many calculus rules without requiring any regularity condition.

Recently, in [7], the following characterization for the subdifferential ∂f is given when $f_t : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, $t \in T$, are proper convex functions and T is arbitrary:

$$\partial f(z) = \bigcap_{\varepsilon>0} \operatorname{cl}\big[\operatorname{co}\big(\bigcup\{\partial_\varepsilon f_t(z) \mid t \in T : f_t(z) \geq f(z) - \varepsilon\}\big) + \mathrm{N}_{\operatorname{dom} f}(z)\big],$$

where $\mathrm{N}_{\operatorname{dom} f}(z)$ stands for the normal cone to the domain of f ($\operatorname{dom} f$) at z, provided that either the f_t's are lsc or that the relative interiors of their (effective) domains have a common point. In this setting, the formula above implies the one given by Volle [28], since $\mathrm{N}_{\operatorname{dom} f}(z) = \{\theta\}$ whenever z is a continuity point of the supremum function f. Further, when dealing with a finite number of functions the term $\mathrm{N}_{\operatorname{dom} f}(x)$ can be removed from the formula above which, consequently, entails the one of Brøndsted [2].

At this step, the purpose of the present paper is twofold. First, we extend the last formula from [7] to the setting of convex functions defined on locally convex spaces and which are not necessarily proper or lsc. To this aim, we consider those collections of functions satisfying the following closedness criterion, which holds for a broad class of convex functions and obviously covers the case of lsc functions:

$$\operatorname{cl} f = \sup_{t\in T} \operatorname{cl} f_t, \tag{1}$$

where $\operatorname{cl} f$ and $\operatorname{cl} f_t$ stand for the lsc hull of the convex functions f and f_t, respectively. Second, we give a unified approach for the framework of calculus rules in convex analysis. In fact, our characterization of ∂f also allows us to obtain formulas for the subdifferential of the resulting function in many operations as the sum of convex functions and the composition of an affine continuous mapping with a convex function. In this way, we provide direct and easier proofs for the basic chain rules when some supplementary qualification conditions are assumed.

The summary of the paper is as follows. In section 2 we introduce the main tools and basic results used in the paper. In section 3 we give the aimed formula for the subdifferential of the supremum of an arbitrary family of convex functions. After a series of auxiliary lemmas the main result is stated in Theorem 4. In it we use a closedness criterion which is studied in Corollary 9. We close this section by deriving some other formulae in Corollaries 7 (for affine functions), 8 (for finite-dimensional spaces or, more generally, when the relative interior of the domain of the supremum function f is not empty), 10 (Volle's formula), and 12 (Brøndsted's formula). In section 4 we introduce a unifying framework for deriving subdifferential calculus rules. Namely, in Theorem 13 we give a formula for the subdifferential of the sum of a convex function and another convex function precomposed with a continuous affine mapping. Theorem 13 constitutes a slight extension of Hiriart-Urruty–Phelps formula (Corollary 14). It also yields an easy derivation of the basic chain rule (Corollary 16) when some supplementary conditions are assumed, namely, the Moreau–Rockafellar constraint qualification.

2. Notations and basic tools. In this paper X and Y stand for (real) separated locally convex spaces. Their topological dual spaces are respectively denoted by X^* and Y^*. The spaces X and X^* (Y and Y^*) are paired in duality by the bilinear form $(x^*,x)\in X^*\times X\mapsto \langle x^*,x\rangle := \langle x,x^*\rangle := x^*(x)$ ($(y^*,y)\in Y^*\times Y\mapsto \langle y^*,y\rangle$, respectively). Throughout the paper, the sole topology defined on the dual spaces is the w^*–topology. The zero vectors in the involved spaces are all denoted by θ, and the neighborhoods of θ are called θ–neighborhoods. We use the notation $\overline{\mathbb{R}} := \mathbb{R}\cup\{-\infty,+\infty\}$.

We first recall some basic results of convex analysis which can be found, e.g., in the books [17] and [32] and the references therein (see also [10] and [22]). Given two nonempty sets A and B in X (or in X^*, Y, Y^*), we define the algebraic (or Minkowski) sum by

$$A+B := \{a+b \mid a\in A,\ b\in B\},\quad A+\emptyset := \emptyset + A := \emptyset; \tag{2}$$

moreover, if $\emptyset\neq\Lambda\subset\mathbb{R}$ we set

$$\Lambda A := \{\lambda a \mid \lambda\in\Lambda,\ a\in A\},\quad \Lambda\emptyset := \emptyset.$$

Furthermore, $\Lambda x := \Lambda\{x\}$, $\lambda A := \{\lambda\}A$, and $x+A := \{x\}+A$.

By $\operatorname{co} A$, $\operatorname{cone} A$, and $\operatorname{aff} A$, we denote the *convex hull,* the *conic hull,* and the *affine hull* of the set A, respectively. Moreover, $\operatorname{int} A$ is the *interior* of A, and $\operatorname{cl} A$ and $\overline{A}$ are indistinctly used for denoting the *closure* of A (w^*–*closure* if $A\subset X^*$ or $A\subset Y^*$). In this way, we set $\overline{\operatorname{co}}A := \operatorname{cl}(\operatorname{co} A)$ and $\overline{\operatorname{cone}}A := \operatorname{cl}(\operatorname{cone} A)$. We use $\operatorname{ri} A$ to denote the (topological) *relative interior* of A (i.e., the interior of A in the topology relative to $\operatorname{aff} A$ if $\operatorname{aff} A$ is closed, and the empty set otherwise). We shall use Greek letters for denoting real numbers.

The following properties are applied very often:

$$\operatorname{cl}(A+B) = \operatorname{cl}(A+\operatorname{cl} B), \tag{3}$$

and if A is convex,

$$\lambda\operatorname{ri} A + (1-\lambda)\operatorname{cl} A\subset \operatorname{ri} A \text{ for every } \lambda\in\ ]0,1]. \tag{4}$$

Associated with $A\neq\emptyset$ we consider the sets

$$\begin{aligned} A^\circ &:= \{x^*\in X^* \mid \langle x^*,x\rangle\geq -1 \text{ for all } x\in A\},\\ A^- &:= -(\operatorname{cone} A)^\circ = \{x^*\in X^* \mid \langle x^*,x\rangle\leq 0 \text{ for all } x\in A\}, \text{ and}\\ A^\perp &:= (-A^-)\cap A^- = \{x^*\in X^* \mid \langle x^*,x\rangle = 0 \text{ for all } x\in A\}, \end{aligned}$$

i.e., the (one-sided) *polar*, the *negative dual cone,* and the *orthogonal subspace* (or *annihilator*) of A, respectively. Observe that A° is a closed convex set containing θ, A^- is a closed convex cone, and $A^\perp$ is a closed linear subspace. Further, by the *bipolar theorem,* we have

$$A^{\circ\circ} = \overline{\operatorname{co}}(A\cup\{\theta\}) \text{ and } A^{--} = \overline{\operatorname{cone}}(\operatorname{co} A). \tag{5}$$

If $A\subset X$ is convex and $x\in X$, we define the *normal cone* to A at x as

$$\mathrm{N}_A(x) := \begin{cases} (A-x)^- & \text{if } x\in A,\\ \emptyset & \text{if } x\in X\setminus A. \end{cases}$$

As a consequence of this definition $N_\emptyset(x) = \emptyset$ for every $x \in X$. If $A \neq \emptyset$ is convex and closed, A_∞ represents its *recession cone* defined as

$$A_\infty := \{y \in X \mid x + \lambda y \in X \text{ for some } x \in X \text{ and all } \lambda \geq 0\}.$$

Given a function $f : X \longrightarrow \overline{\mathbb{R}}$, its *(effective) domain* and *epigraph* are defined by

$$\begin{aligned} \operatorname{dom} f &:= \{x \in X \mid f(x) < +\infty\}, \\ \operatorname{epi} f &:= \{(x, \alpha) \in X \times \mathbb{R} \mid f(x) \leq \alpha\}; \end{aligned}$$

moreover, when f is *proper*, that is, $\operatorname{dom} f \neq \emptyset$ and $f(x) > -\infty$ for all $x \in X$, we consider the *graph* of f as being defined by

$$\operatorname{gph} f := \{(x, f(x)) \in X \times \mathbb{R} \mid x \in \operatorname{dom} f\}.$$

So, for f proper one has $\operatorname{epi} f = \operatorname{gph} f + \mathbb{R}_+(\theta, 1)$. We say that f is *convex* if $\operatorname{epi} f$ is convex. In what follows we shall use the convention $+\infty - \infty := +\infty + (-\infty) := +\infty$. Assume that f is convex. The *lower closure* of f is the function $\operatorname{cl} f : X \longrightarrow \overline{\mathbb{R}}$ defined by

$$(\operatorname{cl} f)(x) := \inf\{t \mid (x, t) \in \operatorname{cl}(\operatorname{epi} f)\}.$$

Clearly we have $\operatorname{epi}(\operatorname{cl} f) = \operatorname{cl}(\operatorname{epi} f)$, which implies that $\operatorname{cl} f$ is a lsc convex function dominated by f; i.e., $\operatorname{cl} f \leq f$. Equivalently, we have

$$(\operatorname{cl} f)(x) = \liminf_{y \to x} f(y) \quad \forall x \in X.$$

Further, it can be checked that $\operatorname{cl}(\operatorname{dom}(\operatorname{cl} f)) = \operatorname{cl}(\operatorname{dom} f)$. If $(\operatorname{cl} f)(x) = f(x)$, then f is lsc at x. If there exists $x_0 \in X$ such that $(\operatorname{cl} f)(x_0) = -\infty$, then $(\operatorname{cl} f)(x) = -\infty$ for all $x \in \operatorname{dom}(\operatorname{cl} f)$. We shall denote by $\Lambda(X)$ the set of all the proper convex functions on X, and by $\Gamma(X)$ the subset of $\Lambda(X)$ consisting of the lsc functions; the sets $\Lambda(X^*)$ and $\Gamma(X^*)$ are defined in a similar way.

The *Fenchel conjugate* of f is the function $f^* : X^* \longrightarrow \overline{\mathbb{R}}$ given by

$$f^*(x^*) := \sup\{\langle x^*, x\rangle - f(x) \mid x \in X\}.$$

The functions f and $\operatorname{cl} f$ have the same conjugate; i.e., $f^* = (\operatorname{cl} f)^*$. The *biconjugate* of f is the function $f^{**} : X \longrightarrow \overline{\mathbb{R}}$ given by

$$f^{**}(x) := \sup\{\langle x^*, x\rangle - f^*(x^*) \mid x^* \in X^*\}.$$

Let us recall here that $f^* \in \Gamma(X^*)$ if and only if $\operatorname{dom} f \neq \emptyset$ and there exist $x^* \in X^*$ and $\alpha \in \mathbb{R}$ such that $f(x) \geq \langle x^*, x\rangle + \alpha$ for all $x \in X$; this happens, for instance, when $f \in \Gamma(X)$ in which case we have $f^{**} = f$.

The *support* and the *indicator* functions of $A \neq \emptyset$ are, respectively, defined as

$$\sigma_A(x^*) := \sup\{\langle x^*, a\rangle \mid a \in A\} \text{ for } x^* \in X^*,$$

and

$$\mathrm{I}_A(x) := \begin{cases} 0 & \text{if } x \in A, \\ +\infty & \text{if } x \in X \setminus A. \end{cases}$$

The function σ_A is sublinear, lsc, and satisfies

$$\sigma_A = \sigma_{\overline{\text{co}}A} = \text{I}^*_{\overline{\text{co}}A}. \tag{6}$$

Moreover, it is known that $(\text{dom}\,\sigma_A)^- = (\overline{\text{co}}A)_\infty$ (e.g., [29, p. 142]) or equivalently, by using (5),

$$\text{cl}(\text{dom}\,\sigma_A) = [(\overline{\text{co}}A)_\infty]^- . \tag{7}$$

If $A_1, \ldots, A_m \subset X$ are nonempty sets ($m \geq 2$), then clearly $\sigma_{A_1} + \cdots + \sigma_{A_m} = \sigma_{A_1+\cdots+A_m}$ and $\max\{\sigma_{A_1}, \ldots, \sigma_{A_m}\} = \sigma_{A_1\cup\cdots\cup A_m}$; moreover, if $1 \leq k < m$, then

$$\sigma_{A_1} + \cdots + \sigma_{A_k} + \max\left\{\sigma_{A_{k+1}}, \ldots, \sigma_{A_m}\right\} = \sigma_{A_1+\cdots+A_k+(A_{k+1}\cup\cdots\cup A_m)}.$$

Hence

$$\text{dom}\,\sigma_{A_1+\cdots+A_m} = \text{dom}\,\sigma_{A_1\cup\cdots\cup A_m} = \text{dom}\,\sigma_{A_1+\cdots+A_k+(A_{k+1}\cup\cdots\cup A_m)}.$$

Using (6) and (7) we get

$$\begin{aligned} [\overline{\text{co}}(A_1 + \cdots + A_m)]_\infty &= [\overline{\text{co}}(A_1 \cup \cdots \cup A_m)]_\infty \\ &= [\overline{\text{co}}\,(A_1 + \cdots + A_k + (A_{k+1} \cup \cdots \cup A_m))]_\infty . \end{aligned} \tag{8}$$

If f is convex and $\varepsilon \geq 0$, the *ε-subdifferential* of f at a point $x \in X$ such that $f(x) \in \mathbb{R}$ is the w^*–closed convex set

$$\partial_\varepsilon f(x) := \{x^* \in X^* \mid f(y) - f(x) \geq \langle x^*, y - x\rangle - \varepsilon \text{ for all } y \in X\}.$$

If $f(x) \notin \mathbb{R}$, then we set $\partial_\varepsilon f(x) := \emptyset$. In particular, for $\varepsilon = 0$ we get $\partial f(x) := \partial_0 f(x)$, the *subdifferential* of f at x. Given $x \in X$ and $\varepsilon \geq 0$ we recall the following properties: $\partial f(x) = \cap_{\varepsilon>0}\partial_\varepsilon f(x)$ and $\partial_\varepsilon f(x) = \partial_\varepsilon f(x) + \text{N}_{\text{dom}\,f}(x)$; moreover, as a simple computation shows (see also [32, Exer. 2.23]),

$$[\partial_\varepsilon f(x)]_\infty = \text{N}_{\text{dom}\,f}(x) \text{ for all } x \in \text{dom}\,f \text{ and } \varepsilon \geq 0 \text{ with } \partial_\varepsilon f(x) \neq \emptyset. \tag{9}$$

If f is not proper, then $\partial_\varepsilon f(x) = \emptyset$ for all $x \in X$. If f is lsc at x and $f(x) \in \mathbb{R}$, then

$$\partial_\varepsilon\,(\text{cl}\,f)\,(x) = \partial_\varepsilon f(x). \tag{10}$$

If $\partial f(x) \neq \emptyset$, then we have

$$(\text{cl}\,f)(x) = f(x) \text{ and } \partial\,(\text{cl}\,f)\,(x) = \partial f(x). \tag{11}$$

If $f \in \Lambda(X)$ and $f(x) \in \mathbb{R}$, then we have $\partial_\varepsilon f(x) \neq \emptyset$ for all $\varepsilon > 0$ if and only if f is lsc at x. Moreover, we have

$$\partial_\varepsilon f(x) = \{x^* \in X^* \mid f(x) + f^*(x^*) \leq \langle x^*, x\rangle + \varepsilon\} \text{ for all } \varepsilon \geq 0. \tag{12}$$

If A is convex and $x \in A$,

$$\partial \text{I}_A(x) = (\text{cone}(A - x))^- = \text{N}_A(x).$$

Finally, if $f \in \Gamma(X)$, then for every $x \in \text{dom}\,f$, $u \in X$ and $\varepsilon > 0$, we have (see [32, Thm. 2.4.11])

$$f'_\varepsilon(x, u) := \inf_{\lambda>0} \frac{f(x + \lambda u) - f(x) + \varepsilon}{\lambda} = \sigma_{\partial_\varepsilon f(x)}(u). \tag{13}$$

3. Calculus rules for the subdifferential of the supremum function. In this section we consider a nonempty family $\{f_t \mid t \in T\}$ of convex functions $f_t : X \to \overline{\mathbb{R}}$ defined on a (separated) real locally convex space X. The corresponding *pointwise supremum function* $f : X \to \overline{\mathbb{R}}$, given by

$$f(x) := \sup\{f_t(x) \mid t \in T\}, \tag{14}$$

is also convex; our main purpose in this section is to provide a formula for the subdifferential ∂f of f in terms exclusively of the data functions f_t, $t \in T$. The following simple example draws aside, in general, the possibility of writing ∂f in terms of ∂f_t, $t \in T$.

Example 1. [11, Ex. 2.1] Let $f_1, f_2 : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be defined by

$$f_1(x) = \begin{cases} -2\sqrt{x} & \text{if } x \geq 0, \\ +\infty & \text{if } x < 0, \end{cases} \quad \text{and } f_2(x) = f_1(-x)$$

so that $f := \max\{f_1, f_2\} = \mathrm{I}_{\{0\}}$. Then, we easily check that $\partial f(0) = \mathbb{R}$ while both $\partial f_1(0)$ and $\partial f_2(0)$ are empty.

Nevertheless, Theorem 4 below provides a characterization of ∂f, which involves the approximate subdifferentials of the data functions. To start with, we first establish two elementary lemmas.

LEMMA 1. *Let $h \in \Lambda(X)$ and $A \subset \operatorname{dom} h$ be a convex set. If $\operatorname{ri} A \neq \emptyset$, then $\inf_A h = \inf_{\operatorname{cl} A} h$.*

Proof. Set $\mu := \inf_A h$. Fix some $x_0 \in \operatorname{ri} A$ and consider $x \in \operatorname{cl} A$. Take $x_n := (1 - \frac{1}{n})x + \frac{1}{n}x_0$ for $n \geq 1$; then

$$\mu \leq h(x_n) \leq (1 - \tfrac{1}{n})h(x) + \tfrac{1}{n}h(x_0).$$

Taking the limit we get $\mu \leq h(x)$; hence $\mu \leq \inf_{\operatorname{cl} A} h$. □

The following simple result is an immediate consequence of (10) and (11).

LEMMA 2. *Let $h \in \Lambda(X)$ and $x \in \operatorname{dom} h$. If $\operatorname{cl} h \in \Lambda(X)$, then*

$$\partial_\varepsilon h(x) = \partial_{\operatorname{cl} h(x) - h(x) + \varepsilon} \operatorname{cl} h(x) \quad \textit{for all} \quad \varepsilon \in \mathbb{R}.$$

Hence $\partial_\varepsilon h(x) \neq \emptyset$ for $\varepsilon > h(x) - \operatorname{cl} h(x)$, and $\partial_\varepsilon h(x) = \emptyset$ for $\varepsilon < h(x) - \operatorname{cl} h(x)$.

From now on, we fix the following notations. Given $z \in X$ and $\varepsilon > 0$ we set

$$\mathcal{F}_z := \{L \subset X \mid L \text{ is a finite-dimensional linear subspace, with } z \in L\},$$

and

$$T_\varepsilon(z) := \{t \in T \mid f_t(z) \geq f(z) - \varepsilon\},$$

where f_t and f are defined as in (14).

The following lemma provides the first extension of Proposition 3 in [7] to general locally convex spaces; [7, Prop. 3] is established in $\mathbb{R}^n$ using subdifferential calculus for support functions. Here we give a direct proof, which, in particular, does not appeal to the Fenchel linearization of the functions f_t.

LEMMA 3. *Let $f_t \in \Gamma(X)$ for $t \in T \neq \emptyset$ and set $f := \sup_{t \in T} f_t$. Assume that $z \in \operatorname{dom} f$ and that $\operatorname{ri}(\operatorname{dom} f) \neq \emptyset$, then*

$$\partial f(z) = \bigcap_{\varepsilon > 0} \operatorname{cl}\left(\operatorname{co}\left(\bigcup_{t \in T_\varepsilon(z)} \partial_{\alpha\varepsilon} f_t(z) \right) + \mathrm{N}_{\operatorname{dom} f}(z) \right) \quad \forall \alpha > 0.$$

Proof. Fix $\alpha > 0$. Denote by A the set in the right-hand side of the above equality. Without loss of generality (w.l.o.g.) we assume that $z = \theta$ and $f(\theta) = 0$. Set

$$T_\varepsilon := T_\varepsilon(\theta), \quad A_\varepsilon := \operatorname{co}\left(\bigcup_{t\in T_\varepsilon(z)} \partial_{\alpha\varepsilon} f_t(z)\right).$$

Note first that, $A_\varepsilon \subset \partial_{(1+\alpha)\varepsilon} f(\theta)$, which together with $\mathrm{N}_{\operatorname{dom} f}(\theta) = (\partial_{(1+\alpha)\varepsilon} f_t(\theta))_\infty$, implies that $\operatorname{cl}(A_\varepsilon + \mathrm{N}_{\operatorname{dom} f}(\theta)) \subset \partial_{(1+\alpha)\varepsilon} f(\theta)$. Indeed,

$$\begin{aligned} \langle x, x^*\rangle &\le f_t(x) - f_t(\theta) + \alpha\varepsilon \\ &\le f(x) + (1+\alpha)\varepsilon \quad \forall t \in T_\varepsilon,\ \forall x^* \in \partial_{\alpha\varepsilon} f_t(\theta),\ \forall x \in X, \end{aligned} \tag{15}$$

whence $x^* \in \partial_{(1+\alpha)\varepsilon} f(\theta)$. Hence $A \subset \cap_{\varepsilon>0} \partial_{(1+\alpha)\varepsilon} f(\theta) = \partial f(\theta)$.

Let us prove now that $\partial f(\theta) \subset A$. Notice that $f = h^*$, where $h := \inf_{t\in T} f_t^* \ge f^* \ge 0$. Moreover, for $x^* \notin A_\varepsilon$ and $\varepsilon > 0$, we have that $h(x^*) \ge (1\wedge\alpha)\varepsilon := \min\{1, \alpha\}\varepsilon$. Indeed, if $t \in T_\varepsilon$, then $x^* \notin \partial_{\alpha\varepsilon} f_t(\theta)$, and so $f_t^*(x^*) \ge f_t(\theta) + f_t^*(x^*) > \langle\theta, x^*\rangle + \alpha\varepsilon = \alpha\varepsilon$, while, for $t \in T \setminus T_\varepsilon$ we have that $f_t^*(x^*) \ge \langle\theta, x^*\rangle - f_t(\theta) > -f(\theta) + \varepsilon = \varepsilon$. Hence $f_t^*(x^*) \ge (1 \wedge \alpha)\varepsilon$ for every $t \in T$, and so $h(x^*) \ge (1 \wedge \alpha)\varepsilon$. Take now $\bar{x}^* \in X^*$, which is not in $\operatorname{cl}(A_\varepsilon + \mathrm{N}_{\operatorname{dom} f}(\theta))$ for (some) $\varepsilon > 0$. Using a separation theorem, there exist $\bar{x} \in X$ and $\gamma > 0$ such that

$$\langle \bar{x}, \bar{x}^*\rangle > \gamma + \langle \bar{x}, x^*\rangle + \langle \bar{x}, u^*\rangle \text{ for all } x^* \in A_\varepsilon \text{ and all } u^* \in \mathrm{N}_{\operatorname{dom} f}(\theta). \tag{16}$$

It follows that $\bar{x} \in (\mathrm{N}_{\operatorname{dom} f}(\theta))^- = \operatorname{cl}(\mathbb{R}_+ \operatorname{dom} f)$. Furthermore, note that from (15) we get $\operatorname{dom} f \subset \operatorname{dom} \sigma_{A_\varepsilon}$, and so $C := \mathbb{R}_+(\operatorname{dom} f) \subset \operatorname{dom} \sigma_{A_\varepsilon} = \operatorname{dom}(\sigma_{A_\varepsilon} - \bar{x}^*)$. Since $\operatorname{aff} C = \operatorname{aff}(\operatorname{dom} f)$ and $\operatorname{ri}(\operatorname{dom} f) \ne \emptyset$, we have that $\operatorname{ri} C \ne \emptyset$. Using Lemma 1 for $\sigma_{A_\varepsilon} - \bar{x}^*$ and C we obtain that one can take $\bar{x} \in \operatorname{dom} f$.

For $\lambda \in]0, 1[$ we have

$$\begin{aligned} f(\lambda\bar{x}) &= \sup\{\langle \lambda\bar{x}, x^*\rangle - h(x^*) \mid x^* \in X^*\} \\ &= \max\left\{ \sup_{x^*\in A_\varepsilon} [\langle \lambda\bar{x}, x^*\rangle - h(x^*)], \sup_{x^*\in X^*\setminus A_\varepsilon} [\langle \lambda\bar{x}, x^*\rangle - h(x^*)] \right\}. \end{aligned}$$

But, $h \ge 0$ and $\langle \bar{x}, \bar{x}^*\rangle \ge \gamma + \sigma_{A_\varepsilon}(\bar{x})$ (being a consequence of (16)) allow us to write

$$\begin{aligned} \sup_{x^*\in A_\varepsilon} [\langle \lambda\bar{x}, x^*\rangle - h(x^*)] &\le \sup_{x^*\in A_\varepsilon} \langle \lambda\bar{x}, x^*\rangle = \lambda\sigma_{A_\varepsilon}(\bar{x}) \\ &\le \lambda(-\gamma + \langle \bar{x}, \bar{x}^*\rangle) < \langle \lambda\bar{x}, \bar{x}^*\rangle, \end{aligned}$$

while the fact that $h \ge (1 \wedge \alpha)\varepsilon$ on $X^* \setminus A_\varepsilon$ implies that

$$\begin{aligned} \sup_{x^*\in X^*\setminus A_\varepsilon} [\langle \lambda\bar{x}, x^*\rangle - h(x^*)] &\le \sup_{x^*\in X^*\setminus A_\varepsilon} \lambda [\langle \bar{x}, x^*\rangle - h(x^*)] + \sup_{x^*\in X^*\setminus A_\varepsilon} (1-\lambda)[-h(x^*)] \\ &\le \lambda h^*(\bar{x}) - (1-\lambda)(1\wedge\alpha)\varepsilon = \lambda f(\bar{x}) - (1-\lambda)(1\wedge\alpha)\varepsilon. \end{aligned}$$

Thus, since

$$\lambda f(\bar{x}) - (1-\lambda)(1\wedge\alpha)\varepsilon < \langle \lambda\bar{x}, x^*\rangle$$

for $\lambda \in]0, 1[$ sufficiently small, for such λ we have $f(\lambda\bar{x}) < \langle \lambda\bar{x}, x^*\rangle$, whence $x^* \notin \partial f(\theta)$ because $f(\theta) = 0$. The proof is complete. □

Now we are ready to give the main result of the paper in which we establish the formula of the subdifferential of the supremum function f defined in (14).

THEOREM 4. *Let $\{f_t \mid t \in T\}$ be a nonempty family of convex functions $f_t : X \to \overline{\mathbb{R}}$ and set $f := \sup_{t\in T} f_t$. Assume that*

$$\operatorname{cl} f = \sup\{\operatorname{cl} f_t \mid t \in T\}.$$

Then, for every $z \in X$, we have

$$\partial f(z) = \bigcap_{L\in\mathcal{F}_z,\varepsilon>0} \operatorname{cl}\left(\operatorname{co}\left(\bigcup_{t\in T_\varepsilon(z)} \partial_{\alpha\varepsilon} f_t(z)\right) + \mathrm{N}_{L\cap\operatorname{dom} f}(z)\right) \quad \textit{for all } \alpha > 0.$$

Proof. Fix $\alpha > 0$ and denote by A the set in the right-hand side of the preceding equality.

Note first that the conclusion holds if $f(z) \notin \mathbb{R}$. Indeed, if $f(z) = +\infty$, then $\partial f(z) = \emptyset = \mathrm{N}_{L\cap\operatorname{dom} f}(z)$ for every $L \in \mathcal{F}_z$, and the conclusion holds trivially (taking into account (2)). If $f(z) = -\infty$, then $f_t(z) = -\infty$ for all $t \in T$, and so $\partial f(z) = \partial_{\alpha\varepsilon} f_t(z) = \emptyset$ for all $t \in T$ and all $\varepsilon > 0$, and again the conclusion holds trivially.

In the rest of the proof we assume that $f(z) \in \mathbb{R}$ and so, w.l.o.g., we take $z = \theta$ and $f(\theta) = 0$. To simplify the writing we use the notation

$$T_\varepsilon := T_\varepsilon(\theta), \quad A_\varepsilon := \operatorname{co}\left(\textstyle\bigcup_{t\in T_\varepsilon} \partial_{\alpha\varepsilon} f_t(\theta)\right), \quad \mathcal{F} := \mathcal{F}_\theta.$$

The inclusion $A \subset \partial f(\theta)$ easily follows by the definition of A_ε. Indeed, fix $x \in \operatorname{dom} f$, and let $L \in \mathcal{F}$. Then, by setting $E := L + \mathbb{R}x$ we get

$$\langle x, x^* + u^*\rangle \le \langle x, x^*\rangle \le f_t(x) - f_t(\theta) + \alpha\varepsilon \le f(x) + (1+\alpha)\varepsilon$$

for all $t \in T_\varepsilon$, $x^* \in \partial_{\alpha\varepsilon} f_t(\theta)$, and $u^* \in \mathrm{N}_{E\cap\operatorname{dom} f}(\theta)$, whence

$$\langle x, v^*\rangle \le f(x) + (1+\alpha)\varepsilon \text{ for all } v^* \in \operatorname{cl}\left(A_\varepsilon + \mathrm{N}_{E\cap\operatorname{dom} f}(\theta)\right).$$

Because $E \in \mathcal{F}$ and $\mathrm{N}_{E\cap\operatorname{dom} f}(\theta) \subset \mathrm{N}_{L\cap\operatorname{dom} f}(\theta)$, we deduce that

$$\langle x, v^*\rangle \le f(x) + (1+\alpha)\varepsilon \text{ for all } \varepsilon > 0 \text{ and } v^* \in A.$$

Hence $\langle x, v^*\rangle \le f(x) - f(\theta)$ for all $x \in \operatorname{dom} f$ and $v^* \in A$. Therefore, $A \subset \partial f(\theta)$. To prove the inclusion $\partial f(\theta) \subset A$ it suffices to assume that $\partial f(\theta) \ne \emptyset$ in which case, by (11),

$$\partial f(\theta) = \partial\left(\operatorname{cl} f\right)(\theta) \text{ and } (\operatorname{cl} f)(\theta) = f(\theta) = 0. \tag{17}$$

For this aim we shall introduce a family of functions satisfying the assumptions of Lemma 3.

Let us set $S := \{t \in T \mid \operatorname{cl} f_t \text{ is not proper}\}$. Then $\operatorname{cl} f_t$ takes its values in $\{-\infty, +\infty\}$ for $t \in S$ and so, because $(\operatorname{cl} f_t)(\theta) \le (\operatorname{cl} f)(\theta) = 0$ for $t \in T$, we obtain that $(\operatorname{cl} f_t)(\theta) = -\infty$ for $t \in S$; using our hypothesis we get $T \setminus S \ne \emptyset$.

Fix $L \in \mathcal{F}$ and define the family of functions $\{g_t \mid t \in T\} \subset \Gamma(X)$ by

$$g_t(x) := \begin{cases} \max\{(\operatorname{cl} f_t)(x), -1\} & \text{for } t \in S, \\ (\operatorname{cl} f_t)(x) & \text{for } t \in T \setminus S \end{cases}$$

and set

$$g(x) := \sup\{g_t(x) + \langle x, x^*\rangle \mid x^* \in L^\perp,\ t \in T\}.$$

(Observe that $g = \sup_{t\in T} g_t + \mathrm{I}_L$.) Then, since $g_t \geq \operatorname{cl} f_t$ for every $t \in T$, the current assumption yields

$$g = \sup_{t\in T} g_t + \mathrm{I}_L \geq \sup_{t\in T} \operatorname{cl} f_t + \mathrm{I}_L = \operatorname{cl} f + \mathrm{I}_L.$$

Furthermore, thanks to (17), there exists a convex neighborhood U of θ such that $(\operatorname{cl} f)(x) > -1$ for every $x \in U$. Hence for $x \in U \cap L$ we have either $(\operatorname{cl} f)(x) = +\infty \geq g(x)$ or $(\operatorname{cl} f)(x) < +\infty$; in this case for $t \in S$ one has $(\operatorname{cl} f_t)(x) = -\infty$, and so $g_t(x) = -1 \leq (\operatorname{cl} f)(x)$, while for $t \in T \setminus S$ one has $g_t(x) = (\operatorname{cl} f_t)(x) \leq (\operatorname{cl} f)(x)$. We deduce that $g(x) \leq (\operatorname{cl} f)(x)$ for $x \in U \cap L$. Therefore,

$$g(x) = (\operatorname{cl} f)(x) + \mathrm{I}_L(x) \quad \text{for every } x \in U. \tag{18}$$

Moreover, because $L \cap U \cap \operatorname{dom} f \subset L \cap U \cap \operatorname{dom}(\operatorname{cl} f) = U \cap \operatorname{dom} g$, we get

$$\mathrm{N}_{\operatorname{dom} g}(\theta) \subset \mathrm{N}_{L\cap \operatorname{dom} f}(\theta). \tag{19}$$

Now set

$$T'_\varepsilon := \{t \in T \mid g_t(\theta) \geq -\varepsilon\}.$$

Then $T'_\varepsilon \subset T_\varepsilon \setminus S$ for $\varepsilon \in]0,1[$. In fact, since $g_t(\theta) = -1$ for $t \in S$, we have that $T'_\varepsilon \subset T \setminus S$. Hence, for $t \in T'_\varepsilon$ we write $0 \geq f_t(\theta) \geq (\operatorname{cl} f_t)(\theta) = g_t(\theta) \geq -\varepsilon$, and so $t \in T_\varepsilon$. Moreover, for $t \in T'_\varepsilon$ we have that $\partial_{\alpha\varepsilon}(\operatorname{cl} f_t)(\theta) \subset \partial_{(1+\alpha)\varepsilon} f_t(\theta)$. Indeed, since we have $f_t(\theta) - (\operatorname{cl} f_t)(\theta) \leq f(\theta) - g_t(\theta) = g(\theta) - g_t(\theta) \leq \varepsilon$, Lemma 2 yields $\partial_{\alpha\varepsilon}(\operatorname{cl} f_t)(\theta) = \partial_{\alpha\varepsilon + f_t(\theta) - (\operatorname{cl} f_t)(\theta)} f_t(\theta) \subset \partial_{(1+\alpha)\varepsilon} f_t(\theta)$. In view of these observations we get

$$\operatorname{co}\left(\bigcup_{t\in T'_\varepsilon} \partial_{\alpha\varepsilon} g_t(\theta)\right) \subset \operatorname{co}\left(\bigcup_{t\in T'_\varepsilon} \partial_{(1+\alpha)\varepsilon} f_t(\theta)\right) \quad \text{for all } \varepsilon \in]0,1[. \tag{20}$$

Now we go back to the proof of the inclusion $\partial f(\theta) \subset A$. We apply Lemma 3 for the family $\{g_{(t,x^*)} \mid (t,x^*) \in T \times L^\perp\} \subset \Gamma(X)$ with $g_{(t,x^*)} := g_t + x^*$ and α (this is possible because $g = \sup\{g_{(t,x^*)} \mid (t,x^*) \in T \times L^\perp\}$ and $\operatorname{dom} g \subset L$, and so $\operatorname{ri}(\operatorname{dom} g) \neq \emptyset$, L being a finite-dimensional space). We obtain

$$\begin{aligned}\partial g(\theta) &= \bigcap_{\varepsilon>0} \operatorname{cl}\left(\operatorname{co}\left(\bigcup_{t\in T'_\varepsilon, x^*\in L^\perp} \partial_{\alpha\varepsilon}(g_t + x^*)(\theta)\right) + \mathrm{N}_{\operatorname{dom} g}(\theta)\right)\\ &= \bigcap_{\varepsilon>0} \operatorname{cl}\left(\operatorname{co}\left(\bigcup_{t\in T'_\varepsilon} \partial_{\alpha\varepsilon} g_t(\theta)\right) + L^\perp + \mathrm{N}_{\operatorname{dom} g}(\theta)\right).\end{aligned}$$

Then in view of the evident fact that $L^\perp + \mathrm{N}_{L\cap\operatorname{dom} f}(\theta) \subset \mathrm{N}_{L\cap\operatorname{dom} f}(\theta)$, and using (19) and (20), we get

$$\begin{aligned}\partial g(\theta) &\subset \bigcap_{\varepsilon\in]0,1[} \operatorname{cl}\left(\operatorname{co}\left(\bigcup_{t\in T'_\varepsilon} \partial_{\alpha\varepsilon} g_t(\theta)\right) + \mathrm{N}_{L\cap\operatorname{dom} f}(\theta)\right)\\ &\subset \bigcap_{\varepsilon\in]0,1[} \operatorname{cl}\left(\operatorname{co}\left(\bigcup_{t\in T_\varepsilon} \partial_{(1+\alpha)\varepsilon} f_t(\theta)\right) + \mathrm{N}_{L\cap\operatorname{dom} f}(\theta)\right).\end{aligned}$$

Hence, for each $\varepsilon \in]0,1[$ we obtain that, taking into account (17) and (18),

$$\begin{aligned}\partial f(\theta) \subset \partial(\operatorname{cl} f)(\theta) + L^{\perp} &= \partial(\operatorname{cl} f)(\theta) + \partial \mathrm{I}_L(\theta) \subset \partial\left(\operatorname{cl} f + \mathrm{I}_L\right)(\theta) \\ &= \partial g(\theta) \subset \operatorname{cl}\left(\operatorname{co}\left(\bigcup_{t\in T_\varepsilon} \partial_{(1+\alpha)\varepsilon} f_t(\theta)\right) + \mathrm{N}_{L\cap \operatorname{dom} f}(\theta)\right)\end{aligned}$$

for all $\varepsilon \in]0,1[$. Since $\delta := \frac{\alpha}{1+\alpha}\varepsilon \in]0,\varepsilon[\subset]0,1[$ (for $\varepsilon \in]0,1[$), we also have that

$$\partial f(\theta) \subset \operatorname{cl}\left(\operatorname{co}\left(\bigcup_{t\in T_\delta} \partial_{\alpha\varepsilon} f_t(\theta)\right) + \mathrm{N}_{L\cap \operatorname{dom} f}(\theta)\right) \subset \operatorname{cl}\left(A_\varepsilon + \mathrm{N}_{L\cap \operatorname{dom} f}(\theta)\right)$$

for all $\varepsilon \in]0,1[$. Since $\varepsilon \in]0,1[$ and $L \in \mathcal{F}$ were arbitrarily chosen, we obtain that

$$\partial f(\theta) \subset \bigcap_{L\in\mathcal{F},\varepsilon\in]0,1[} \operatorname{cl}\left(A_\varepsilon + \mathrm{N}_{L\cap \operatorname{dom} f}(\theta)\right) = \bigcap_{L\in\mathcal{F},\varepsilon>0} \operatorname{cl}\left(A_\varepsilon + \mathrm{N}_{L\cap \operatorname{dom} f}(\theta)\right) = A.$$

The proof is complete. □

Theorem 4 provides a complete description for ∂f only in terms of the data functions f_t, $t \in T$. Other descriptions will be provided in Theorem 6 below. We first establish the following lemma, which provides a straightforward infinite-dimensional extension of the corresponding statements in [7, Prop. 4].

LEMMA 5. *Let $T \neq \emptyset$ and $\{f_t \mid t \in T\} \subset \Gamma(X)$, and set $f := \sup\{f_t \mid t \in T\}$. Then, for every $z \in \operatorname{dom} f$, we have that*

$$\mathrm{N}_{\operatorname{dom} f}(z) = \{v^* \in X^* \mid (v^*, \langle v^*, z\rangle) \in [\overline{\operatorname{co}}\,(\cup_{t\in T} \operatorname{gph} f_t^*)]_\infty\} \tag{21}$$

$$= \{v^* \in X^* \mid (v^*, \langle v^*, z\rangle) \in [\overline{\operatorname{co}}\,(\cup_{t\in T} \operatorname{epi} f_t^*)]_\infty\} \tag{22}$$

$$= \{v^* \in X^* \mid (v^*, \langle v^*, z\rangle) \in (\operatorname{epi} f^*)_\infty\} \tag{23}$$

$$= \{v^* \in X^* \mid (v^*, \langle v^*, z\rangle) \in \operatorname{epi}(\sigma_{\operatorname{dom} f})\}. \tag{24}$$

Proof. We assume that f is proper. Statement (24) is just the definition of $\mathrm{N}_{\operatorname{dom} f}(z)$. As seen in Lemma 3, we have that

$$(\inf_{t\in T} f_t^*)^* = \sup_{t\in T} f_t^{**} = \sup_{t\in T} f_t = f.$$

Since f is proper we obtain that

$$f^* = (\inf_{t\in T} f_t^*)^{**} = \overline{\operatorname{co}}\left(\inf_{t\in T} f_t^*\right),$$

that is, $\operatorname{epi} f^* = \overline{\operatorname{co}}(\cup_{t\in T} \operatorname{epi} f_t^*)$; moreover, by [32, Exer. 2.23] one has $(\operatorname{epi} f^*)_\infty = \operatorname{epi}(\sigma_{\operatorname{dom} f})$. Using these two relations we get statements (22) and (23). To finish the proof, it suffices to establish the equality between the sets appearing in the right-hand sides of (21) and (22), say, $E_1(z)$ and $E_2(z)$, respectively, or simply the inclusion $E_2(z) \subset E_1(z)$, since the opposite inclusion is trivial. Indeed, because for any proper function $g : X \to \overline{\mathbb{R}}$ one has $\operatorname{gph} g + \mathbb{R}_+(\theta,1) = \operatorname{epi} g$, we obtain that

$$\overline{\operatorname{co}}\,(\cup_{t\in T} \operatorname{gph} f_t^*) \subset \operatorname{cl}\,[\overline{\operatorname{co}}\,(\cup_{t\in T} \operatorname{gph} f_t^*) + \mathbb{R}_+(\theta,1)] = \overline{\operatorname{co}}\,(\cup_{t\in T} \operatorname{epi} f_t^*) = \operatorname{epi} f^*.$$

Since f^* is proper, we have $[\overline{\operatorname{co}}\,(\cup_{t\in T} \operatorname{gph} f_t^*)]_\infty \cap -[\mathbb{R}_+(\theta,1)]_\infty = \{(\theta,0)\}$, and so by [30, Cor. 3.12] (see also [16, Thm. 1.1]), we obtain that $\overline{\operatorname{co}}\,(\cup_{t\in T} \operatorname{gph} f_t^*) + \mathbb{R}_+(\theta,1)$ is closed, whence $\overline{\operatorname{co}}\,(\cup_{t\in T} \operatorname{gph} f_t^*) + \mathbb{R}_+(\theta,1) = \overline{\operatorname{co}}\,(\cup_{t\in T} \operatorname{epi} f_t^*)$, and

$$\begin{aligned}[\overline{\operatorname{co}}\,(\cup_{t\in T} \operatorname{epi} f_t^*)]_\infty &= [\overline{\operatorname{co}}\,(\cup_{t\in T} \operatorname{gph} f_t^*) + \mathbb{R}_+(\theta,1)]_\infty \\ &= [\overline{\operatorname{co}}\,(\cup_{t\in T} \operatorname{gph} f_t^*)]_\infty + \mathbb{R}_+(\theta,1).\end{aligned}$$

Take $v^* \in E_2(z)$; using the preceding relation, $(v^*, \langle v^*, z\rangle) = (x^*, \eta + \lambda)$ for some $(x^*, \eta) \in [\overline{\text{co}}\,(\cup_{t\in T} \text{gph}\, f_t^*)]_\infty$, and $\lambda \geq 0$. Moreover, since $\text{dom}\, f \times \{-1\} \subset \text{dom}\,(\sigma_{\text{epi}\, f^*}) \subset [(\text{epi}\, f^*)_\infty]^-$, we obtain that

$$\text{dom}\, f \times \{-1\} \subset [(\overline{\text{co}}\,(\cup_{t\in T} \text{epi}\, f_t^*))_\infty]^- \subset [(\overline{\text{co}}\,(\cup_{t\in T} \text{gph}\, f_t^*))_\infty]^-,$$

and so $\langle (x^*, \eta), (z, -1)\rangle \leq 0$. Since $v^* = x^*$, it follows that

$$\lambda = \langle (v^*, \eta), (z, -1)\rangle = \langle (x^*, \eta), (z, -1)\rangle \leq 0;$$

hence $\lambda = 0$, and so $(v^*, \langle v^*, z\rangle) = (x^*, \eta) \in [\overline{\text{co}}\,(\cup_{t\in T} \text{gph}\, f_t^*)]_\infty$. This shows that $v^* \in E_1(z)$. ∎

We have the following theorem in which, for simplicity, we suppose that $f_t \in \Gamma(X)$ for all $t \in T$.

THEOREM 6. *Let $T \neq \emptyset$ and $\{f_t \mid t \in T\} \subset \Gamma(X)$, and set $f := \sup_{t\in T} f_t$. Then, for every $z \in X$ and every $\alpha > 0$, we have that*

$$\partial f(z) = \bigcap_{L\in\mathcal{F}_z, \varepsilon>0} \overline{\text{co}}\Big(A_L + \bigcup_{t\in T_\varepsilon(z)} \partial_{\alpha\varepsilon} f_t(z)\Big) = \bigcap_{L\in\mathcal{F}_z, \varepsilon>0} \overline{\text{co}}\Big(B_L + \bigcup_{t\in T_\varepsilon(z)} \partial_{\alpha\varepsilon} f_t(z)\Big),$$

where

$$A_L := \left\{ v^* \in X^* \,\middle|\, (v^*, \langle v^*, z\rangle) \in \left[\overline{\text{co}}\Big((L^\perp \times \mathbb{R}_+) \cup \Big(\bigcup_{t\in T} \text{epi}\, f_t^*\Big)\Big)\right]_\infty \right\},$$

$$B_L := \left\{ v^* \in X^* \,\middle|\, (v^*, \langle v^*, z\rangle) \in \left[\overline{\text{co}}\Big((L^\perp \times \{0\}) \cup \Big(\bigcup_{t\in T} \text{gph}\, f_t^*\Big)\Big)\right]_\infty \right\}.$$

Proof. According to Theorem 4 it suffices to write $\text{N}_{L\cap\text{dom}\, f}(z)$ in terms of the data functions f_t for each $L \in \mathcal{F}_z$. Indeed, by Lemma 5 applied to the family $\{f_t \mid t \in T\} \cup \{\text{I}_L\} \subset \Gamma(X)$, we have $\text{N}_{L\cap\text{dom}\, f}(z) = A_L = B_L$; we used the fact that $(\text{I}_L)^* = \text{I}_{L^\perp}$, and so $\text{epi}\,(\text{I}_L)^* = \text{epi}\,(\text{I}_{L^\perp}) = L^\perp \times \mathbb{R}_+$ and $\text{gph}\,(\text{I}_L)^* = \text{gph}\,(\text{I}_{L^\perp}) = L^\perp \times \{0\}$. ∎

In the affine case (f_t affine) our formula takes a simpler form.

COROLLARY 7. *Assume that $T \neq \emptyset$ and $f := \sup\{\langle a_t^*, \cdot\rangle - \beta_t \mid t \in T\}$, with $a_t^* \in X^*$ and $\beta_t \in \mathbb{R}$. Then, for every $z \in X$, we have that*

$$\partial f(z) = \bigcap_{L\in\mathcal{F}_z, \varepsilon>0} \text{cl}\,(\text{co}\{a_t^* \mid t \in T_\varepsilon(z)\} + B_L),$$

where $T_\varepsilon(z) := \{t \in T \mid \langle a_t^*, z\rangle - \beta_t \geq f(z) - \varepsilon\}$ and

$$B_L := \left\{ v^* \in X^* \;\middle|\; (v^*, \langle v^*, z\rangle) \in \left[\overline{\text{co}}\left((L^\perp \times \{0\}) \cup \{(a_t^*, \beta_t) \mid t \in T\}\right)\right]_\infty \right\}.$$

In particular, for a given nonempty set $A \subset X^*$, we have that

$$\partial\sigma_A(z) = \bigcap_{L\in\mathcal{F}_z, \varepsilon>0} \text{cl}\left(\text{co}(A_\varepsilon) + \left[\overline{\text{co}}\left(L^\perp \cup A\right)\right]_\infty \cap \{z\}^\perp\right),$$

where $A_\varepsilon := \{a^* \in A \mid \langle z, a^*\rangle \geq \sigma_A(z) - \varepsilon\}$.

Proof. These formulae easily follow by Theorem 6, similarly as in [7, Prop. 1]. ∎

The following corollary gives us a simplified representation for the subdifferential set of f when $\text{ri}(\text{dom}\, f) \neq \emptyset$. This is also an extension of Lemma 3 when the functions f_t are not necessarily lsc.

COROLLARY 8. *Let $\{f_t \mid t \in T\}$ be a nonempty family of convex functions $f_t : X \to \overline{\mathbb{R}}$, and set $f := \sup_{t\in T} f_t$. Assume that* $\operatorname{ri}(\operatorname{dom} f) \neq \emptyset$. *Then, for every $z \in X$ and $\alpha > 0$, we have that*

$$\partial f(z) = \bigcap_{\varepsilon>0} \operatorname{cl}\left(\operatorname{co}\left(\bigcup_{t\in T_\varepsilon(z)} \partial_{\alpha\varepsilon} f_t(z) \right) + \mathrm{N}_{\operatorname{dom} f}(z) \right).$$

Proof. The inclusion "$\supset$" follows immediately by Theorem 4, since we have $\mathrm{N}_{\operatorname{dom} f}(z) \subset \mathrm{N}_{L\cap\operatorname{dom} f}(z)$ for every $L \in \mathcal{F}_z$. To prove the inclusion "$\subset$", let $\alpha > 0$ be fixed, and let $\partial f(z) \neq \emptyset$ (otherwise the inclusion is obvious). We (may) assume that $z = \theta$ and $f(\theta) = 0$. Then it suffices to show that $\partial f(\theta) \subset \operatorname{cl}\big(\operatorname{co}\big(\bigcup_{t\in T_\varepsilon(\theta)} \partial_{\alpha\varepsilon} f_t(\theta)\big) + \mathrm{N}_{\operatorname{dom} f}(\theta)\big)$ for any given $\varepsilon > 0$. Let $V \in \mathcal{V}$, that is, V is a θ–neighborhood in X^*, and $L \in \mathcal{F}_\theta$ be such that $L^\perp \subset V$. We may suppose w.l.o.g. that $L \cap \operatorname{ri}(\operatorname{dom} f) \neq \emptyset$, which in particular, implies that $L \cap \operatorname{ri}(\mathbb{R}_+ \operatorname{dom} f) \neq \emptyset$. Using (4) we obtain that $\operatorname{cl}(L \cap \mathbb{R}_+ \operatorname{dom} f) = L \cap \operatorname{cl}(\mathbb{R}_+ \operatorname{dom} f)$; this implies that (see [32, p. 7])

$$\mathrm{N}_{L\cap\operatorname{dom} f}(\theta) = (L \cap \operatorname{cl}(\mathbb{R}_+ \operatorname{dom} f))^- = \operatorname{cl}(L^- + (\mathbb{R}_+ \operatorname{dom} f)^-) = \operatorname{cl}\left(L^\perp + \mathrm{N}_{\operatorname{dom} f}(\theta)\right).$$

So, by using once again Theorem 4 and (3), we obtain that

$$\begin{aligned} \partial f(\theta) &\subset \operatorname{cl}\left[\operatorname{co}\left(\textstyle\bigcup_{t\in T_\varepsilon(\theta)} \partial_{\alpha\varepsilon} f_t(\theta)\right) + \mathrm{N}_{L\cap\operatorname{dom} f}(\theta)\right] \\ &= \operatorname{cl}\left[\operatorname{co}\left(\textstyle\bigcup_{t\in T_\varepsilon(\theta)} \partial_{\alpha\varepsilon} f_t(\theta)\right) + L^\perp + \mathrm{N}_{\operatorname{dom} f}(\theta)\right] \\ &\subset \operatorname{co}\left(\textstyle\bigcup_{t\in T_\varepsilon(\theta)} \partial_{\alpha\varepsilon} f_t(\theta)\right) + \mathrm{N}_{\operatorname{dom} f}(\theta) + V. \end{aligned}$$

As V is an arbitrary θ–neighborhood, we get that

$$\begin{aligned} \partial f(\theta) &\subset \bigcap_{V\in\mathcal{V}} \left(\operatorname{co}\left(\bigcup_{t\in T_\varepsilon(\theta)} \partial_{\alpha\varepsilon} f_t(\theta) \right) + \mathrm{N}_{\operatorname{dom} f}(\theta) + V \right) \\ &= \operatorname{cl}\left(\operatorname{co}\left(\textstyle\bigcup_{t\in T_\varepsilon(\theta)} \partial_{\alpha\varepsilon} f_t(\theta)\right) + \mathrm{N}_{\operatorname{dom} f}(\theta)\right), \end{aligned}$$

which finishes the proof. □

From a geometric point of view the closedness criterion given in Theorem 4 is equivalent to

$$\operatorname{cl}\left(\bigcap_{t\in T} \operatorname{epi} f_t \right) = \bigcap_{t\in T} \operatorname{cl}(\operatorname{epi} f_t), \tag{25}$$

which is itself satisfied by a wide variety of convex functions as the following result shows.

COROLLARY 9. *Let $\{f_t \mid t \in T\}$ be a nonempty family of convex functions $f_t : X \to \overline{\mathbb{R}}$, and set $f := \sup_{t\in T} f_t$. Assume that one of the following conditions holds*:

(i) *All of the functions f_t, with $t \in T$ are lsc.*
(ii) *There exists $x_0 \in \operatorname{dom} f$ such that f_t is continuous at x_0 for every $t \in T$.*
(iii) *$T := \{1, \ldots, k, k+1\}$, and there exists $x_0 \in \cap_{i=1}^{k+1} \operatorname{dom} f_i$ such that $f_1, \ldots, f_k$ are continuous at x_0.*
(iv) *$X = \mathbb{R}^n$ and $\operatorname{dom} f \cap (\cap_{t\in T} \operatorname{ri}(\operatorname{dom} f_t))$ is nonempty.*

Then, we have that

$$\operatorname{cl} f = \sup\{\operatorname{cl} f_t \mid t \in T\},$$

and, consequently, for every $z \in X$ and $\alpha > 0$, it holds that

$$\partial f(z) = \bigcap_{L \in \mathcal{F}_z, \varepsilon > 0} \operatorname{cl}\left(\operatorname{co}\left(\bigcup_{t \in T_\varepsilon(z)} \partial_{\alpha\varepsilon} f_t(z) \right) + \mathrm{N}_{L \cap \operatorname{dom} f}(z) \right).$$

Proof. Setting $A_t := \operatorname{epi} f_t$ for $t \in T$ and $A := \operatorname{epi} f$, one has always $A = \cap_{t \in T} A_t$, and we have to show that $\operatorname{cl} A = \cap_{t \in T} \operatorname{cl}(A_t)$. The inclusion $\operatorname{cl} A \subset \cap_{t \in T} \operatorname{cl}(A_t)$ being obvious, it remains to prove that $\operatorname{cl} A \supset \cap_{t \in T} \operatorname{cl}(A_t)$ in each of the following cases.

(i) It is immediate.

(ii) First observe that [31, Lem. 13] is valid even if f is not proper. Consider $\mu > f(x_0)$. Applying this result we obtain that $y_0 := (x_0, \mu) \in \bigcap_{t \in T} \operatorname{int} A_t$. Now if $x \in \bigcap_{t \in T} \operatorname{cl} A_t$, then $(1 - \lambda)x + \lambda y_0 \in \bigcap_{t \in T} \operatorname{int} A_t \subset A$ for every $\lambda \in]0, 1[$, whence $x \in \operatorname{cl} A$.

(iii) Set $B := \bigcap_{t=1}^k A_t$. Then, similarly as in (ii), we can show that $y_0 := (x_0, \mu) \in A_{k+1} \cap \operatorname{int} B$. Hence

$$\operatorname{cl}\left(\bigcap_{t \in T} A_t\right) = \operatorname{cl}\left(A_{k+1} \cap B\right) = \operatorname{cl} A_{k+1} \cap \operatorname{cl} B = \operatorname{cl} A_{k+1} \cap \left(\bigcap_{t=1}^k \operatorname{cl} A_t\right) = \bigcap_{t \in T} \operatorname{cl} A_t.$$

(iv) This is practically [22, Thm. 9.4].

Taking into account Theorem 4, the final conclusion follows. □

The following result (for $\alpha = 1$) is due to Volle (see, e.g., [28, Thm. A]) and is originally established in the context of normed spaces.

COROLLARY 10. *Let $\{f_t \mid t \in T\}$ be a nonempty family of convex functions $f_t : X \to \overline{\mathbb{R}}$, and set $f := \sup_{t \in T} f_t$. Assume that f is finite and continuous at $z \in X$. Then, we have*

$$\partial f(z) = \bigcap_{\varepsilon > 0} \overline{\operatorname{co}}\left(\bigcup_{t \in T_\varepsilon(z)} \partial_{\alpha\varepsilon} f_t(z) \right) \quad \text{for all } \alpha > 0.$$

Proof. Because f is finite and continuous at z, we have that $z \in \operatorname{int}(\operatorname{dom} f)$, and so $\mathrm{N}_{\operatorname{dom} f}(z) = \{\theta\}$. Further, as $z \in \cap_{t \in T} \operatorname{int}(\operatorname{dom} f)$ Condition (ii) of Corollary 9 yields $\operatorname{cl} f = \sup\{\operatorname{cl} f_t \mid t \in T\}$. Of course, $\operatorname{ri}(\operatorname{dom} f) = \operatorname{int}(\operatorname{dom} f) \neq \emptyset$, and so the conclusion follows from Corollary 8. □

In order to derive Brøndsted's formula (Corollary 12 below) we shall need the following result on normal cones.

LEMMA 11. (i) *Let $g_1, \ldots, g_k \in \Gamma(X)$, $f \in \Gamma(Y)$, and consider a continuous affine mapping $A : X \to Y$, where X and Y are (separated) locally convex spaces. Then, for every $z \in \operatorname{dom}(g_1 + \cdots + g_k + f \circ A)$ and all $\varepsilon, \varepsilon_1, \ldots, \varepsilon_k > 0$, we have that*

$$\mathrm{N}_{\operatorname{dom}(g_1 + \cdots + g_k + f \circ A)}(z) = \left[\operatorname{cl}\left(\partial_{\varepsilon_1} g_1(z) + \cdots + \partial_{\varepsilon_k} g_k(z) + A_0^* \partial_\varepsilon f(Az) \right) \right]_\infty,$$

where A_0 is the linear part of A, and A_0^* is the adjoint of A_0.

(ii) *Let $\{f_1, \ldots, f_m\} \subset \Gamma(X)$, with $m \geq 2$ and $0 \leq k \leq m$. Then, for all $z \in \bigcap_{t=1}^m \operatorname{dom} f_t$ and all $\varepsilon_1, \ldots, \varepsilon_m > 0$, we have that*

$$\mathrm{N}_{\cap_{t=1}^m \operatorname{dom} f_t}(z) = \left[\operatorname{cl}\left(\partial_{\varepsilon_1} f_1(z) + \cdots + \partial_{\varepsilon_k} f_k(z) + \operatorname{co}\left(\partial_{\varepsilon_{k+1}} f_{k+1}(z) \cup \cdots \cup \partial_{\varepsilon_m} f_m(z) \right) \right) \right]_\infty,$$

where $C_1 + \cdots + C_k := \emptyset$ if $k = 0$ and $C_{k+1} \cup \cdots \cup C_m := \emptyset$ if $k = m$.

Proof. (i) Using (7) and (13), as well as the fact that $\mathbb{R}_+(B\cap C)=\mathbb{R}_+B\cap\mathbb{R}_+C$ when B and C are convex sets containing θ, we get that

$$\begin{aligned}
&[(\mathrm{cl}\,(\partial_{\varepsilon_1}g_1(z)+\cdots+\partial_{\varepsilon_k}g_k(z)+A_0^*\partial_\varepsilon f(Az)))_\infty]^-\\
&\quad=\mathrm{cl}(\mathrm{dom}(\sigma_{\partial_{\varepsilon_1}g_1(z)}+\cdots+\sigma_{\partial_{\varepsilon_k}g_k(z)}+\sigma_{\partial_\varepsilon f(Az)}\circ A_0))\\
&\quad=\mathrm{cl}\left(\mathrm{dom}((g_1)'_{\varepsilon_1}(z,\cdot))\cap\cdots\cap\mathrm{dom}((g_k)'_{\varepsilon_k}(z,\cdot))\cap A_0^{-1}\,\mathrm{dom}(f'_\varepsilon(Az,\cdot))\right)\\
&\quad=\mathrm{cl}\left(\mathbb{R}_+\left(\mathrm{dom}\,g_1-z\right)\cap\cdots\cap\mathbb{R}_+(\mathrm{dom}\,g_k-z)\cap A_0^{-1}\left(\mathbb{R}_+(\mathrm{dom}\,f-Az)\right)\right)\\
&\quad=\mathrm{cl}\left(\mathbb{R}_+\left(\mathrm{dom}(g_1+\cdots+g_k+f\circ A)-z\right)\right),
\end{aligned}$$

whence the conclusion follows using (5).

(ii) Taking $f=0$ in (i) and observing that $\mathrm{dom}(g_1+\cdots+g_k)=\bigcap_{t=1}^k\mathrm{dom}\,g_t$, we get that $\mathrm{N}_{\cap_{t=1}^k\mathrm{dom}\,g_t}(z)=[\mathrm{cl}\,(\partial_{\varepsilon_1}g_1(z)+\cdots+\partial_{\varepsilon_k}g_k(z))]_\infty$. The conclusion follows now using (8). □

The following result is due to Brøndsted (e.g., [2]); see also [7, Prop. 7] where such a formula is extended to families of infinitely many convex functions defined on $\mathbb{R}^n$.

COROLLARY 12. *Consider the convex functions $f_i:X\to\overline{\mathbb{R}}$ for $i=1,\dots,k$, and set $f:=\max\{f_1,\dots,f_k\}$. Assume that*

$$\mathrm{cl}\,f=\max\{\mathrm{cl}\,f_1,\dots,\mathrm{cl}\,f_k\}.$$

Given $z\in X$ such that $(\mathrm{cl}\,f)(z)=(\mathrm{cl}\,f_i)(z)$ for $i=1,\dots,k$, we have that

$$\partial f(z)=\bigcap\nolimits_{\varepsilon>0}\overline{\mathrm{co}}\left(\bigcup\nolimits_{i=1}^k\partial_\varepsilon f_i(z)\right).$$

Proof. It suffices to establish the inclusion "$\subset$" in the nontrivial case $\partial f(z)\neq\emptyset$. According to (11), the function f is proper and satisfies $f(z)=(\mathrm{cl}\,f)(z)\in\mathbb{R}$ and $\partial f(z)=\partial\,(\mathrm{cl}\,f)\,(z)$. Because

$$(\mathrm{cl}\,f_i)(z)\le f_i(z)\le f(z)=(\mathrm{cl}\,f)(z)=(\mathrm{cl}\,f_i)(z),$$

we obtain that $(\mathrm{cl}\,f_i)(z)=f_i(z)=f(z)\in\mathbb{R}$ for all $i\in T:=\{1,\dots,k\}$; hence the functions $\mathrm{cl}\,f_i$, with $i\in T$, are proper. Furthermore, using (10) we get

$$\partial_\varepsilon(\mathrm{cl}\,f_i)(z)=\partial_\varepsilon f_i(z)\ \text{for all}\ \varepsilon>0\ \text{and}\ i\in T. \tag{26}$$

Fix $\varepsilon>0$; it is clear that $T_\varepsilon(z)=T$. Let $V\in\mathcal{V}$, that is, V is a convex θ–neighborhood in X^*, and take $L\in\mathcal{F}_z$ such that $L^\perp\subset V$ ($\Leftrightarrow L^\perp\subset\frac{1}{2}V$). Applying Theorem 4 for $\{\mathrm{cl}\,f_1,\dots,\mathrm{cl}\,f_k\}$ and $\alpha=1$, we have that

$$\partial\,(\mathrm{cl}\,f)\,(z)\subset\mathrm{cl}\left(\mathrm{co}\left(\bigcup\nolimits_{i\in T}\partial_\varepsilon(\mathrm{cl}\,f_i)(z)\right)+\mathrm{N}_{L\cap\mathrm{dom}(\mathrm{cl}\,f)}(z)\right).$$

But Lemma 11(ii) applied to $\{\mathrm{cl}\,f_1,\dots,\mathrm{cl}\,f_k,\mathrm{I}_L\}$ implies that

$$\mathrm{N}_{L\cap\mathrm{dom}(\mathrm{cl}\,f)}(z)=\left[\overline{\mathrm{co}}\left(L^\perp+\left(\bigcup\nolimits_{i\in T}\partial_\varepsilon(\mathrm{cl}\,f_i)(z)\right)\right)\right]_\infty,$$

where we used the property $\partial_\varepsilon\mathrm{I}_L(z)=L^\perp$. Thus, taking into account (3) and (26), we get that

$$\begin{aligned}
\partial f(z)=\partial\,(\mathrm{cl}\,f)\,(z)&\subset\mathrm{cl}\left(\overline{\mathrm{co}}\left(\bigcup\nolimits_{i\in T}\partial_\varepsilon(\mathrm{cl}\,f_i)(z)\right)+\left[\overline{\mathrm{co}}\left(L^\perp+\left(\bigcup\nolimits_{i\in T}\partial_\varepsilon(\mathrm{cl}\,f_i)(z)\right)\right)\right]_\infty\right)\\
&\subset\mathrm{cl}\left(\overline{\mathrm{co}}\left(L^\perp+\left(\bigcup\nolimits_{i\in T}\partial_\varepsilon f_i(z)\right)\right)+\left[\overline{\mathrm{co}}\left(L^\perp+\left(\bigcup\nolimits_{i\in T}\partial_\varepsilon f_i(z)\right)\right)\right]_\infty\right)\\
&=\overline{\mathrm{co}}\left(L^\perp+\left(\bigcup\nolimits_{i\in T}\partial_\varepsilon f_i(z)\right)\right)=\mathrm{cl}\left(L^\perp+\mathrm{co}\left(\bigcup\nolimits_{i\in T}\partial_\varepsilon f_i(z)\right)\right)\\
&\subset L^\perp+\mathrm{co}\left(\bigcup\nolimits_{i\in T}\partial_\varepsilon f_i(z)\right)+\tfrac{1}{2}V\subset\mathrm{co}\left(\bigcup\nolimits_{i\in T}\partial_\varepsilon f_i(z)\right)+V.
\end{aligned}$$

Consequently,

$$\partial f(z) \subset \bigcap_{V\in\mathcal{V}} \left(\operatorname{co}\left(\bigcup_{i\in T}\partial_\varepsilon f_i(z)\right)+V\right) = \overline{\operatorname{co}}\left(\bigcup_{i\in T}\partial_\varepsilon f_i(z)\right).$$

Finally, the conclusion follows by taking the intersection over $\varepsilon > 0$. ∎

4. Other calculus rules. Throughout this section, we consider two convex functions $f : Y \to \overline{\mathbb{R}}$ and $g : X \to \overline{\mathbb{R}}$, where X and Y are (separated) real locally convex spaces, and a continuous affine mapping $A : X \to Y$ defined by

$$Ax = A_0x + b,$$

where A_0 is the linear part of A and $b \in Y$. We denote by A_0^* the adjoint operator of A_0.

We show that our rule given in Theorem 4, providing formulas for the subdifferential of the supremum function, also gives calculus rules for other operations expressed by means of the convex function $g+f\circ A$. The resulting formulas are not new, but our aim here is to highlight the unifying character of Theorem 4, which also yields alternative proofs that do not rely on the commonly used approach based on conjugation theory [23].

At the first stage, we derive in the following theorem a slight extension of the Hiriart-Urruty–Phelps formula [11]. This allows us to express the subdifferential of $g + f \circ A$ in terms of the approximate subdifferentials of f and g. For comparative purposes, when the involved spaces X and Y are Banach, this is equivalent to writing $\partial(g + f \circ A)$ in terms of the subdifferentials of the data functions at nearby points (e.g., [14], [18], and [25]).

THEOREM 13. *Let us consider two convex functions $f : Y \to \overline{\mathbb{R}}$ and $g : X \to \overline{\mathbb{R}}$, where X and Y are (separated) real locally convex spaces, and a continuous affine mapping $A : X \to Y$, i.e., $Ax = A_0x + b$, where A_0 is the linear part of A and $b \in Y$. Assume that the following holds (when it makes sense):*

$$\operatorname{cl}(g + f \circ A) = (\operatorname{cl} g) + (\operatorname{cl} f) \circ A.$$

Then, for every $z \in X$, we have that

$$\partial(g + f \circ A)(z) = \bigcap_{\varepsilon>0} \operatorname{cl}\left(\partial_\varepsilon g(z) + A_0^*\partial_\varepsilon f(Az)\right),$$

where A_0^* is the adjoint operator of A_0.

Proof. Let us set $\varphi := g + f \circ A$, and $\psi := (\operatorname{cl} g) + (\operatorname{cl} f) \circ A$. The inclusion " $\supset$ " always holds, and consequently, it suffices to establish the opposite one when $\partial\varphi(z) \neq \emptyset$. In such a case, by (11) and the current assumption, we have

$$(\operatorname{cl} g)(z) + (\operatorname{cl} f)(Az) = (\operatorname{cl}\varphi)(z) = \varphi(z) = g(z) + f(Az) \in \mathbb{R},$$

and

$$\partial\varphi(z) = \partial(\operatorname{cl}\varphi)(z) = \partial((\operatorname{cl} g) + (\operatorname{cl} f) \circ A)(z) = \partial\psi(z). \tag{27}$$

Hence, $(\operatorname{cl} g)(z) = g(z) \in \mathbb{R}$ and $(\operatorname{cl} f)(Az) = f(Az) \in \mathbb{R}$, and so $\operatorname{cl} f \in \Gamma(Y)$ and $\operatorname{cl} g \in \Gamma(X)$. Furthermore, according to (10), for every $\varepsilon \geq 0$, one has $\partial_\varepsilon(\operatorname{cl} g)(z) = \partial_\varepsilon g(z)$ and $\partial_\varepsilon(\operatorname{cl} f)(Az) = \partial_\varepsilon f(Az)$.

Now, by the Legendre–Fenchel linearization of $\operatorname{cl} f$, we write that for every $x \in X$,

$$\begin{aligned}\psi(x) &= (\operatorname{cl} g)(x) + (\operatorname{cl} f)(Ax)\\ &= (\operatorname{cl} g)(x) + \sup\{\langle y^*, Ax\rangle - f^*(y^*) \mid y^* \in \operatorname{dom} f^*\}\\ &= \sup\{(\operatorname{cl} g)(x) + \langle A_0^* y^*, x\rangle + \langle y^*, b\rangle - f^*(y^*) \mid y^* \in \operatorname{dom} f^*\}.\end{aligned}$$

So, applying Theorem 4 (with $\alpha = 1$) together with Corollary 9(i),

$$\partial\psi(z) = \bigcap_{L\in\mathcal{F}_z,\varepsilon>0} \operatorname{cl}\left(\operatorname{co}\left(\bigcup_{y^*\in T_\varepsilon(z)} \left(\partial_\varepsilon(\operatorname{cl} g)(z) + A_0^* y^*\right)\right) + \mathrm{N}_{L\cap\operatorname{dom}\psi}(z)\right),$$

where, by (12),

$$\begin{aligned}T_\varepsilon(z) &= \{y^* \in Y^* \mid (\operatorname{cl} g)(z) + \langle A_0^* y^*, z\rangle + \langle y^*, b\rangle - f^*(y^*) \geq \psi(z) - \varepsilon\}\\ &= \{y^* \in Y^* \mid (\operatorname{cl} f)(Az) + f^*(y^*) \leq \langle y^*, Az\rangle + \varepsilon\} = \partial_\varepsilon(\operatorname{cl} f)(Az).\end{aligned}$$

Hence

$$\partial\psi(z) = \bigcap_{L\in\mathcal{F}_z,\varepsilon>0} \operatorname{cl}\left(\partial_\varepsilon(\operatorname{cl} g)(z) + A_0^*\partial_\varepsilon f(Az) + \mathrm{N}_{L\cap\operatorname{dom}\psi}(z)\right).$$

Now let $V \in \mathcal{V}$ (that is, V is a convex θ–neighborhood in X^*), and let $L \in \mathcal{F}_z$ be such that $L^\perp \subset V$. Then, for every $\varepsilon > 0$, from Lemma 11(i) we get

$$\mathrm{N}_{L\cap\operatorname{dom}\psi}(z) = \left[\operatorname{cl}\left(\partial_\varepsilon(\operatorname{cl} g)(z) + A_0^*\partial_\varepsilon(\operatorname{cl} f)(Az) + L^\perp\right)\right]_\infty,$$

so that, by taking into account (3), (27) leads us to

$$\begin{aligned}\partial\varphi(z) = \partial\psi(z) &\subset \operatorname{cl}\left(\operatorname{cl}\left(\partial_\varepsilon(\operatorname{cl} g)(z) + A_0^*\partial_\varepsilon(\operatorname{cl} f)(Az) + L^\perp\right)\right.\\ &\qquad \left. + \left[\operatorname{cl}\left(\partial_\varepsilon(\operatorname{cl} g)(z) + A_0^*\partial_\varepsilon(\operatorname{cl} f)(Az) + L^\perp\right)\right]_\infty\right)\\ &= \operatorname{cl}\left(\partial_\varepsilon(\operatorname{cl} g)(z) + A_0^*\partial_\varepsilon(\operatorname{cl} f)(Az) + L^\perp\right)\\ &\subset \partial_\varepsilon(\operatorname{cl} g)(z) + A_0^*\partial_\varepsilon(\operatorname{cl} f)(Az) + V\\ &= \partial_\varepsilon g(z) + A_0^*\partial_\varepsilon f(Az) + V,\end{aligned}$$

and consequently,

$$\partial\varphi(z) \subset \textstyle\bigcap_{\varepsilon>0}\bigcap_{V\in\mathcal{V}}\left(\partial_\varepsilon g(z) + A_0^*\partial_\varepsilon f(Az) + V\right) = \bigcap_{\varepsilon>0}\operatorname{cl}\left(\partial_\varepsilon g(z) + A_0^*\partial_\varepsilon f(Az)\right).$$

The proof is complete. □

Taking f and g to be lsc in Theorem 13 we obtain the following result of Hiriart-Urruty–Phelps [9].

COROLLARY 14. *Let f, g, and A be as in Theorem* 13. *If f and g are, in addition, lsc, then for every $z \in X$, we have that*

$$\partial(g + f\circ A)(z) = \bigcap_{\varepsilon>0} \operatorname{cl}\left(\partial_\varepsilon g(z) + A_0^*\partial_\varepsilon f(Az)\right).$$

In Corollary 16 below we derive the well-known Moreau–Rockafellar's formula on the sum (e.g., [19], p. 47). But, first, we need the following lemma, which gives us information about the closure of convex functions. Its proof does not appeal to the framework of Fenchel duality.

LEMMA 15. *Let $f : Y \to \overline{\mathbb{R}}$ and $g : X \to \overline{\mathbb{R}}$ be convex functions, and $A : X \to Y$ be a continuous affine mapping. Assume that f is finite and continuous at Ax_0 for some $x_0 \in (\operatorname{dom} g) \cap A^{-1}(\operatorname{dom} f)$. Then*

$$\operatorname{cl}(f \circ A + g) = (\operatorname{cl} f) \circ A + (\operatorname{cl} g).$$

Proof. Because $\operatorname{cl} f \leq f$, $\operatorname{cl} g \leq g$, and $(\operatorname{cl} f) \circ A + (\operatorname{cl} g)$ is lsc, one has $(\operatorname{cl} f) \circ A + (\operatorname{cl} g) \leq \operatorname{cl}(f \circ A + g)$. Moreover, in our hypothesis f and $\operatorname{cl} f$ are proper. To establish the converse inequality it suffices to take

$$x \in (\operatorname{dom}(\operatorname{cl} g)) \cap A^{-1}(\operatorname{dom}(\operatorname{cl} f)) \subset (\operatorname{dom}(\operatorname{cl} g)) \cap A^{-1}(\operatorname{cl}(\operatorname{dom} f))$$

such that $(\operatorname{cl}(f \circ A + g))(x) > -\infty$.

Let us fix $\lambda \in]0, 1[$ and set $x_\lambda := \lambda x_0 + (1-\lambda)x$ $[\in (\operatorname{dom}(\operatorname{cl} g)) \cap A^{-1}(\operatorname{cl}(\operatorname{dom} f))]$. Since $Ax_0 \in \operatorname{int}(\operatorname{dom} f)$ and $Ax \in \operatorname{cl}(\operatorname{dom} f)$, (4) yields

$$Ax_\lambda = A(\lambda x_0 + (1-\lambda)x) = \lambda Ax_0 + (1-\lambda)Ax \in \operatorname{int}(\operatorname{dom} f),$$

and so f is continuous at Ax_λ. Now let $(x_i)_{i\in I} \subset X$ be a net which converges to x and satisfies $(\operatorname{cl} g)(x_\lambda) = \lim_i g(\lambda x_0 + (1-\lambda)x_i)$. Since $\lim_i f(\lambda Ax_0 + (1-\lambda)Ax_i) = f(Ax_\lambda) = (\operatorname{cl} f)(Ax_\lambda)$, we obtain that

$$\begin{aligned}(\operatorname{cl}(f \circ A + g))(x_\lambda) &\leq \liminf_i \big(f(\lambda Ax_0 + (1-\lambda)Ax_i) + g(\lambda x_0 + (1-\lambda)x_i)\big)\\ &= (\operatorname{cl} f)(\lambda Ax_0 + (1-\lambda)Ax) + (\operatorname{cl} g)(x_\lambda)\\ &\leq \lambda\big((\operatorname{cl} f)(Ax_0) + (\operatorname{cl} g)(x_0)\big) + (1-\lambda)\big((\operatorname{cl} f)(Ax) + (\operatorname{cl} g)(x)\big).\end{aligned}$$

Whence, as $\lambda \downarrow 0$ we get

$$\liminf_{\lambda\to 0}(\operatorname{cl}(f \circ A + g))(x_\lambda) \leq (\operatorname{cl} f)(Ax) + (\operatorname{cl} g)(x),$$

and so $(\operatorname{cl}(f \circ A + g))(x) \leq (\operatorname{cl} f)(Ax) + (\operatorname{cl} g)(x)$. The proof is complete. ∎

COROLLARY 16. *Let $f : Y \to \overline{\mathbb{R}}$ and $g : X \to \overline{\mathbb{R}}$ be convex functions, and $A : X \to Y$ be a continuous affine mapping with linear part A_0. Assume that f is finite and continuous at Ax_0 for some $x_0 \in (\operatorname{dom} g) \cap A^{-1}(\operatorname{dom} f)$. Then, for every $z \in X$, we have that*

$$\partial(f \circ A + g)(z) = A_0^* \partial f(Az) + \partial g(z).$$

Proof. It is enough to show that $\partial(f \circ A + g)(z) \subset A_0^* \partial f(Az) + \partial g(z)$. Taking into account Theorem 13 and Lemma 15, it suffices to prove that

$$\bigcap_{\varepsilon>0} \operatorname{cl}\big(A_0^* \partial_\varepsilon f(Az) + \partial_\varepsilon g(z)\big) \subset A_0^* \partial f(Az) + \partial g(z) \tag{28}$$

for the nontrivial case $\partial(g + f \circ A)(z) \neq \emptyset$; hence $z \in (\operatorname{dom} g) \cap A^{-1}(\operatorname{dom} f)$ and $g(z), f(Az) \in \mathbb{R}$.

Indeed, for x^* in the set from the left-hand side of (28) and for each $r = 1, 2, \ldots$, there are nets $(v_i^*)_{i\in I} \subset \partial_{1/r} f(Az)$ and $(u_i^*)_{i\in I} \subset \partial_{1/r} g(z)$ such that $u_i^* + A_0^* v_i^* \to x^*$; thus we may assume that, for every $i \in I$,

$$\langle u_i^* + A_0^* v_i^*, z - x_0\rangle \leq \langle x^*, z - x_0\rangle + 1.$$

Since $u_i^* \in \partial_{1/r} g(z)$ and $r \geq 1$, this implies that

$$\langle v_i^*, Az - Ax_0 \rangle \leq \langle u_i^*, x_0 - z \rangle + \langle x^*, z - x_0 \rangle + 1 \leq g(x_0) - g(z) + \langle x^*, z - x_0 \rangle + 2.$$

Because f is continuous at Ax_0, there exists a symmetric θ–neighborhood $U \subset Y$ such that $\sup_{y \in U} f(y + Ax_0) \leq f(Ax_0) + 1$. Hence, for all $y \in U$,

$$\begin{aligned} \langle v_i^*, y \rangle &= \langle v_i^*, Az - Ax_0 \rangle + \langle v_i^*, y + Ax_0 - Az \rangle \\ &\leq \langle v_i^*, Az - Ax_0 \rangle + f(y + Ax_0) - f(Az) + 1 \\ &\leq g(x_0) - g(z) + \langle x^*, z - x_0 \rangle + f(Ax_0) - f(Az) + 4 \leq \mu \end{aligned}$$

for some $\mu > 0$. This shows that $\inf\{\langle v_i^*, y \rangle \mid y \in U\} \geq -\mu$, and so $(v_i^*)_{i \in I} \subset (\mu^{-1} U)^\circ$. Hence, by Alaoglu–Bourbaki's Theorem we may suppose w.l.o.g. that $(v_i^*)_{i \in I}$ and $(u_i^*)_{i \in I}$ w^*–converge to some $v_r^* \in \partial_{1/r} f(Az) \cap (\mu^{-1} U)^\circ$ and $u_r^* \in \partial_{1/r} g(z)$, respectively, and so $x^* = u_r^* + A_0^* v_r^*$. By the same argument we may suppose that $(v_r^*)_r$ and $(u_r^*)_r$ also w^*–converge to some $v^* \in \partial f(Az)$ and $u^* \in \partial g(z)$ and $x^* = u^* + A_0^* v^* \in \partial g(z) + A_0^* \partial f(Az)$. The proof is complete. □

Concluding remarks. (1) The preceding proof still works under more general regularity conditions, as those studied in Theorem 2.8.3 of [32].

(2) It should be noted that Lemma 5 can be easily deduced from Corollary 2.6.3 of [32], which is itself an extension of Corollary 14.

(3) Our main result in section 3 gives the formula for the subdifferential of the pointwise supremum $f := \sup_{t \in T} f_t$ of an arbitrary family of convex functions $f_t : X \to \overline{\mathbb{R}}$, $t \in T$. An important special case, which commonly appears in applications, corresponds to the so-called continuous model (e.g., [13], [24], and [32, Thm. 2.4.18]); see also [6]. There, the index set T is a (separated) compact space, and the parametrized mappings $t \to f_t(x)$ are upper semicontinuous for every $x \in X$. Such a situation is intermediate between the finite ([29]) and the general cases, and it is approached in a forthcoming paper.

(4) For further examples (in $\mathbb{R}^n$) in relation with our formula given in Theorem 4, the reader is addressed to references [6] and [7].

REFERENCES

[1] A. Auslender and M. Teboulle, *Asymptotic Cones and Functions in Optimization and Variational Inequalities*, Springer, New York, 2003.

[2] A. Brøndsted, *On the subdifferential of the supremum of two convex functions*, Math. Scand., 31 (1972), pp. 225–230.

[3] J. M. Danskin, *The Theory of Max-Min and its Applications to Weapons Allocations Problems*, Springer, New York, 1967.

[4] M. A. Goberna and M. A. López, *Optimal value function in semi-infinite programming*, J. Optim. Theory Appl., 59 (1988), pp. 261–279.

[5] M. A. Goberna and M. A. López, *Linear Semi-Infinite Optimization*, John Wiley and Sons, Chichester, UK, 1998.

[6] A. Hantoute, *Subdifferential set of the supremum of lower semicontinuous convex functions and the conical hull property*, Top, 14 (2006), pp. 355–374.

[7] A. Hantoute and M. A. López, *A complete characterization of the subdifferential set of the supremum of an arbitrary family of convex functions*, J. Convex Anal., 15 (2008), to appear.

[8] A. Hantoute and M. A. López, *Characterization of total ill-posedness in semi-infinite linear optimization*, J. Comput. Appl. Math., 217 (2008), pp. 350–364.

[9] J.-B. Hiriart-Urruty and R. R. Phelps, *Subdifferential calculus using ε-subdifferentials*, J. Funct. Anal., 118 (1993), pp. 154–166.

[10] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* I, II, Springer, Berlin, 1993.
[11] J.-B. Hiriart-Urruty, M. Moussaoui, A. Seeger, and M. Volle, *Subdifferential calculus without qualification conditions, using approximate subdifferentials: A survey*, Nonlinear Anal., 24 (1995), pp. 1727–1754.
[12] A. D. Ioffe and V. L. Levin, *Subdifferentials of convex functions*, Trudy Mos. Mat. Obs., 26 (1972), pp. 3–73 (Russian).
[13] A. D. Ioffe and V. H. Tikhomirov, *Theory of Extremal Problems*, in Stud. Math. Appl. 6, North-Holland, Amsterdam, 1979.
[14] F. Jules and M. Lassonde, *Formulas for subdifferentials of sums of convex functions*, J. Convex Anal., 9 (2002), pp. 519–533.
[15] V. L. Levin, *An application of Helly's theorem in convex programming, problems of best approximation and related questions*, Mat. Sb., 79(121) (1969), pp. 250–263. Math. USSR, Sb. 8 (1969), pp. 235–247 (in English).
[16] D. T. Luc, *Recession cones and the domination property in vector optimization*, Math. Program., 49 (1990), pp. 113–122.
[17] J.-J. Moreau, *Fonctionnelles Convexes*, Instituto Poligrafico e Zecca dello Stato, Rome, Italy, 2003.
[18] J.-P. Penot, *Subdifferential calculus without qualification assumptions*, J. Convex Anal., 3 (1996), pp. 207–219.
[19] R. R. Phelps, *Convex Functions, Monotone Operators and Differentiability*, 2nd ed., Lecture Notes in Math. 1364, Springer, Berlin, 1993.
[20] B. N. Pschenichnyi, *Convex programming in a normalized space*, Kibern., 5 (1965), pp. 46–54 (in Russian); translated as Cybern., 1 (1965), pp. 46–57.
[21] R. T. Rockafellar, *Directionally Lipschitzian functions and subdifferential calculus*, Proc. London Math. Soc., 39 (1979), pp. 331–355.
[22] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.
[23] R. T. Rockafellar, *Conjugate Duality and Optimization*, in CBMS Reg. Conf. Ser. Appl. Math. 16, SIAM, Philadelphia, 1974.
[24] V. N. Solov'ev, *The subdifferential and the directional derivatives of the maximum of a family of convex functions*, Izv. Ross Akad. Nauk Ser. Mat., 65 (2001), pp. 107–132.
[25] L. Thibault, *Sequential convex subdifferential calculus and sequential Lagrange multipliers*, SIAM J. Control Optim., 35 (1997), pp. 1434–1444.
[26] V. M. Tikhomirov, *Analysis* II*, Convex Analysis and Approximation Theory*, Encyclopedia Math. Sci. 14, R. V. Gamkrelidze, ed., Springer, New York, 1987.
[27] M. Valadier, *Sous-différentiels d'une borne supérieure et d'une somme continue de fonctions convexes*, C. R. Acad. Sci. Paris Sér. A-B Math., 268 (1969), pp. 39–42.
[28] M. Volle, *Sous-différentiel d'une enveloppe supérieure de fonctions convexes*, C. R. Acad. Sci. Paris Sér. I Math., 317 (1993), pp. 845–849.
[29] M. Volle, *On the subdifferential of an upper envelope of convex functions*, Acta Math. Vietnam., 19 (1994), pp. 137–148.
[30] C. Zălinescu, *Stability for a class of nonlinear optimization problems and applications*, in Nonsmooth Optimization and Related Topics, Ettore Majorana Internat. Sci. Ser. Phys. Sci., 43, Plenum, New York, 1989, pp. 437–458.
[31] C. Zălinescu, *On several results about convex set functions*, J. Math. Anal. Appl., 328 (2007), pp. 1451–1470.
[32] C. Zălinescu, *Convex Analysis in General Vector Spaces*, World Scientific, Singapore, 2002.

SIAM J. OPTIM.
Vol. 19, No. 2, pp. 883–910

A CLASS OF INTERIOR PROXIMAL-LIKE ALGORITHMS FOR CONVEX SECOND-ORDER CONE PROGRAMMING*

SHAOHUA PAN† AND JEIN-SHAN CHEN‡

Abstract. We propose a class of interior proximal-like algorithms for the second-order cone program, which is to minimize a closed proper convex function subject to general second-order cone constraints. The class of methods uses a distance measure generated by a twice continuously differentiable strictly convex function on $(0, +\infty)$, and includes as a special case the entropy-like proximal algorithm [Eggermont, *Linear Algebra Appl.*, 130 (1990), pp. 25–42], which was originally proposed for minimizing a convex function subject to nonnegative constraints. Particularly, we consider an approximate version of these methods, allowing the inexact solution of subproblems. Like the entropy-like proximal algorithm for convex programming with nonnegative constraints, we, under some mild assumptions, establish the global convergence expressed in terms of the objective values for the proposed algorithm, and we show that the sequence generated is bounded, and every accumulation point is a solution of the considered problem. Preliminary numerical results are reported for two approximate entropy-like proximal algorithms, and numerical comparisons are also made with the merit function approach [Chen and Tseng, *Math. Program.*, 104 (2005), pp. 293–327], which verify the effectiveness of the proposed method.

Key words. proximal method, measure of distance, second-order cone, second-order cone-convexity

AMS subject classifications. 65K05, 90C30

DOI. 10.1137/070685683

1. Introduction.

We consider the following convex second-order cone programming (CSOCP):

$$\begin{array}{l} \min\ f(\zeta) \\ \text{subject to (s.t.)}\ \ A\zeta + b \succeq_{\mathcal{K}} 0, \end{array} \tag{1}$$

where $f : \mathbb{R}^m \to (-\infty, +\infty]$ is a closed proper convex function; A is an $n \times m$ matrix, with $n \geq m$; b is a vector in $\mathbb{R}^n$; $x \succeq_{\mathcal{K}} 0$ means $x \in \mathcal{K}$; and $\mathcal{K}$ is the Cartesian product of second-order cones (SOCs), also called Lorentz cones [14]. In other words,

$$\mathcal{K} = \mathcal{K}^{n_1} \times \mathcal{K}^{n_2} \times \cdots \times \mathcal{K}^{n_N}, \tag{2}$$

where $N, n_1, \ldots, n_N \geq 1$, $n_1 + n_2 + \cdots + n_N = n$, and

$$\mathcal{K}^{n_i} := \left\{(x_1, x_2) \in \mathbb{R} \times \mathbb{R}^{n_i - 1} \mid x_1 \geq \|x_2\|\right\},$$

with $\|\cdot\|$ denoting the Euclidean norm and $\mathcal{K}^1$ denoting the set of nonnegative reals $\mathbb{R}_+$. The CSOCP, as an extension of the standard second-order cone programming, has a wide range of applications from engineering, control, and finance to robust optimization and combinatorial optimization; see [1, 21, 23] and references therein.

*Received by the editors March 19, 2007; accepted for publication (in revised form) April 25, 2008; published electronically August 13, 2008.

http://www.siam.org/journals/siopt/19-2/68568.html

†School of Mathematical Sciences, South China University of Technology Guangzhou 510640, China (shhpan@cut.edu.cn). This author's work is partially supported by the Doctoral Starting-up Foundation (B13B6050640) of GuangDong Province.

‡Department of Mathematics, National Taiwan Normal University Taipei 11677, Taiwan (jschen@math.ntnu.edu.tw). Member of Mathematics Division, National Center for Theoretical Sciences, Taipei Office. This author's work is partially supported by the National Science Council of Taiwan.

Recently, the second-order cone programming (SOCP) and the SOC complementarity problem have received much attention in optimization. There exist many methods for solving the CSOCP, including the smoothing methods [10, 15], the smoothing-regularization method [17], the semismooth Newton method [22], and the merit function approach [8]. All of these methods are proposed by using some SOC complementarity function or merit function to reformulate the KKT optimality conditions of the CSOCP as a nonsmooth (or smoothing) system of equations or an unconstrained minimization problem. Notice that the CSOCP is a typical convex programming problem which has extensive applications. But, to the best of our knowledge, there are few convex programming methods developed for (or extended to) the CSOCP except the interior point method [33]. Hence, it is worthy to explore other types of convex programming methods for the CSOCP which are different from the aforementioned methods.

One such method is the proximal point algorithm for minimizing a convex function $f(\zeta)$ over $\mathbb{R}^m$, which generates a sequence $\{\zeta^k\}$ by the following iterative scheme:

$$\zeta^k = \operatorname*{argmin}_{\zeta\in\mathbb{R}^m}\left\{f(\zeta) + \frac{1}{2\mu_k}\|\zeta - \zeta^{k-1}\|^2\right\}, \tag{3}$$

where μ_k is a sequence of positive numbers. The method was originally introduced by Martinet [24] with the Moreau proximal approximation of f (see [25]), and then further developed by Rockafellar [30, 31]. Later, some researchers [5, 13, 32] proposed and studied nonquadratic proximal point algorithms by replacing the quadratic distance in (3) with a Bregman distance or an entropy-like distance.

The entropy-like proximal algorithm was designed for minimizing a convex function $f(\zeta)$ subject to nonnegative constraints $\zeta \geq 0$. In [12], Eggermont first introduced the Kullback–Leibler relative entropy, defined by

$$^1d(\zeta,\xi) = \sum_{i=1}^m \zeta_i \ln(\zeta_i/\xi_i) + \zeta_i - \xi_i \quad \forall \zeta \geq 0,\ \xi > 0,$$

and established the following entropy-like proximal point algorithm:

$$\begin{cases} \zeta^0 > 0, \\ \zeta^k = \operatorname*{argmin}_{\zeta>0}\left\{f(\zeta) + {\mu_k}^{-1} d(\zeta^{k-1},\zeta)\right\}. \end{cases} \tag{4}$$

Later, Teboulle [32] proposed to replace the usual Kullback–Leibler relative entropy with a new type of distance-like function, called φ-divergence, to define the entropy-like proximal map. Let $\varphi : \mathbb{R} \to (-\infty, +\infty]$ be a closed proper convex function satisfying certain conditions (see [18, 32]). The φ-divergence induced by φ is defined as

$$d_\varphi(\zeta,\xi) := \sum_{i=1}^m \xi_i \varphi(\zeta_i/\xi_i). \tag{5}$$

Based on the φ-divergence, Isume et al. [18, 19] generalized Eggermont's algorithm as

$$\begin{cases} \zeta^0 > 0, \\ \zeta^k = \operatorname*{argmin}_{\zeta>0}\left\{f(\zeta) + {\mu_k}^{-1} d_\varphi(\zeta,\zeta^{k-1})\right\}, \end{cases} \tag{6}$$

[1] The convention of $0 \ln 0 = 0$ is used throughout this paper.

and they obtained the convergence theorems under weaker assumptions. Clearly, when

$$\varphi(t) = -\ln t + t - 1 \quad (t > 0),$$

we have that $d_\varphi(\zeta, \xi) = d(\xi, \zeta)$, and consequently the algorithm reduces to Eggermont's.

Observing that the proximal-like algorithm (6) associated with $\varphi(t) = -\ln t + t - 1$ inherits the features of the interior point method as well as the proximal point method, Auslender [2] extended the algorithm to general linearly constrained convex minimization problems and variational inequalities on polyhedra. Then, is it possible to extend the algorithm to nonpolyhedra symmetric conic optimization problems and establish the corresponding convergence results? In this paper, we will explore its extension to the setting of SOCs and establish a class of interior proximal-like algorithms for the CSOCP. We should mention that the algorithm (6) with the entropy function $t \ln t - t + 1$ $(t \geq 0)$ was recently extended to convex semidefinite programming [11].

For simplicity, in the rest of this paper, we focus on the case where $\mathcal{K} = \mathcal{K}^n$. All of the analysis can be carried over to the general case where $\mathcal{K}$ has the direct product structure as (2). It is known that $\mathcal{K}^n$ is a closed convex cone with the interior given by

$$\operatorname{int}(\mathcal{K}^n) := \left\{(x_1, x_2) \in \mathbb{R} \times \mathbb{R}^{n-1} \mid x_1 > \|x_2\|\right\}.$$

For any x, y in $\mathbb{R}^n$, we write $x \succeq_{\mathcal{K}^n} y$ if $x - y \in \mathcal{K}^n$; and write $x \succ_{\mathcal{K}^n} y$ if $x - y \in \operatorname{int}(\mathcal{K}^n)$. In other words, we have that $x \succeq_{\mathcal{K}^n} 0$ if and only if $x \in \mathcal{K}^n$ and $x \succ_{\mathcal{K}^n} 0$ if and only if $x \in \operatorname{int}(\mathcal{K}^n)$. We denote $\mathcal{F}$ by the constraint set of the CSOCP, i.e.,

$$\mathcal{F} := \left\{\zeta \in \mathbb{R}^m \mid A\zeta + b \succeq_{\mathcal{K}^n} 0\right\}. \tag{7}$$

It is not difficult to verify that $\mathcal{F}$ is convex, and its interior $\operatorname{int}(\mathcal{F})$ is given by

$$\operatorname{int}(\mathcal{F}) := \left\{\zeta \in \mathbb{R}^m \mid A\zeta + b \succ_{\mathcal{K}^n} 0\right\}.$$

The proximal-like algorithm that we propose for the CSOCP is defined as follows:

$$\begin{cases} \zeta^0 \in \operatorname{int}(\mathcal{F}), \\ \zeta^k = \underset{\zeta \in \operatorname{int}(\mathcal{F})}{\operatorname{argmin}} \left\{ f(\zeta) + \mu_k^{-1} D(A\zeta + b, A\zeta^{k-1} + b) \right\}, \end{cases} \tag{8}$$

where $D : \mathbb{R}^n \times \mathbb{R}^n \to (-\infty, +\infty]$ is a closed proper convex function generated by a class of twice continuously differentiable strictly convex functions on $(0, +\infty)$, and the specific expression is given in section 3. The class of distance measures, as will be shown in section 3, includes as a special case the natural extension of $d_\varphi(x, y)$, with $\varphi(t) = -\ln t + t - 1$ to the SOCs. For the proximal-like algorithm (8), we particularly consider an approximate version which allows an inexact minimization of the subproblem (8) and establish its global convergence results under some mild assumptions. Numerical results are reported for two approximate entropy-like proximal algorithms, which verify the effectiveness of the proximal method proposed. In addition, numerical comparisons with the merit function approach [8] indicate that the condition number of the Hessian matrix $\nabla^2 f(\zeta)$ has a great influence on the numerical performance of the proximal-like algorithm and the merit function approach, but the

former seems to have no direct relation with the dense degree of test problems, but the latter tends to more function evaluations as the density increases.

The outline of this paper is as follows. In section 2, we review some basic concepts and properties associated with SOCs. In section 3, we state the definition of $D(x,y)$ and present some specific examples. Some favorable properties of $D(x,y)$ are investigated in section 4. In section 5, we describe an approximate proximal-like algorithm allowing inexact minimization in (8) and establish the global convergence of the algorithm. In section 6, we report our numerical experiences for the proposed proximal-like algorithm by solving some convex SOCPs. Finally, we conclude this paper in section 7.

Throughout this paper, I represents an identity matrix of suitable dimension, and $\mathbb{R}^n$ denotes the space of n-dimensional real column vectors. For a differentiable function h on $\mathbb{R}$, we denote h', h'', and h''' by its first, second, and third derivative, respectively. Given a set S, we denote $\bar{S}, \mathrm{int}(S)$, and $\mathrm{bd}(S)$ by the closure, the interior and the boundary of S, respectively. Note that a function is closed if and only if it is lower semicontinuous, and a function is proper if $f(\zeta) < \infty$ for at least one $\zeta \in \mathbb{R}^m$ and $f(\zeta) > -\infty$ for all $\zeta \in \mathbb{R}^m$. For a closed proper convex function $f : \mathbb{R}^m \to (-\infty, +\infty]$, we denote its domain by $\mathrm{dom} f := \{\, \zeta \in \mathbb{R}^m \mid f(\zeta) < \infty\}$ and the subdifferential of f at $\widehat{\zeta}$ by

$$\partial f(\widehat{\zeta}) := \left\{ w \in \mathbb{R}^m \mid f(\zeta) \geq f(\widehat{\zeta}) + \langle w, \zeta - \widehat{\zeta} \rangle \ \forall \zeta \in \mathbb{R}^m \right\}.$$

If f is differentiable at ζ, the notation $\nabla f(\zeta)$ represents the gradient at ζ of f.

2. Preliminaries. This section recalls some basic concepts and preliminary results related to SOCs that will be used in the subsequent analysis. For any $x = (x_1, x_2) \in \mathbb{R} \times \mathbb{R}^{n-1}$ and $y = (y_1, y_2) \in \mathbb{R} \times \mathbb{R}^{n-1}$, we define their *Jordan product* as

$$x \circ y := (\langle x, y \rangle,\ y_1 x_2 + x_1 y_2). \tag{9}$$

We write x^2 to mean $x \circ x$ and write $x + y$ to mean the usual componentwise addition of vectors. Then $\circ$, $+$, and $e = (1, 0, \ldots, 0)^T \in \mathbb{R}^n$ have the following basic properties (see [14, 15]): (1) $e \circ x = x$ for all $x \in \mathbb{R}^n$. (2) $x \circ y = y \circ x$ for all $x, y \in \mathbb{R}^n$. (3) $x \circ (x^2 \circ y) = x^2 \circ (x \circ y)$ for all $x, y \in \mathbb{R}^n$. (4) $(x+y) \circ z = x \circ z + y \circ z$ for all $x, y, z \in \mathbb{R}^n$. The Jordan product is not associative. For example, for $n = 3$, let $x = (1, -1, 1)$ and $y = z = (1, 0, 1)$, then we have that $(x \circ y) \circ z = (4, -1, 4) \neq x \circ (y \circ z) = (4, -2, 4)$. However, it is power associated, i.e., $x \circ (x \circ x) = (x \circ x) \circ x$ for all $x \in \mathbb{R}^n$. Thus, we may, without fear of ambiguity, write x^m for the product of m copies of x and $x^{m+n} = x^m \circ x^n$ for all positive integers m and n. We stipulate that $x^0 = e$. Besides, $\mathcal{K}^n$ is not closed under Jordan product. For example, $x = (1, 1, 0), y = (2, -1, 3) \in \mathcal{K}^n$, but $x \circ y = (1, 1, 3) \notin \mathcal{K}^n$.

For each $x = (x_1, x_2) \in \mathbb{R} \times \mathbb{R}^{n-1}$, the *determinant* and the *trace* of x are defined by

$$\det(x) = x_1^2 - \|x_2\|^2, \quad \mathrm{tr}(x) = 2x_1. \tag{10}$$

In general, $\det(x \circ y) \neq \det(x)\det(y)$ unless $x_2 = \alpha y_2$ for some $\alpha \in \mathbb{R}$. A vector $x = (x_1, x_2) \in \mathbb{R} \times \mathbb{R}^{n-1}$ is said to be *invertible* if $\det(x) \neq 0$. If x is invertible, then there exists a unique $y = (y_1, y_2) \in \mathbb{R} \times \mathbb{R}^{n-1}$ satisfying $x \circ y = y \circ x = e$. We call

this y the inverse of x and denote it by x^{-1}. In fact, we have that

$$x^{-1} = \frac{1}{x_1^2 - \|x_2\|^2}(x_1, -x_2) = \frac{1}{\det(x)}(\mathrm{tr}(x)e - x). \tag{11}$$

Hence, $x \in \mathrm{int}(\mathcal{K}^n)$ if and only if $x^{-1} \in \mathrm{int}(\mathcal{K}^n)$, and $(x^k)^{-1}$ is well-defined if $x \in \mathrm{int}(\mathcal{K}^n)$.

In the following, we recall from [15] that each $x = (x_1, x_2) \in \mathbb{R} \times \mathbb{R}^{n-1}$ admits a spectral factorization associated with $\mathcal{K}^n$ of the form

$$x = \lambda_1(x) \cdot u_x^{(1)} + \lambda_2(x) \cdot u_x^{(2)},$$

where $\lambda_i(x)$ and $u_x^{(i)}$ for $i = 1, 2$ are the spectral values and the associated spectral vectors of x, respectively, given by

$$\begin{aligned} \lambda_i(x) &= x_1 + (-1)^i \|x_2\|, \\ u_x^{(i)} &= \begin{cases} \frac{1}{2}\left(1,\ (-1)^i \frac{x_2}{\|x_2\|}\right) & \text{if } x_2 \neq 0; \\ \frac{1}{2}\left(1,\ (-1)^i \bar{x}_2\right) & \text{if } x_2 = 0, \end{cases} \end{aligned} \tag{12}$$

with $\bar{x}_2$ being any vector in $\mathbb{R}^{n-1}$ such that $\|\bar{x}_2\| = 1$. If $x_2 \neq 0$, then the factorization is unique. The spectral decomposition along with the Jordan algebra associated with SOC has some basic properties, whose proofs can be found in [14, 15]. Here, we list four of them that will often be used in the subsequent sections.

PROPERTY 2.1. *For any $x = (x_1, x_2) \in \mathbb{R} \times \mathbb{R}^{n-1}$ with the spectral values $\lambda_1(x), \lambda_2(x)$ and spectral vectors $u_x^{(1)}, u_x^{(2)}$ given as in* (12)*, the following results hold:*

(a) *$u_x^{(1)}$ and $u_x^{(2)}$ are orthogonal under Jordan product and have length $1/\sqrt{2}$, i.e.,*

$$u_x^{(1)} \circ u_x^{(2)} = 0, \quad \|u_x^{(1)}\| = \|u_x^{(2)}\| = 1/\sqrt{2}.$$

(b) *$u_x^{(1)}$ and $u_x^{(2)}$ are idempotent under Jordan product, i.e., $u_x^{(i)} \circ u_x^{(i)} = u_x^{(i)}$ for $i = 1, 2$.*

(c) *The determinant, the trace, and the Euclidean norm of x can be denoted by $\lambda_1(x), \lambda_2(x)$:*

$$\det(x) = \lambda_1(x)\lambda_2(x), \quad \mathrm{tr}(x) = \lambda_1(x) + \lambda_2(x), \quad \|x\|^2 = \frac{[\lambda_1(x)]^2 + [\lambda_2(x)]^2}{2}.$$

(d) *$\lambda_1(x)$ are nonnegative (positive) if and only if $x \in \mathcal{K}^n$ ($x \in \mathrm{int}(\mathcal{K}^n)$).*

LEMMA 2.1.

(a) *For any $x \in \mathbb{R}^n$, $x \succeq_{\mathcal{K}^n} 0 \Longleftrightarrow \langle x, y\rangle \geq 0$ for any $y \succeq_{\mathcal{K}^n} 0$.*

(b) *For any $x \in \mathbb{R}^n$, $x \succ_{\mathcal{K}^n} 0 \Longleftrightarrow \langle x, y\rangle > 0$ for any $y \succeq_{\mathcal{K}^n} 0$ and $y \neq 0$.*

(c) *For any $x, y \in \mathbb{R}^n$, let $\lambda_i(x)$ and $\lambda_i(y)$ for $i = 1, 2$ be their spectral values. Then,*

$$\lambda_1(x)\lambda_2(y) + \lambda_2(x)\lambda_1(y) \leq \mathrm{tr}(x \circ y) \leq \lambda_1(x)\lambda_1(y) + \lambda_2(x)\lambda_2(y).$$

Proof. Part (a) is direct by the self-duality of $\mathcal{K}^n$, and we next consider parts (b) and (c).

(b) Let $x = (x_1, x_2), y = (y_1, y_2) \in \mathbb{R} \times \mathbb{R}^{n-1}$. The necessity follows from

$$\langle x, y\rangle = x_1 y_1 + x_2^T y_2 \geq x_1 y_1 - \|x_2\|\|y_2\| \geq x_1 y_1 - y_1\|x_2\| = y_1(x_1 - \|x_2\|) > 0,$$

where the first inequality is by Cauchy–Schwartz, the second is due to $y \succeq_{\mathcal{K}^n} 0$, and the third is since $x \succ_{\mathcal{K}^n} 0$ and $y \neq 0,\ y \succeq_{\mathcal{K}^n} 0$. Next, we prove the sufficiency. First, from $\langle x, y\rangle > 0$ for any $y \succeq_{\mathcal{K}^n} 0$ and $y \neq 0$, we deduce that $x_1 > 0$ by setting $y = e$. If $x_2 = 0$, then the conclusion follows. If $x_2 \neq 0$, then we set $y = (1, -\frac{x_2}{\|x_2\|})$. Clearly, $y \succeq_{\mathcal{K}^n} 0,\ y \neq 0$, and $0 < \langle x, y\rangle = x_1 - \|x_2\| = \lambda_1(x)$. By Property 2.1 (d), we then have $x \succ_{\mathcal{K}^n} 0$.

(c) For any $x = (x_1, x_2), y = (y_1, y_2) \in \mathbb{R} \times \mathbb{R}^{n-1}$, by (12) we can compute that

$$\lambda_1(x)\lambda_2(y) + \lambda_2(x)\lambda_1(y) = 2x_1y_1 - 2\|x_2\|\|y_2\| \leq 2(x_1y_1 + x_2^Ty_2) = \mathrm{tr}(x \circ y),$$
$$\lambda_1(x)\lambda_1(y) + \lambda_2(x)\lambda_2(y) = 2x_1y_1 + 2\|x_2\|\|y_2\| \geq 2(x_1y_1 + x_2^Ty_2) = \mathrm{tr}(x \circ y)\ .$$

Combining with the two inequalities above then yields the desired result. □

For any $h : \mathbb{R} \to \mathbb{R}$, the following vector-valued function was considered in [6, 15]:

$$h^{\mathrm{soc}}(x) = h[\lambda_1(x)] \cdot u_x^{(1)} + h[\lambda_2(x)] \cdot u_x^{(2)} \quad \forall x = (x_1, x_2) \in \mathbb{R} \times \mathbb{R}^{n-1}. \tag{13}$$

If h is defined only on a subset of $\mathbb{R}$, then h^{soc} is defined on the corresponding subset of $\mathbb{R}^n$. The definition in (13) is unambiguous whether $x_2 \neq 0$ or $x_2 = 0$. For the vector-valued function h^{soc} induced by h, we have the following results.

LEMMA 2.2. *Given a function $h : \mathrm{I}_{\mathbb{R}} \to \mathbb{R}$, let $h^{\mathrm{soc}} : S \to \mathbb{R}^n$ be the vector-valued function induced by h as in* (13)*, where $\mathrm{I}_{\mathbb{R}} \subseteq \mathbb{R}$ and $S \subseteq \mathbb{R}^n$. Then, the following results hold:*

(a) *For any $x \in S$, $\lambda_i[h^{\mathrm{soc}}(x)] = h[\lambda_i(x)]$ for $i = 1, 2$ and $\mathrm{tr}[h^{\mathrm{soc}}(x)] = \sum_{i=1}^2 h[\lambda_i(x)]$.*

(b) *If h is continuously differentiable on $\mathrm{I}_{\mathbb{R}}$, then h^{soc} is continuously differentiable on the set S, and its transposed Jacobian at $x = (x_1, x_2) \in S$ is given by the formula*

$$\nabla h^{\mathrm{soc}}(x) = h'(x_1)I \tag{14}$$

if $x_2 = 0$, and otherwise

$$\nabla h^{\mathrm{soc}}(x) = \begin{bmatrix} b & c\dfrac{x_2^T}{\|x_2\|} \\ c\dfrac{x_2}{\|x_2\|} & aI + (b-a)\dfrac{x_2x_2^T}{\|x_2\|^2} \end{bmatrix}, \tag{15}$$

where

$$a = \frac{h[\lambda_2(x)] - h[\lambda_1(x)]}{\lambda_2(x) - \lambda_1(x)},\ b = \frac{h'[\lambda_2(x)] + h'[\lambda_1(x)]}{2},\ c = \frac{h'[\lambda_2(x)] - h'[\lambda_1(x)]}{2}.$$

(c) *If h is continuously differentiable on $\mathrm{I}_{\mathbb{R}}$, then $\mathrm{tr}[h^{\mathrm{soc}}(x)]$ is continuously differentiable on the set S, and its gradient $\nabla \mathrm{tr}[h^{\mathrm{soc}}(x)] = 2\nabla h^{\mathrm{soc}}(x) \cdot e = 2(h')^{\mathrm{soc}}(x)$.*

(d) *If h is (strictly) convex on $\mathrm{I}_{\mathbb{R}}$, then $\mathrm{tr}[h^{\mathrm{soc}}(x)]$ is (strictly) convex on the set S.*

Proof. (a) The proof is direct by the definition of h^{soc} and the spectral value.

(b) The conclusion follows directly from [15, Propostion 5.2] or [6, Proposition 4].

(c) Since $\mathrm{tr}[h^{\mathrm{soc}}(x)] = 2\langle h^{\mathrm{soc}}(x), e\rangle$, by part (b) $\mathrm{tr}[h^{\mathrm{soc}}(x)]$ is obviously continuously differentiable. Applying the chain rule for the inner product of two functions yields

$$\nabla \mathrm{tr}[h^{\mathrm{soc}}(x)] = 2\nabla h^{\mathrm{soc}}(x) \cdot e,$$

where $\nabla h^{\mathrm{soc}}(x)$ is given by (14)–(15). By a simple computation, it is easy to verify that

$$\nabla h^{\mathrm{soc}}(x) \cdot e = h'[\lambda_1(x)]u_x^{(1)} + h'[\lambda_2(x)]u_x^{(2)} = (h')^{\mathrm{soc}}(x).$$

Combining the last two equalities immediately gives the second part of the conclusions.

(d) The proof is similar to that of [26, Lemma 3.2 (d)], and so we omit it. □

To close this section, we review the definition of SOC-convexity and SOC-monotonicity. The two concepts, such as the matrix-convexity and the matrix-monotonicity in the semidefinite programming, play an important role in the solution methods of SOCPs.

DEFINITION 2.1 (see [7]). *Given a function $h : \mathrm{I}_{\mathbb{R}} \to \mathbb{R}$, let $h^{\mathrm{soc}} : S \to \mathbb{R}^n$ be the vector-valued function defined as in* (13)*, where $\mathrm{I}_{\mathbb{R}} \subseteq \mathbb{R}$ and $S \subseteq \mathbb{R}^n$. Then,*

(a) *h is said to be SOC-monotone of order n on $\mathrm{I}_{\mathbb{R}}$ if for any $x, y \in S$,*

$$x \succeq_{\mathcal{K}^n} y \Longrightarrow h^{\mathrm{soc}}(x) \succeq_{\mathcal{K}^n} h^{\mathrm{soc}}(y).$$

(b) *h is said to be SOC-convex of order n on $\mathrm{I}_{\mathbb{R}}$ if for any $x, y \in S$ and $0 \le \beta \le 1$,*

$$h^{\mathrm{soc}}\Big(\beta x + (1-\beta)y\Big) \preceq_{\mathcal{K}^n} \beta h^{\mathrm{soc}}(x) + (1-\beta)h^{\mathrm{soc}}(y). \tag{16}$$

We say that h is SOC-convex (respectively, SOC-monotone) on $\mathrm{I}_{\mathbb{R}}$ if h is SOC-convex of all order n (respectively, SOC-monotone of all order n) on $\mathrm{I}_{\mathbb{R}}$. A function h is said to be SOC-concave on $\mathrm{I}_{\mathbb{R}}$ whenever $-h$ is SOC-convex on $\mathrm{I}_{\mathbb{R}}$. When h is continuous on $\mathrm{I}_{\mathbb{R}}$, the condition in (16) can be replaced by the more special condition:

$$h^{\mathrm{soc}}\left(\frac{x+y}{2}\right) \preceq_{\mathcal{K}^n} \frac{1}{2}\Big(h^{\mathrm{soc}}(x) + h^{\mathrm{soc}}(y)\Big). \tag{17}$$

Obviously, the set of SOC-monotone functions and the set of SOC-convex functions are both closed under positive linear combinations and under pointwise limits.

3. Distance-like functions in SOCs. In this section, we present the definition of the distance-like function $D(x, y)$ involved in the proximal-like algorithm (8) and some specific examples. Let $\phi : \mathbb{R} \to (-\infty, +\infty]$ be a closed proper convex function with $\mathrm{dom}\phi = [0, +\infty)$ and assume that

(C.1) ϕ is strictly convex on its domain.

(C.2) ϕ is twice continuously differentiable on $\mathrm{int}(\mathrm{dom}\phi)$, with $\lim_{t\to 0^+} \phi''(t) = +\infty$.

(C.3) $\phi'(t)t - \phi(t)$ is convex on $\mathrm{int}(\mathrm{dom}\phi)$.

(C.4) ϕ' is SOC-concave on $\mathrm{int}(\mathrm{dom}\phi)$.

In what follows, we denote by Φ the class of functions satisfying Conditions C.1–C.4.

Given a $\phi \in \Phi$, let ϕ^{soc} and $(\phi')^{\mathrm{soc}}$ be the vector-valued function given as in (13). We define $D(x, y)$ involved in the proximal-like algorithm (8) by

$$D(x,y) := \begin{cases} \mathrm{tr}\Big[\phi^{\mathrm{soc}}(y) - \phi^{\mathrm{soc}}(x) - (\phi')^{\mathrm{soc}}(x) \circ (y-x)\Big] & \forall x \in \mathrm{int}(\mathcal{K}^n), y \in \mathcal{K}^n, \\ +\infty & \text{otherwise.} \end{cases} \tag{18}$$

The function, as will be shown in the next section, possesses some favorable properties. Particularly, $D(x, y) \ge 0$ for any $x, y \in \mathrm{int}(\mathcal{K}^n)$, and $D(x, y) = 0$ if and only if $x = y$. Hence, $D(x, y)$ can be used to measure the distance between the two points in $\mathrm{int}(\mathcal{K}^n)$.

In the following, we concentrate on the examples of the distance-like function $D(x,y)$. For this purpose, we first give another characterization for Condition C.3.

LEMMA 3.1. *Let $\phi : \mathbb{R} \to (-\infty, +\infty]$ be a closed proper function with $\mathrm{dom}\phi = [0, +\infty)$. If ϕ is thrice continuously differentiable on $\mathrm{int}(\mathrm{dom}\phi)$, then ϕ satisfies Condition C.3 if and only if its derivative function ϕ' is exponentially convex,*[2] *or*

$$\phi'(t_1 t_2) \le \frac{1}{2}\Big(\phi'(t_1^2) + \phi'(t_2^2)\Big) \quad \forall t_1, t_2 > 0. \tag{19}$$

Proof. Since the function ϕ is thrice continuously differentiable on $\mathrm{int}(\mathrm{dom}\phi)$, ϕ satisfies Condition C.3 if and only if

$$\phi''(t) + t\phi'''(t) \ge 0 \quad (\forall t > 0).$$

Observe that the inequality is also equivalent to

$$t\phi''(t) + t^2\phi'''(t) \ge 0 \quad (\forall t > 0),$$

and hence substituting by $t = \exp(\theta)$ for $\theta \in \mathbb{R}$ into the inequality yields that

$$\exp(\theta)\phi''(\exp(\theta)) + \exp(2\theta)\phi'''(\exp(\theta)) \ge 0 \quad \forall \theta \in \mathbb{R}.$$

Since the left-hand side of this inequality is exactly $[\phi'(\exp(\theta))]''$, it means that $\phi'(\exp(\cdot))$ is convex on $\mathbb{R}$. Consequently, the first part of the conclusions follows.

Note that the convexity of $\phi'(\exp(\cdot))$ on $\mathbb{R}$ is equivalent to saying, for any $\theta_1, \theta_2 \in \mathbb{R}$,

$$\phi'(\exp(r\theta_1 + (1-r)\theta_2)) \le r\phi'(\exp(\theta_1)) + (1-r)\phi'(\exp(\theta_2)), \quad r \in [0,1],$$

which, by letting $t_1 = \exp(\theta_1)$ and $t_2 = \exp(\theta_2)$, can be rewritten as

$$\phi'(t_1^r t_2^{1-r}) \le r\phi'(t_1) + (1-r)\phi'(t_2) \ \ \forall t_1, t_2 > 0 \text{ and } r \in [0,1].$$

This is clearly equivalent to the statement in (19) due to the continuity of ϕ'. □

Remark 3.1. The exponential convexity was also used in the definition of the *self-regular* function [27] in which the authors denote Ω by the set of functions whose elements are twice continuously differentiable and exponentially convex on $(0, +\infty)$. By Lemma 3.1, clearly, if $h \in \Omega$, then the function $\int_0^t h(\theta)d\theta$ necessarily satisfies Condition C.3. For example, $\ln t$ belongs to Ω, and so $\int_0^t \ln\theta d\theta = t\ln t$ satisfies Condition C.3.

For the characterizations of the SOC-concavity, interested readers may refer to [7, 9]. Here, we present a lemma which states that the composition of two SOC-concave functions is SOC-concave under some conditions. By this lemma, we may conveniently obtain some new SOC-concave functions from the existing ones.

LEMMA 3.2. *Let $g : \mathrm{J}_{\mathbb{R}} \to \mathbb{R}$ and $h : \mathrm{I}_{\mathbb{R}} \to \mathrm{J}_{\mathbb{R}}$, where $\mathrm{J}_{\mathbb{R}} \subseteq \mathbb{R}$ and $\mathrm{I}_{\mathbb{R}} \subseteq \mathbb{R}$. If g is SOC-concave and SOC-monotone on $\mathrm{J}_{\mathbb{R}}$ and h is SOC-concave on $\mathrm{I}_{\mathbb{R}}$, then their composition $g(h(\cdot))$ is also SOC-concave on $\mathrm{I}_{\mathbb{R}}$. If, in addition, h is SOC-monotone on $\mathrm{I}_{\mathbb{R}}$, then $g(h(\cdot))$ is also SOC-monotone on $\mathrm{I}_{\mathbb{R}}$.*

Proof. For the sake of notation, let $g^{\mathrm{soc}} : \widehat{S} \to \mathbb{R}^n$ and $h^{\mathrm{soc}} : S \to \widehat{S}$ be the vector-valued functions associated with g and h, respectively, where $S \subseteq \mathbb{R}^n$ and $\widehat{S} \subseteq \mathbb{R}^n$.

[2] Which means the function $\phi'(\exp(\cdot)) : \mathbb{R} \to \mathbb{R}$ is convex on $\mathbb{R}$,

Define $\widehat{g}(t) = g(h(t))$. Then, for any $x \in S$, it follows from (11) and (13) that

$$\begin{aligned} g^{\rm soc}(h^{\rm soc}(x)) &= g^{\rm soc}\Big[h(\lambda_1(x))u_x^{(1)} + h(\lambda_2(x))u_x^{(2)}\Big] \\ &= g[h(\lambda_1(x))]u_x^{(1)} + g[h(\lambda_2(x))]u_x^{(2)} \\ &= \widehat{g}^{\rm soc}(x). \end{aligned} \tag{20}$$

We next prove that $\widehat{g}(t)$ is SOC-concave on $\mathrm{I}_{\mathbb{R}}$. For any $x, y \in S$ and $0 \le \beta \le 1$, from the SOC-concavity of $h(t)$ it follows that

$$h^{\rm soc}(\beta x + (1-\beta)y) \succeq_{\mathcal{K}^n} \beta h^{\rm soc}(x) + (1-\beta)h^{\rm soc}(y).$$

Using the SOC-monotonicity and SOC-concavity of g, we then obtain that

$$\begin{aligned} g^{\rm soc}\Big[h^{\rm soc}(\beta x + (1-\beta)y)\Big] &\succeq_{\mathcal{K}^n} g^{\rm soc}\Big[\beta h^{\rm soc}(x) + (1-\beta)h^{\rm soc}(y)\Big] \\ &\succeq_{\mathcal{K}^n} \beta g^{\rm soc}[h^{\rm soc}(x)] + (1-\beta)g^{\rm soc}[h^{\rm soc}(y)]. \end{aligned}$$

This together with (20) implies that for any $x, y \in S$ and $0 \le \beta \le 1$,

$$\widehat{g}^{\rm soc}\Big(\beta x + (1-\beta)y\Big) \succeq_{\mathcal{K}^n} \beta\widehat{g}^{\rm soc}(x) + (1-\beta)\widehat{g}^{\rm soc}(y).$$

Consequently, the function $\widehat{g}(t)$, i.e., $g(h(\cdot))$ is SOC-concave on $\mathrm{I}_{\mathbb{R}}$. The second part of the conclusions is obvious. □

PROPOSITION 3.1. (a) *The function $h(t) = t^r$, with $0 \le r \le 1$ is both SOC-concave and SOC-monotone on $[0, +\infty)$.*

(b) *$h(t) = -t^{-r}$, with $0 \le r \le 1$ is SOC-concave and SOC-monotone on $(0, +\infty)$.*

(c) *For all $u \le 0$, $h(t) = \frac{1}{u-t}$ is SOC-concave as well as SOC-monotone on $(0, +\infty)$.*

(d) *The function $\ln t$ is SOC-concave and SOC-monotone on $(0, +\infty)$.*

Proof. (a) The proof has been given by [7, Proposition 3.7], and we here omit it.

(b) The conclusion follows directly from [9, Corollary 4.2].

(c) Let $g(t) = -1/t$ and $\widehat{h}(t) = t - u$. Then, $h(t) = 1/(u-t)$ is exactly the composition of the two functions, i.e., $h(t) = g(\widehat{h}(t))$. From part (b), $g(t)$ is SOC-monotone and SOC-concave on $(0, +\infty)$; whereas by [7, Proposition 3.1 (b)] $\widehat{h}(t)$ is SOC-monotone and SOC-concave on $(0, +\infty)$. Thus, applying Lemma 3.2, we readily obtain the conclusion.

(d) The proof can be found in [9]. In view of the importance of $\ln t$, we here present a different proof by following the same line as [3]. Noting that

$$\ln t = \int_{-\infty}^{0} \left[\frac{1}{u-t} - \frac{u}{u^2+1}\right] du \quad (t > 0),$$

we have for any $x \in \mathrm{int}(\mathcal{K}^n)$ that

$$\ln x = \int_{-\infty}^{0} \left[(ue - x)^{-1} - \frac{u}{u^2+1}e\right] du. \tag{21}$$

For any $x = (x_1, x_2),\ y = (y_1, y_2) \in \mathrm{int}(\mathcal{K}^n)$ and any $0 \le \beta \le 1$, let

$$w = \ln(\beta x + (1-\beta)y) - \beta \ln x - (1-\beta)\ln y.$$

Then, by the definition of SOC-concavity, proving the SOC-concavity of $\ln t$ on $(0, +\infty)$ is equivalent to showing that $w \in \mathcal{K}^n$. From (21) and (11), it follows that

$$\begin{aligned} w &= \int_{-\infty}^{0} \left[(ue - \beta x - (1-\beta)y)^{-1} - \beta(ue - x)^{-1} - (1-\beta)(ue - y)^{-1}\right] du \\ &= \begin{pmatrix} \displaystyle\int_{-\infty}^{0} \left[\frac{u - \beta x_1 - (1-\beta)y_1}{\det(ue - \beta x - (1-\beta)y)} - \frac{\beta(u - x_1)}{\det(ue - x)} - \frac{(1-\beta)(u - y_1)}{\det(ue - y)}\right] du \\ \displaystyle\int_{-\infty}^{0} \left[\frac{\beta x_2 + (1-\beta)y_2}{\det(ue - \beta x - (1-\beta)y)} - \frac{\beta x_2}{\det(ue - x)} - \frac{(1-\beta)y_2}{\det(ue - y)}\right] du \end{pmatrix} \\ &:= \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}, \end{aligned}$$

where $w_1 \in \mathbb{R}$ and $w_2 \in \mathbb{R}^{n-1}$. However, by Proposition 3.1 (c) and Definition 2.1,

$$\Big(ue - \beta x - (1-\beta)y\Big)^{-1} - \beta(ue - x)^{-1} - (1-\beta)(ue - y)^{-1} \in \mathcal{K}^n,$$

which implies that

$$\frac{u - \beta x_1 - (1-\beta)y_1}{\det(ue - \beta x - (1-\beta)y)} - \frac{\beta(u - x_1)}{\det(ue - x)} - \frac{(1-\beta)(u - y_1)}{\det(ue - y)} \geq 0$$

and

$$\begin{aligned} &\left\| \frac{\beta x_2 + (1-\beta)y_2}{\det(ue - \beta x - (1-\beta)y)} - \frac{\beta x_2}{\det(ue - x)} - \frac{(1-\beta)y_2}{\det(ue - y)} \right\| \\ &\leq \frac{u - \beta x_1 - (1-\beta)y_1}{\det(ue - \beta x - (1-\beta)y)} - \frac{\beta(u - x_1)}{\det(ue - x)} - \frac{(1-\beta)(u - y_1)}{\det(ue - y)}. \end{aligned}$$

As a consequence,

$$\begin{aligned} w_1 &= \int_{-\infty}^{0} \left[\frac{u - \beta x_1 - (1-\beta)y_1}{\det(ue - \beta x - (1-\beta)y)} - \frac{\beta(u - x_1)}{\det(ue - x)} - \frac{(1-\beta)(u - y_1)}{\det(ue - y)}\right] du \\ &\geq 0 \end{aligned}$$

and

$$\begin{aligned} \|w_2\| &\leq \int_{-\infty}^{0} \left\| \left[\frac{\beta x_2 + (1-\beta)y_2}{\det(ue - \beta x - (1-\beta)y)} - \frac{\beta x_2}{\det(ue - x)} - \frac{(1-\beta)y_2}{\det(ue - y)}\right] \right\| du \\ &\leq \int_{-\infty}^{0} \left[\frac{u - \beta x_1 - (1-\beta)y_1}{\det(ue - \beta x - (1-\beta)y)} - \frac{\beta(u - x_1)}{\det(ue - x)} - \frac{(1-\beta)(u - y_1)}{\det(ue - y)}\right] du \\ &= w_1. \end{aligned}$$

This shows that $w \in \mathcal{K}^n$, and consequently $\ln t$ is SOC-concave on $(0, +\infty)$. By a similar argument, we can prove that $\ln t$ is SOC-monotone on $(0, +\infty)$. □

From Lemma 3.2 and Proposition 3.1, we may obtain the following corollary, which particularly shows that the modified logarithmic barrier function is SOC-concave.

COROLLARY 3.1. (a) *The modified logarithmic barrier function* $\ln(\alpha + t)$ *for* $\alpha > 0$ *is both SOC-concave and SOC-monotone on* $(-\alpha, +\infty)$.

(b) *For any* $\alpha > 0$ *and* $\beta > 0$, *the functions* $\ln(\alpha + \beta t^r)$, *with* $0 \leq r \leq 1$ *are SOC-concave and SOC-monotone on* $[0, +\infty)$.

(c) *For any $u > 0$, the functions $\frac{t}{u+t}$ are SOC-concave and SOC-monotone on $(0, +\infty)$.*

(d) *For all $u > 0$, the functions $\frac{-1}{\sqrt{u+t}}$ are SOC-concave and SOC-monotone on $(-u, +\infty)$.*

Proof. (a) The proof is due to Proposition 3.1(d), [7, Proposition 3.1], and Lemma 3.2 by letting $g : (0, +\infty) \to \mathbb{R}$ be $g(t) = \ln t$, and $h : (-a, +\infty) \to (0, +\infty)$ be $h(t) = a + t$.

(b) Let $g : (0, +\infty) \to \mathbb{R}$ be $g(t) = \ln t$, and $h : (0, +\infty) \to (0, +\infty)$ be $h(t) = a + \beta t^r$. The result follows from Proposition 3.1(a), Proposition 3.1(d), and Lemma 3.2.

(c) Let $g : (-1, 0) \to (0, 1)$ be $g(t) = 1 + t$, and $h : (0, +\infty) \to (-1, 0)$ be $h(t) = -u/(u+t)$. Then, we obtain the result from Proposition 3.1(c), [7, Proposition 3.1], and Lemma 3.2. The result also extends the conclusion of [7, Proposition 3.4].

(d) Let $g : (0, +\infty) \to (0, +\infty)$ be $g(t) = \sqrt{t}$, and $h : (-u, +\infty) \to (0, +\infty)$ be $h(t) = u + t$. Then, from Lemma 3.2 it follows that $g(h(t)) = \sqrt{u + t}$ is SOC-concave and SOC-monotone on $(-u, +\infty)$. Using Lemma 3.2 again with $g(t) = -1/t$ and $h(t) = \sqrt{u + t}$, we obtain the desired result. □

Now we present serval examples of $D(x, y)$ to close this section. From these examples, we may see that the conditions required by $\phi \in \Phi$ are not so strict, and the construction of the distance-like functions in SOCs can be completed by selecting a class of single variate convex functions.

Example 3.1. Let $\phi(t) = t \ln t - t + 1$ if $t \geq 0$, and $\phi(t) = +\infty$ if $t < 0$. It is easy to verify that ϕ satisfies Conditions C.1–C.3. Also, by Proposition 3.1(d), Condition C.4 also holds. From formula (13), it follows that, for any $y \in \mathcal{K}^n$ and $x \in \mathrm{int}(\mathcal{K}^n)$,

$$\phi^{\mathrm{soc}}(y) = y \circ \ln y - y + e \quad \text{and} \quad (\phi')^{\mathrm{soc}}(x) = \ln x.$$

Consequently, the distance-like function induced by ϕ is given by

$$D_1(x, y) = \mathrm{tr}\,(y \circ \ln y - y \circ \ln x + x - y) \quad \forall x \in \mathrm{int}(\mathcal{K}^n), y \in \mathcal{K}^n.$$

This function is precisely the natural extension of the entropy-like distance $d_\varphi(\cdot, \cdot)$, with $\varphi(t) = -\ln t + t - 1$ to the SOCs. In addition, comparing $D_1(x, y)$ with the distance-like function $H(x, y)$ in Example 3.1 of [26], we note that $D_1(x, y) = H(y, x)$, but the proximal-like algorithms corresponding to them are completely different.

Example 3.2. Let $\phi(t) = t \ln t + (1 + t) \ln(1 + t) - (1 + t) \ln 2$ if $t \geq 0$, and $\phi(t) = +\infty$ if $t < 0$. By computing, we can show that ϕ satisfies Conditions C.1–C.3. Furthermore, from Proposition 3.1(d) and Corollary 3.1(a), we learn that ϕ also satisfies Condition C.4. This means that $\phi \in \Phi$. For any $y \in \mathcal{K}^n$ and $x \in \mathrm{int}(\mathcal{K}^n)$, we can compute that

$$\begin{aligned} \phi^{\mathrm{soc}}(y) &= y \circ \ln y + (e + y) \circ \ln(e + y) - \ln 2(e + y), \\ (\phi')^{\mathrm{soc}}(x) &= (2 - \ln 2)e + \ln x + \ln(e + x). \end{aligned}$$

Therefore, the distance-like function generated by such a ϕ is given by

$$D_2(x, y) = \mathrm{tr}\Big[-\ln(e + x) \circ (e + y) + y \circ (\ln y - \ln x) + (e + y) \circ \ln(e + y) - 2(y - x)\Big]$$

for any $x \in \mathrm{int}(\mathcal{K}^n)$ and $y \in \mathcal{K}^n$. It should be pointed out that $D_2(x, y)$ is not the extension of $d_\varphi(\cdot, \cdot)$, with $\varphi(t) = \phi(t)$ given by [18] to the SOCs.

Example 3.3. Take $\phi(t) = t^{\frac{2r+3}{2}} + t^2$, with $0 \leq r < \frac{1}{2}$ if $t \geq 0$, and $\phi(t) = +\infty$ if $t < 0$. It is easy to verify that ϕ satisfies Conditions C.1–C.3. Furthermore, from

Proposition 3.1(a) it follows that ϕ satisfies Condition C.4. Thus, $\phi \in \Phi$. By a simple computation,

$$\phi^{\rm soc}(y) = y^{\frac{2r+3}{2}} + y^2 \quad \forall y \in \mathcal{K}^n \ \text{ and } \ (\phi')^{\rm soc}(x) = \frac{2r+3}{2} x^{\frac{2r+1}{2}} + 2x \quad \forall x \in {\rm int}(\mathcal{K}^n).$$

Hence, the distance-like function induced by ϕ has the following expression:

$$D_3(x,y) = {\rm tr}\left[\frac{2r+1}{2} x^{\frac{2r+3}{2}} + x^2 - y \circ \Big(\frac{2r+3}{2} x^{\frac{2r+1}{2}} + 2x\Big) + y^{\frac{2r+3}{2}} + y^2\right].$$

Example 3.4. Let $\phi(t) = t^{a+1} + at\ln t - at$, with $0 < a \le 1$ if $t \ge 0$, and $\phi(t) = +\infty$ if $t < 0$. It is easily shown that ϕ satisfies Conditions C.1–C.3. By Proposition 3.1(a) and Proposition 3.1(d), ϕ' is SOC-concave on $(0, +\infty)$. Hence, $\phi \in \Phi$. For any $y \in \mathcal{K}^n$ and $x \in {\rm int}(\mathcal{K}^n)$,

$$\phi^{\rm soc}(y) = y^{a+1} + ay \circ \ln y - ay \quad \text{and} \quad (\phi')^{\rm soc}(x) = (a+1)x^a + a\ln x.$$

Consequently, the distance-like function induced by ϕ has the following expression:

$$D_4(x,y) = {\rm tr}\Big[ax^{a+1} + ax - y \circ \Big((a+1)x^a + a\ln x\Big) + y^{a+1} + ay \circ \ln y - ay\Big].$$

4. Properties of distance-like functions. In what follows, we study some favorable properties of the function $D(x,y)$. We begin with two technical lemmas that will be used in the subsequent analysis. Among others, the first lemma is a direct consequence of Lemma 2.2 and the definition of Φ.

LEMMA 4.1. *Given a $\phi \in \Phi$, let $\phi^{\rm soc}$ and $(\phi')^{\rm soc}$ be the vector-valued functions given as in* (13)*. Then, we have the following results:*

(a) *$\phi^{\rm soc}(x)$ and $(\phi')^{\rm soc}(x)$ are well-defined on $\mathcal{K}^n$ and ${\rm int}(\mathcal{K}^n)$, respectively, and*

$$\lambda_i[\phi^{\rm soc}(x)] = \phi[\lambda_i(x)], \quad \lambda_i[(\phi')^{\rm soc}(x)] = \phi'[\lambda_i(x)], \quad i = 1, 2.$$

(b) *$\phi^{\rm soc}(x)$ and $(\phi')^{\rm soc}(x)$ are continuously differentiable on ${\rm int}(\mathcal{K}^n)$, with the transposed Jacobian at x given as in formulas* (14)–(15)*.*

(c) *${\rm tr}[\phi^{\rm soc}(x)]$ and ${\rm tr}[(\phi')^{\rm soc}(x)]$ are continuously differentiable on ${\rm int}(\mathcal{K}^n)$, and*

$$\begin{aligned} \nabla {\rm tr}\Big[\phi^{\rm soc}(x)\Big] &= 2\nabla\phi^{\rm soc}(x)\cdot e = 2(\phi')^{\rm soc}(x), \\ \nabla {\rm tr}\Big[(\phi')^{\rm soc}(x)\Big] &= 2\nabla(\phi')^{\rm soc}(x)\cdot e = 2(\phi'')^{\rm soc}(x). \end{aligned} \tag{22}$$

(d) *The function ${\rm tr}[\phi^{\rm soc}(x)]$ is strictly convex on ${\rm int}(\mathcal{K}^n)$.*

LEMMA 4.2. *Given a $\phi \in \Phi$ and a fixed point $z \in \mathbb{R}^n$, let $\phi_z : {\rm int}(\mathcal{K}^n) \to \mathbb{R}$ be given by*

$$\phi_z(x) := {\rm tr}\Big[-z \circ (\phi')^{\rm soc}(x)\Big]. \tag{23}$$

Then, the function $\phi_z(x)$ possesses the following properties:

(a) *$\phi_z(x)$ is continuously differentiable on ${\rm int}(\mathcal{K}^n)$, with $\nabla\phi_z(x) = -2\nabla(\phi')^{\rm soc}(x)\cdot z$.*

(b) *$\phi_z(x)$ is convex over ${\rm int}(\mathcal{K}^n)$ when $z \in \mathcal{K}^n$, and furthermore, it is strictly convex over ${\rm int}(\mathcal{K}^n)$ when $z \in {\rm int}(\mathcal{K}^n)$.*

Proof. (a) Since $\phi_z(x) = -2\langle(\phi')^{\rm soc}(x), z\rangle$ for any $x \in {\rm int}(\mathcal{K}^n)$, we have that $\phi_z(x)$ is continuously differentiable on ${\rm int}(\mathcal{K}^n)$ by Lemma 4.1(c). Moreover, applying the chain rule for the inner product of two functions readily yields $\nabla\phi_z(x) = -2\nabla(\phi')^{\rm soc}(x)\cdot z$.

(b) By the continuous differentiability of $\phi_z(x)$, to prove the convexity of ϕ_z on ${\rm int}(\mathcal{K}^n)$, it suffices to prove the following inequality:

$$\phi_z\left(\frac{x+y}{2}\right) \le \frac{1}{2}\Big(\phi_z(x)+\phi_z(y)\Big) \quad \forall x, y \in {\rm int}(\mathcal{K}^n). \tag{24}$$

By Condition C.4, ϕ' is SOC-concave on $(0,+\infty)$. Therefore, we have that

$$-(\phi')^{\rm soc}\left(\frac{x+y}{2}\right) \succeq_{\mathcal{K}^n} -\frac{1}{2}\Big[(\phi')^{\rm soc}(x)+(\phi')^{\rm soc}(y)\Big],$$

i.e.,

$$(\phi')^{\rm soc}\left(\frac{x+y}{2}\right) - \frac{1}{2}(\phi')^{\rm soc}(x) - \frac{1}{2}(\phi')^{\rm soc}(y) \succeq_{\mathcal{K}^n} 0.$$

Using Lemma 2.1(a) and the fact that $z \in \mathcal{K}^n$, we then obtain that

$$\left\langle z,\ (\phi')^{\rm soc}\left(\frac{x+y}{2}\right) - \frac{1}{2}(\phi')^{\rm soc}(x) - \frac{1}{2}(\phi')^{\rm soc}(y)\right\rangle \ge 0, \tag{25}$$

which in turn implies that

$$\Big\langle -z,\ (\phi')^{\rm soc}\left(\frac{x+y}{2}\right)\Big\rangle \le \frac{1}{2}\Big\langle -z,\ (\phi')^{\rm soc}(x)\Big\rangle + \frac{1}{2}\Big\langle -z,\ (\phi')^{\rm soc}(y)\Big\rangle.$$

The last inequality is exactly the one in (24). Hence, ϕ_z is convex on ${\rm int}(\mathcal{K}^n)$ for $z \in \mathcal{K}^n$.

To prove the second part of the conclusions, we need only to prove that the inequality in (25) holds strictly for any $x, y \in {\rm int}(\mathcal{K}^n)$ and $x \ne y$. By Lemma 2.1(b), this is also equivalent to proving the vector $(\phi')^{\rm soc}\left(\frac{x+y}{2}\right) - \frac{1}{2}(\phi')^{\rm soc}(x) - \frac{1}{2}(\phi')^{\rm soc}(y)$ is nonzero, since

$$(\phi')^{\rm soc}\left(\frac{x+y}{2}\right) - \frac{1}{2}(\phi')^{\rm soc}(x) - \frac{1}{2}(\phi')^{\rm soc}(y) \in \mathcal{K}^n \quad \text{and} \quad z \in {\rm int}(\mathcal{K}^n).$$

From Condition C.4, it follows that ϕ' is concave on $(0,+\infty)$, since the SOC-concavity implies the concavity. This, together with the strict monotonicity of ϕ', implies that ϕ' is strictly concave on $(0,+\infty)$. Using Lemma 2.2(d), we then have that ${\rm tr}[(\phi')^{\rm soc}(x)]$ is strictly concave on ${\rm int}(\mathcal{K}^n)$. This means that, for any $x, y \in {\rm int}(\mathcal{K}^n)$ and $x \ne y$,

$${\rm tr}\left[(\phi')^{\rm soc}\Big(\frac{x+y}{2}\Big)\right] - \frac{1}{2}{\rm tr}\left[(\phi')^{\rm soc}(x)\right] - \frac{1}{2}{\rm tr}\left[(\phi')^{\rm soc}(y)\right] > 0. \tag{26}$$

In addition, we note that the first element of $(\phi')^{\rm soc}\left(\frac{x+y}{2}\right) - \frac{1}{2}(\phi')^{\rm soc}(x) - \frac{1}{2}(\phi')^{\rm soc}(y)$ is

$$\frac{\phi'\Big(\lambda_1\Big(\frac{x+y}{2}\Big)\Big) + \phi'\Big(\lambda_2\Big(\frac{x+y}{2}\Big)\Big)}{2} - \frac{\phi'(\lambda_1(x)) + \phi'(\lambda_2(x))}{4} - \frac{\phi'(\lambda_1(y)) + \phi'(\lambda_2(y))}{4},$$

which, by Property 2.1(c), can be rewritten as

$$\frac{1}{2}\mathrm{tr}\left[(\phi')^{\mathrm{soc}}\Big(\frac{x+y}{2}\Big)\right]-\frac{1}{4}\mathrm{tr}\left[(\phi')^{\mathrm{soc}}(x)\right]-\frac{1}{4}\mathrm{tr}\left[(\phi')^{\mathrm{soc}}(y)\right].$$

This together with (26) shows that $(\phi')^{\mathrm{soc}}\left(\frac{x+y}{2}\right)-\frac{1}{2}(\phi')^{\mathrm{soc}}(x)-\frac{1}{2}(\phi')^{\mathrm{soc}}(y)$ is nonzero for any $x,y\in\mathrm{int}(\mathcal{K}^n)$ and $x\neq y$. Consequently, ϕ_z is strictly convex on $\mathrm{int}(\mathcal{K}^n)$. □

Now we are in a position to study the properties of the distance-like function $D(x,y)$.

PROPOSITION 4.1. *Given a $\phi\in\Phi$, let $D(x,y)$ be defined as in* (18)*. Then,*

(a) *$D(x,y)\geq 0$ for any $x\in\mathrm{int}(\mathcal{K}^n)$ and $y\in\mathcal{K}^n$, and $D(x,y)=0$ if and only if $x=y$;*

(b) *for any fixed $y\in\mathcal{K}^n$, $D(\cdot,y)$ is continuously differentiable on* $\mathrm{int}(\mathcal{K}^n)$*, with*

$$\nabla_x D(x,y)=2\nabla(\phi')^{\mathrm{soc}}(x)\cdot(x-y); \tag{27}$$

(c) *for any fixed $y\in\mathcal{K}^n$, the function $D(\cdot,y)$ is convex over* $\mathrm{int}(\mathcal{K}^n)$*, and for any fixed $y\in\mathrm{int}(\mathcal{K}^n)$, $D(\cdot,y)$ is strictly convex over* $\mathrm{int}(\mathcal{K}^n)$*;*

(d) *for any fixed $y\in\mathrm{int}(\mathcal{K}^n)$, the function $D(\cdot,y)$ is essentially smooth;*

(e) *for any fixed $y\in\mathcal{K}^n$, the level sets $L_D(y,\gamma):=\{x\in\mathrm{int}(\mathcal{K}^n):D(x,y)\leq\gamma\}$ for all $\gamma\geq 0$ are bounded.*

Proof. (a) By Lemma 4.1(c), for any $x\in\mathrm{int}(\mathcal{K}^n)$ and $y\in\mathcal{K}^n$, we can rewrite $D(x,y)$ as

$$D(x,y)=\mathrm{tr}[\phi^{\mathrm{soc}}(y)]-\mathrm{tr}[\phi^{\mathrm{soc}}(x)]-\langle\nabla\mathrm{tr}[\phi^{\mathrm{soc}}(x)],\ y-x\rangle.$$

Notice that $\mathrm{tr}[\phi^{\mathrm{soc}}(x)]$ is strictly convex on $\mathrm{int}(\mathcal{K}^n)$ by Lemma 4.1(d), and hence $D(x,y)\geq 0$ for any $x\in\mathrm{int}(\mathcal{K}^n)$ and $y\in\mathcal{K}^n$, and $D(x,y)=0$ if and only if $x=y$.

(b) By Lemma 4.1(b) and Lemma 4.1(c), the functions $\mathrm{tr}[\phi^{\mathrm{soc}}(x)]$ and $\langle(\phi')^{\mathrm{soc}}(x),x\rangle$ are continuously differentiable on $\mathrm{int}(\mathcal{K}^n)$. Noting that, for any $x\in\mathrm{int}(\mathcal{K}^n)$ and $y\in\mathcal{K}^n$,

$$D(x,y)=\mathrm{tr}[\phi^{\mathrm{soc}}(y)]-\mathrm{tr}[\phi^{\mathrm{soc}}(x)]-2\langle(\phi')^{\mathrm{soc}}(x),y-x\rangle;$$

we then have the continuous differentiability of $D(\cdot,y)$ on $\mathrm{int}(\mathcal{K}^n)$. Furthermore,

$$\begin{aligned}\nabla_x D(x,y)&=-\nabla\mathrm{tr}[\phi^{\mathrm{soc}}(x)]-2\nabla(\phi')^{\mathrm{soc}}(x)\cdot(y-x)+2(\phi')^{\mathrm{soc}}(x)\\&=-2(\phi')^{\mathrm{soc}}(x)+2\nabla(\phi')^{\mathrm{soc}}(x)\cdot(x-y)+2(\phi')^{\mathrm{soc}}(x)\\&=2\nabla(\phi')^{\mathrm{soc}}(x)\cdot(x-y).\end{aligned}$$

(c) By the definition of ϕ_z given as in (23), $D(x,y)$ can be rewritten as

$$D(x,y)=\mathrm{tr}[(\phi')^{\mathrm{soc}}(x)\circ x-\phi^{\mathrm{soc}}(x)]+\phi_y(x)+\mathrm{tr}[\phi^{\mathrm{soc}}(y)].$$

Thus, to prove the (strict) convexity of $D(\cdot,y)$ on $\mathrm{int}(\mathcal{K}^n)$, it suffices to show that

$$\mathrm{tr}[(\phi')^{\mathrm{soc}}(x)\circ x-\phi^{\mathrm{soc}}(x)]+\phi_y(x)$$

is (strictly) convex on $\mathrm{int}(\mathcal{K}^n)$. Let $\psi:(0,+\infty)\to\mathbb{R}$ be the function defined by

$$\psi(t):=\phi'(t)t-\phi(t). \tag{28}$$

Then, the vector-valued function induced by ψ via (13) is $(\phi')^{\rm soc}(x)\circ x-\phi^{\rm soc}(x)$, i.e.,

$$\psi^{\rm soc}(x)=(\phi')^{\rm soc}(x)\circ x-\phi^{\rm soc}(x). \tag{29}$$

From Condition C.3 and Lemma 2.2(d), it follows that $\mathrm{tr}[(\phi')^{\rm soc}(x)\circ x-\phi^{\rm soc}(x)]$ is convex over $\mathrm{int}(\mathcal{K}^n)$. In addition, by Lemma 4.2(b), $\phi_y(x)$ is convex on $\mathrm{int}(\mathcal{K}^n)$ if $y\in\mathcal{K}^n$, and it is strictly convex if $y\in\mathrm{int}(\mathcal{K}^n)$. Thus, we get the desired results.

(d) From [29, p. 251] and parts (a)–(b), to prove that $D(\cdot,y)$ is essentially smooth for any fixed $y\in\mathrm{int}(\mathcal{K}^n)$, it suffices to show that $\|\nabla_x D(x^k,y)\|\to+\infty$ for any $\{x^k\}\subset\mathrm{int}(\mathcal{K}^n)$, with $x^k\to x\in\mathrm{bd}(\mathcal{K}^n)$. We next prove the conclusion by the following two cases: $x_1>0$ and $x_1=0$. For the sake of notation, let $x^k=(x_1^k,x_2^k)\in\mathbb{R}\times\mathbb{R}^{n-1}$.

Case 1. $x_1>0$. In this case, $\|x_2\|=x_1>0$, since $x\in\mathrm{bd}(\mathcal{K}^n)$. Noting that $x^k\to x$, we have that $x_2^k\neq 0$ for all sufficiently large k. From the gradient formula (27),

$$\|\nabla_x D(x^k,y)\|=\|2\nabla(\phi')^{\rm soc}(x^k)\cdot(x^k-y)\|\geq\left|2[\nabla(\phi')^{\rm soc}(x^k)\cdot(x^k-y)]_1\right|, \tag{30}$$

where $[\nabla(\phi')^{\rm soc}(x^k)\cdot(x^k-y)]_1$ denotes the first element of the vector $\nabla(\phi')^{\rm soc}(x^k)\cdot(x^k-y)$. By the gradient formula (15), we can compute that

$$\begin{aligned}2[\nabla(\phi')^{\rm soc}(x^k)\cdot(x^k-y)]_1&=[\phi''(\lambda_2(x^k))+\phi''(\lambda_1(x^k))](x_1^k-y_1)\\&\quad+[\phi''(\lambda_2(x^k))-\phi''(\lambda_1(x^k))]\frac{(x_2^k-y_2)^Tx_2^k}{\|x_2^k\|}\\&=\phi''(\lambda_2(x^k))\left(\lambda_2(x^k)-y_1-y_2^Tx_2^k/\|x_2^k\|\right)\\&\quad-\phi''(\lambda_1(x^k))\left(y_1-y_2^Tx_2^k/\|x_2^k\|-\lambda_1(x^k)\right).\end{aligned} \tag{31}$$

Therefore,

$$\begin{aligned}\left|2[\nabla(\phi')^{\rm soc}(x^k)\cdot(x^k-y)]_1\right|&\geq\left|\phi''(\lambda_1(x^k))\left(y_1-y_2^Tx_2^k/\|x_2^k\|-\lambda_1(x^k)\right)\right|\\&\quad-\left|\phi''(\lambda_2(x^k))\left(\lambda_2(x^k)-y_1-y_2^Tx_2^k/\|x_2^k\|\right)\right|\\&\geq\left|\phi''(\lambda_1(x^k))\right|\cdot\left(\left|y_1-y_2^Tx_2^k/\|x_2^k\|\right|-\lambda_1(x^k)\right)\\&\quad-\left|\phi''(\lambda_2(x^k))\right|\cdot\left|\lambda_2(x^k)-y_1-y_2^Tx_2^k/\|x_2^k\|\right|\\&\geq\left|\phi''(\lambda_1(x^k))\right|\cdot\left(\lambda_1(y)-\lambda_1(x^k)\right)\\&\quad-\left|\phi''(\lambda_2(x^k))\right|\cdot\left|\lambda_2(x^k)-y_1-y_2^Tx_2^k/\|x_2^k\|\right|.\end{aligned}$$

Noting that $\lambda_1(x^k)\to\lambda_1(x)=0$, $\lambda_2(x^k)\to\lambda_2(x)>0$, and $\frac{y_2^Tx_2^k}{\|x_2^k\|}\to\frac{y_2^Tx_2}{\|x_2\|}$ as $k\to+\infty$, the second term in the right-hand side of the last inequality converges to a finite value, whereas the first term approaches to $+\infty$, since $|\phi''(\lambda_1(x^k))|\to+\infty$ by Condition C.2 and $\lambda_1(y)-\lambda_1(x^k)\to\lambda_1(y)>0$. This implies that as $k\to+\infty$,

$$\left|2[\nabla(\phi')^{\rm soc}(x^k)\cdot(x^k-y)]_1\right|\to+\infty.$$

Combining with the inequality (30) immediately yields $\|\nabla_x D(x^k,y)\|\to+\infty$.

Case 2. $x_1=0$. In this case, we necessarily have that $x=0$, since $x\in\mathcal{K}^n$. Considering that $x^k\to x$, it then follows that $x_2^k=0$ or $x_2^k>0$ for all sufficiently large k. If $x_2^k=0$ for all sufficiently large k, then from (14) we have that

$$\|\nabla_x D(x^k,y)\|=\|2\phi''(x_1^k)(x^k-y)\|\geq 2|\phi''(x_1^k)|\cdot|x_1^k-y_1|.$$

Since $y_1 > 0$ by $y \in \mathrm{int}(\mathcal{K}^n)$ and $x_1^k \to x_1 = 0$, applying Condition C.2 yields that the right-hand side tends to $+\infty$, and consequently $\|\nabla_x D(x^k, y)\| \to +\infty$ when $k \to +\infty$.

Next, we consider the case that $x_2^k > 0$ for all sufficiently large k. In this case, the inequalities (30)–(31) still hold. By Cauchy–Schwarz inequality,

$$\begin{aligned}\lambda_2(x^k) - y_1 - y_2^T x_2^k/\|x_2^k\| &\geq \lambda_2(x^k) - y_1 - \|y_2\| = \lambda_2(x^k) - \lambda_2(y),\\ y_1 - y_2^T x_2^k/\|x_2^k\| - \lambda_1(x^k) &\geq y_1 - \|y_2\| - \lambda_1(x^k) = \lambda_1(y) - \lambda_1(x^k).\end{aligned}$$

Since $\lambda_1(x^k), \lambda_2(x^k) \to 0$ as $k \to +\infty$ and $\lambda_1(y), \lambda_2(y) > 0$ by $y \in \mathrm{int}(\mathcal{K}^n)$, the last two inequalities imply that

$$\begin{aligned}\lambda_2(x^k) - y_1 - y_2^T x_2^k/\|x_2^k\| &\to -\lambda_2(y) < 0,\\ y_1 - y_2^T x_2^k/\|x_2^k\| - \lambda_1(x^k) &\to \lambda_1(y) > 0.\end{aligned}$$

On the other hand, by Condition C.2, when $k \to +\infty$,

$$\phi''(\lambda_2(x^k)) \to +\infty, \quad \phi''(\lambda_1(x^k)) \to +\infty.$$

The two sides show that the right-hand side of (31) approaches to $-\infty$ as $k \to +\infty$, and consequently, $2|[\nabla(\phi')^{\mathrm{soc}}(x^k) \cdot (x^k - y)]_1| \to +\infty$. Thus, from (30), it follows that $\|\nabla_x D(x^k, y)\| \to +\infty$ as $k \to +\infty$.

(e) From the definition of $D(x, y)$, it follows that, for any $x, y \in \mathrm{int}(\mathcal{K}^n)$,

$$\begin{aligned}D(x,y) &= \mathrm{tr}[\phi^{\mathrm{soc}}(y)] - \mathrm{tr}[\phi^{\mathrm{soc}}(x)] - \mathrm{tr}[(\phi')^{\mathrm{soc}}(x) \circ y] + \mathrm{tr}[(\phi')^{\mathrm{soc}}(x) \circ x]\\ &= \sum_{i=1}^{2} \phi(\lambda_i(y)) - \sum_{i=1}^{2} \phi(\lambda_i(x)) - \mathrm{tr}[(\phi')^{\mathrm{soc}}(x) \circ y] + \mathrm{tr}[(\phi')^{\mathrm{soc}}(x) \circ x], \qquad (32)\end{aligned}$$

where the second equality is from Lemma 4.1(a) and Property 2.1(c). Since

$$\begin{aligned}(\phi')^{\mathrm{soc}}(x) \circ x &= \left[\phi'(\lambda_1(x))u_x^{(1)} + \phi'(\lambda_2(x))u_x^{(2)}\right] \circ \left[\lambda_1(x)u_x^{(1)} + \lambda_2(x)u_x^{(2)}\right]\\ &= \phi'(\lambda_1(x))\lambda_1(x)u_x^{(1)} + \phi'(\lambda_2(x))\lambda_2(x)u_x^{(2)},\end{aligned}$$

we have from Lemma 2.2(a) that

$$\mathrm{tr}[(\phi')^{\mathrm{soc}}(x) \circ x] = \sum_{i=1}^{2} \phi'(\lambda_i(x))\lambda_i(x).$$

In addition, by Lemma 2.1(c) and Lemma 4.1(a), we have that

$$\mathrm{tr}[(\phi')^{\mathrm{soc}}(x) \circ y] \leq \sum_{i=1}^{2} \phi'(\lambda_i(x))\lambda_i(y).$$

Combining the last two inequalities with (32) yields that

$$\begin{aligned}D(x,y) &\geq \sum_{i=1}^{2} \left[\phi(\lambda_i(y)) - \phi(\lambda_i(x)) - \phi'(\lambda_i(x))\lambda_i(y) + \phi'(\lambda_i(x))\lambda_i(x)\right]\\ &= \sum_{i=1}^{2} \left[\phi(\lambda_i(y)) - \phi(\lambda_i(x)) - \phi'(\lambda_i(x))(\lambda_i(y) - \lambda_i(x))\right]\\ &= \sum_{i=1}^{2} d_B(\lambda_i(y), \lambda_i(x)),\end{aligned}$$

where $d_B : \mathbb{R}_+ \times \mathbb{R}_{++} \to \mathbb{R}$ is the function defined by

$$d_B(s,t) = \phi(s) - \phi(t) - \phi'(t)(s-t).$$

This implies that, for any fixed $y \in \mathcal{K}^n$ and $\gamma \geq 0$,

$$L_D(y,\gamma) \subseteq \left\{ x \in \mathrm{int}(\mathcal{K}^n) \;\Big|\; \sum_{i=1}^{2} d_B(\lambda_i(y), \lambda_i(x)) \leq \gamma \right\}. \tag{33}$$

Note that, for any fixed $s \geq 0$, the set $\{t > 0 : d_B(s,t) \leq 0\}$ equals to $\{s\}$ or $\emptyset$, and hence it is bounded. Thus, from [29, Corollary 8.7.1] and Condition C.3, it follows that the level sets $\{t > 0 : d_B(s,t) \leq \gamma\}$ for any fixed $s \geq 0$ are bounded. This together with (33) implies that the level sets $L_D(y,\gamma)$ are bounded for all $\gamma \geq 0$. □

PROPOSITION 4.2. *Given a $\phi \in \Phi$, let $D(x,y)$ be defined as in* (18). *Then, for all $x, y \in \mathrm{int}(\mathcal{K}^n)$ and $z \in \mathcal{K}^n$, we have the following inequality:*

$$\begin{aligned} D(x,z) - D(y,z) &\geq 2\langle \nabla(\phi')^{\mathrm{soc}}(y) \cdot (z-y),\ y-x \rangle \\ &= 2\langle \nabla(\phi')^{\mathrm{soc}}(y) \cdot (y-x),\ z-y \rangle. \end{aligned} \tag{34}$$

Proof. From the definition of $D(x,y)$ and $\phi_z(x)$ and equality (29), it follows that

$$\begin{aligned} D(x,z) - D(y,z) &= \mathrm{tr}[(\phi')^{\mathrm{soc}}(x) \circ x - \phi^{\mathrm{soc}}(x)] + \phi_z(x) \\ &\quad -\mathrm{tr}[(\phi')^{\mathrm{soc}}(y) \circ y - \phi^{\mathrm{soc}}(y)] - \phi_z(y) \\ &= \mathrm{tr}[\psi^{\mathrm{soc}}(x)] - \mathrm{tr}[\psi^{\mathrm{soc}}(y)] + \phi_z(x) - \phi_z(y) \\ &\geq \langle \nabla \mathrm{tr}[\psi^{\mathrm{soc}}(y)],\ x-y \rangle + \langle \nabla \phi_z(y),\ x-y \rangle \\ &= \langle 2(\psi')^{\mathrm{soc}}(y),\ x-y \rangle - \langle 2\nabla(\phi')^{\mathrm{soc}}(y) \cdot z,\ x-y \rangle, \end{aligned} \tag{35}$$

where the inequality is due to the convexity of $\mathrm{tr}[\psi^{\mathrm{soc}}(x)]$ and $\phi_z(x)$, and the last equality follows from Lemma 2.2(c) and Lemma 4.2(a). From the definition of ψ given as in (28), it is easy to compute that

$$\langle (\psi')^{\mathrm{soc}}(y),\ x-y \rangle = \langle (\phi'')^{\mathrm{soc}}(y) \circ y,\ x-y \rangle. \tag{36}$$

In addition, by the gradient formulas in (14)–(15), we can compute that

$$\nabla(\phi')^{\mathrm{soc}}(y) \cdot y = (\phi'')^{\mathrm{soc}}(y) \circ y, \tag{37}$$

which in turn implies that

$$\begin{aligned} &\langle \nabla(\phi')^{\mathrm{soc}}(y) \cdot z,\ x-y \rangle \\ &= \langle \nabla(\phi')^{\mathrm{soc}}(y) \cdot (y+z-y),\ x-y \rangle \\ &= \langle \nabla(\phi')^{\mathrm{soc}}(y) \cdot y,\ x-y \rangle + \langle \nabla(\phi')^{\mathrm{soc}}(y) \cdot (z-y),\ x-y \rangle \\ &= \langle (\phi'')^{\mathrm{soc}}(y) \circ y,\ x-y \rangle + \langle \nabla(\phi')^{\mathrm{soc}}(y) \cdot (z-y),\ x-y \rangle. \end{aligned}$$

This, together with (36) and (35), yields the first inequality in (34), whereas the second inequality follows from the symmetry of the matrix $\nabla(\phi')^{\mathrm{soc}}(y)$. □

Propositions 4.1–4.2 indicate that $D(x,y)$ possesses some favorable properties similar to those for d_φ. In the next section, we will employ these properties to establish the convergence for an approximate version of the proximal-like algorithm (8).

5. Approximate proximal-like algorithm. The proximal-like algorithm described as (8) for the CSOCP consists of a sequence of exact minimization. However, in practical computations, it is impossible to obtain the exact solution of these minimization problems. In this section, we consider an approximate version of this algorithm, which allows the inexact solution of the subproblems (8). Throughout this section, we make the following assumptions for the CSOCP:

(A1) $\inf\{f(\zeta) \mid \zeta \in \mathcal{F}\} := f_* > -\infty$ and $\mathrm{dom} f \cap \mathrm{int}(\mathcal{F}) \neq \emptyset$.

(A2) The matrix A is of maximal rank m.

Remark 5.1. Assumption A1 is elementary for the existence of the solution of the CSOCP. Assumption A2 is common in the solution of the SOCPs, which is clearly satisfied when $\mathcal{F} = \{\zeta \in \mathbb{R}^n \mid \zeta \succeq_{\mathcal{K}^n} 0\}$. Moreover, if we consider the linear SOCP

$$\begin{array}{l} \min\ \bar{c}^T x \\ \text{s.t. } \bar{A}x = \bar{b}, \quad x \in \mathcal{K}^n, \end{array} \tag{38}$$

where $\bar{A} \in \mathbb{R}^{m\times n}$, with $m \leq n$, $\bar{b} \in \mathbb{R}^m$ and $\bar{c} \in \mathbb{R}^n$, the assumption that $\bar{A}$ has full row rank m is standard. Consequently, its dual problem, given by

$$\begin{array}{l} \max\ \bar{b}^T y \\ \text{s.t. } \bar{c} - \bar{A}^T y \succeq_{\mathcal{K}^n} 0 \end{array} \tag{39}$$

satisfies assumption A2. This shows that we can solve the linear SOCP by applying the approximate proximal-like algorithm described below to the dual problem (39). In addition, from Lemma 1 in the appendix, we know that the recession cone of $\mathcal{F}$ is given by $0^+\mathcal{F} = \{d \in \mathbb{R}^m | \ Ad \succeq_{\mathcal{K}^n} 0\}$. This implies that assumption A2 is also satisfied when $\mathcal{F}$ is supposed to be bounded, since its recession cone $0^+\mathcal{F}$ now reduces to zero.

For the sake of notation, in what follows, we denote $\mathcal{D} : \mathrm{int}(\mathcal{F}) \times \mathcal{F} \to \mathbb{R}$ by

$$\mathcal{D}(\zeta, \xi) := D(A\zeta + b, A\xi + b). \tag{40}$$

From Proposition 4.1, we readily obtain the following properties of $\mathcal{D}(\zeta, \xi)$.

LEMMA 5.1. *Let $\mathcal{D}(\zeta, \xi)$ be defined by* (40)*. Then, under assumption* A2*, we have that*

(a) $\mathcal{D}(\zeta, \xi) \geq 0$ *for any* $\zeta \in \mathrm{int}(\mathcal{F})$ *and* $\xi \in \mathcal{F}$*, and* $\mathcal{D}(\zeta, \xi) = 0$ *if and only if* $\zeta = \xi$*;*

(b) *the function* $\mathcal{D}(\cdot, \xi)$ *for any fixed* $\xi \in \mathcal{F}$ *is continuously differentiable on* $\mathrm{int}(\mathcal{F})$*, with*

$$\nabla_\zeta \mathcal{D}(\zeta, \xi) = 2A^T \nabla(\phi')^{\mathrm{soc}}(A\zeta + b)A(\zeta - \xi); \tag{41}$$

(c) *for any fixed* $\xi \in \mathcal{F}$*, the function* $\mathcal{D}(\cdot, \xi)$ *is convex on* $\mathrm{int}(\mathcal{F})$*, and for any fixed* $\xi \in \mathrm{int}(\mathcal{F})$*, then* $\mathcal{D}(\cdot, \xi)$ *is strictly convex over* $\mathrm{int}(\mathcal{F})$*;*

(d) *for any fixed* $\xi \in \mathrm{int}(\mathcal{F})$*, the function* $\mathcal{D}(\cdot, \xi)$ *is essentially smooth;*

(e) *for any fixed* $\xi \in \mathcal{F}$*, the level sets* $L(\xi, \gamma) = \left\{\zeta \in \mathrm{int}(\mathcal{F}) : \mathcal{D}(\zeta, \xi) \leq \gamma\right\}$ *for all* $\gamma \geq 0$ *are bounded.*

Now we describe an approximate version of the proximal-like algorithm (APM) (8).

The APM. *Given a starting point* $\zeta^0 \in \mathrm{int}(\mathcal{F})$ *and constants* $\epsilon_k \geq 0$ *and* $\mu_k > 0$*, generate the sequence* $\{\zeta^k\} \subset \mathrm{int}(\mathcal{F})$ *satisfying*

$$g^k \in \partial_{\epsilon_k} f(\zeta^k), \tag{42}$$

$$\mu_k g^k + \nabla_\zeta \mathcal{D}(\zeta^k, \zeta^{k-1}) = 0, \tag{43}$$

where $\partial_\epsilon f$ *represents the* ϵ*-subdifferential of* f*.*

Remark 5.2. The APM can be regarded as an approximate version of the proximal algorithm (8) in the following sense. From the relation in (42) and the convexity of $\mathcal{D}(\cdot,\xi)$ over $\mathrm{int}(\mathcal{F})$ for any fixed $\xi \in \mathrm{int}(\mathcal{F})$, it follows that, for any $u \in \mathrm{int}(\mathcal{F})$,

$$f(u) \geq f(\zeta^k) + \langle u - \zeta^k, g^k \rangle - \epsilon_k$$

and

$$\mu_k^{-1}\mathcal{D}(u,\zeta^{k-1}) \geq \mu_k^{-1}\mathcal{D}(\zeta^k,\zeta^{k-1}) + \mu_k^{-1}\langle \nabla_\zeta \mathcal{D}(\zeta^k,\zeta^{k-1}),\ u-\zeta^k\rangle.$$

Adding the last two inequalities and using (43) yields that

$$f(u) + \mu_k^{-1}\mathcal{D}(u,\zeta^{k-1}) \geq f(\zeta^k) + \mu_k \mathcal{D}(\zeta^k,\zeta^{k-1}) - \epsilon_k.$$

This implies that

$$\zeta^k \in \epsilon_k - \mathrm{argmin}\left\{f(\zeta) + \mu_k^{-1}\mathcal{D}(\zeta,\zeta^{k-1})\right\}, \tag{44}$$

where, for a given function F and $\epsilon \geq 0$, the notation

$$\epsilon - \mathrm{argmin}\ F(\zeta) := \left\{\zeta^* : F(\zeta^*) \leq \inf F(\zeta) + \epsilon\right\}. \tag{45}$$

In the rest of this section, we focus on the convergence of the APM under assumptions A1 and A2. First, we prove that the APM generates a sequence $\{\zeta^k\} \subset \mathrm{int}(\mathcal{F})$, and consequently the APM is well-defined. This is implied by the following lemma.

LEMMA 5.2. *For any $\xi \in \mathrm{int}(\mathcal{F})$ and $\mu > 0$, we have that the following results hold:*

(a) *The function $F(\cdot) := f(\cdot) + \mu^{-1}\mathcal{D}(\cdot,\xi)$ has bounded level sets under assumption* A1.

(b) *If, in addition, assumption* A2 *holds, then there has a unique $\widehat{\zeta} \in \mathrm{int}(\mathcal{F})$ such that*

$$\widehat{\zeta} = \mathop{\mathrm{argmin}}_{\zeta \in \mathrm{int}(\mathcal{F})} \left\{f(\zeta) + \mu^{-1}\mathcal{D}(\zeta,\xi)\right\}, \tag{46}$$

and moreover, the minimum in the right-hand side is attained at $\widehat{\zeta}$ satisfying

$$-2\mu^{-1}A^T\nabla(\phi')^{\mathrm{soc}}(A\widehat{\zeta}+b)A(\widehat{\zeta}-\xi) \in \partial f(\widehat{\zeta}). \tag{47}$$

Proof. (a) Fix $\xi \in \mathrm{int}(\mathcal{F})$ and $\mu > 0$. By assumption A1 and the nonnegativity of $\mathcal{D}(\zeta,\xi)$, to show that $F(\zeta)$ has bounded level sets, it suffices to show that, for all $\nu \geq f_*$, the level sets $L(\nu) := \{\zeta \in \mathrm{int}(\mathcal{F}) \mid F(\zeta) \leq \nu\}$ are bounded. Notice that $L(\nu) \subseteq L(\xi, \mu(\nu - f_*))$ and $L(\xi,\gamma) := \{\zeta \in \mathrm{int}(\mathcal{F}) \mid \mathcal{D}(\zeta,\xi) \leq \gamma\}$ are bounded for all $\gamma \geq 0$ by Lemma 5.1 (e). Therefore, the sets $L(\nu)$ all $\nu \geq f_*$ are bounded.

(b) By Lemma 5.1(b), $F(\zeta)$ is a closed proper strictly convex function. Hence, if the minimum exists, it must be unique. From part (a), the minimizer $\widehat{\zeta}$ exists, and so it is unique. Under assumption A2, using the gradient formula in (41) and the optimality conditions for (46) then yields that

$$0 \in \partial f(\widehat{\zeta}) + 2\mu^{-1}A^T\nabla(\phi')^{\mathrm{soc}}(A\widehat{\zeta}+b)A(\widehat{\zeta}-\xi) + \partial\delta(\widehat{\zeta} \mid \mathcal{F}), \tag{48}$$

where $\delta(u \mid \mathcal{F}) = 0$ if $u \in \mathcal{F}$ and $+\infty$ otherwise. By Lemma 5.1(c) and [29, Theorem 26.1], we have that $\partial_\zeta \mathcal{D}(\zeta,\xi) = \emptyset$ for all $\zeta \in \mathrm{bd}(\mathcal{F})$. Hence, the relation in (48) implies that $\widehat{\zeta} \in \mathrm{int}(\mathcal{F})$. On the other hand, from [29, p. 226], we know that

$$\partial\delta(u \mid \mathcal{F}) = \{v \in \mathbb{R}^n \mid v \preceq_{\mathcal{K}^n} 0,\ \mathrm{tr}(v \circ u) = 0\}.$$

Using Lemma 2.1, we then obtain $\partial\delta(\widehat{\zeta} \mid \mathcal{F}) = \{0\}$. Thus, the proof is complete. ∎

Next, we investigate the properties of the sequence $\{\zeta^k\}$ generated by the APM.

PROPOSITION 5.1. *Let $\{\mu_k\}$ be any sequence of positive numbers and $\sigma_n = \sum_{k=1}^n \mu_k$. Let $\{\zeta^k\}$ be the sequence generated by the APM. Then,*

(a) $\mu_k[f(\zeta^k) - f(\zeta)] \le \mathcal{D}(\zeta^{k-1}, \zeta) - \mathcal{D}(\zeta^k, \zeta) + \mu_k\epsilon_k$ *for all* $\zeta \in \mathcal{F}$.

(b) $\mathcal{D}(\zeta^k, \zeta) \le D(\zeta^{k-1}, \zeta) + \mu_k\epsilon_k$ *for all* $\zeta \in \mathcal{F}$ *subject to* $f(\zeta) \le f(\zeta^k)$.

(c) $\sigma_n(f(\zeta^n) - f(\zeta)) \le \mathcal{D}(\zeta^0, \zeta) - D(\zeta^n, \zeta) + \sum_{k=1}^n \sigma_k\epsilon_k$ *for all* $\zeta \in \mathcal{F}$.

Proof. (a) For any $\zeta \in \mathcal{F}$, using the definition of the ϵ-subdifferential, we have that

$$f(\zeta) \ge f(\zeta^k) + \langle g^k,\ \zeta - \zeta^k\rangle - \epsilon_k, \tag{49}$$

where $g^k \in \partial_{\epsilon_k} f(\zeta^k)$. However, from (43) and (41), it follows that

$$g^k = -2\mu_k^{-1} A^T \nabla(\phi')^{\mathrm{soc}}(A\zeta^k + b)A(\zeta^k - \zeta^{k-1}).$$

Substituting this g^k into (49), we then obtain that

$$\mu_k[f(\zeta^k) - f(\zeta)] \le 2\Big\langle A^T \nabla(\phi')^{\mathrm{soc}}(A\zeta^k + b)A(\zeta^k - \zeta^{k-1}),\ \zeta - \zeta^k\Big\rangle + \mu_k\epsilon_k.$$

On the other hand, applying Proposition 4.2 at the points $x = A\zeta^{k-1} + b$, $y = A\zeta^k + b$, and $z = A\zeta + b$ and using the definition of $\mathcal{D}(\zeta, \xi)$ given by (40) yields that

$$\mathcal{D}(\zeta^{k-1}, \zeta) - \mathcal{D}(\zeta^k, \zeta) = 2\Big\langle A^T \nabla(\phi')^{\mathrm{soc}}(A\zeta^k + b)A(\zeta^k - \zeta^{k-1}),\ \zeta - \zeta^k\Big\rangle.$$

Combining the last two equations, we immediately obtain the result.

(b) The result follows directly from part (a) for any $\zeta \in \mathcal{F}$ such that $f(\zeta^k) \ge f(\zeta)$.

(c) First, from (44), it follows that

$$\zeta^k \in \epsilon_k - \operatorname{argmin}\left\{f(\zeta) + \mu_k^{-1}\mathcal{D}(\zeta, \zeta^{k-1})\right\}.$$

This implies that, for any $\zeta \in \operatorname{int}(\mathcal{F})$,

$$f(\zeta) + \mu_k^{-1}\mathcal{D}(\zeta, \zeta^{k-1}) \ge f(\zeta^k) + \mu_k^{-1}\mathcal{D}(\zeta^k, \zeta^{k-1}) - \epsilon_k.$$

Setting $\zeta = \zeta^{k-1}$ in this inequality and using Lemma 5.1(d) then yields that

$$f(\zeta^{k-1}) - f(\zeta^k) \ge \mu_k^{-1}\mathcal{D}(\zeta^k, \zeta^{k-1}) - \epsilon_k \ge -\epsilon_k.$$

Multiplying the above inequality by σ_{k-1} and summing over $k = 1, 2, \ldots, n$, we get

$$\sum_{k=1}^n \left[\sigma_{k-1} f(\zeta^{k-1}) - (\sigma_k - \mu_k) f(\zeta^k)\right] \ge -\sum_{k=1}^n \sigma_{k-1}\epsilon_k,$$

which, by noting that $\sigma_k = \mu_k + \sigma_{k-1}$ (with $\sigma_0 \equiv 0$), can be reduced to

$$\sigma_n f(\zeta^n) - \sum_{k=1}^n \mu_k f(\zeta^k) \le \sum_{k=1}^n \sigma_{k-1}\epsilon_k.$$

On the other hand, using part (a) and summing over $k = 1, 2, \ldots, n$, we have that

$$-\sigma_n f(\zeta) + \sum_{k=1}^n \mu_k f(\zeta^k) \le \mathcal{D}(\zeta^0, \zeta) - D(\zeta^n, \zeta) + \sum_{k=1}^n \mu_k\epsilon_k \quad \forall \zeta \in \mathcal{F}.$$

Adding the last two inequalities yields

$$\sigma_n(f(\zeta^n) - f(\zeta)) \le \mathcal{D}(\zeta^0, \zeta) - D(\zeta^n, \zeta) + \sum_{k=1}^{n} (\mu_k + \sigma_{k-1})\epsilon_k,$$

which proves (c) because $\mu_k + \sigma_{k-1} = \sigma_k$. ∎

We are now in a position to prove our main convergence result for the APM. For convenience, we denote the optimal set of the CSOCP by $\mathcal{X} := \{\zeta \mid f(\zeta) = f_*\}$.

PROPOSITION 5.2. *Let $\{\zeta^k\}$ be the sequence generated by the APM and $\sigma_n = \sum_{k=1}^{n} \mu_k$. Then, under assumptions* A1 *and* A2*, the following results hold.*

(a) *If $\sigma_n \to +\infty$ and $\mu_k^{-1}\sigma_k\epsilon_k \to 0$, then $\lim_{n\to+\infty} f(\zeta^n) \to f_*$.*

(b) *If the optimal set $\mathcal{X} \neq \emptyset$, $\sigma_n \to \infty$ and $\sum_{k=1}^{\infty} \mu_k\epsilon_k < \infty$, then the sequence ζ^k is bounded, and every accumulation point is a solution of the CSOCP.*

Proof. (a) From Proposition 5.1(c) and the nonnegativity of $\mathcal{D}(\zeta^n, \zeta)$, it follows that

$$f(\zeta^n) - f(\zeta) \le \sigma_n^{-1}\mathcal{D}(\zeta^0, \zeta) + \sigma_n^{-1} \sum_{k=1}^{n} \sigma_k\epsilon_k \quad \forall \zeta \in \mathcal{F}.$$

Taking the limit $\sigma_n \to +\infty$ to the two sides of the last inequality, we immediately have that the first term in the right-hand side goes to zero. In addition, applying Lemma 2 in the appendix with $a_{nk} := \sigma_n^{-1}\mu_k$ if $k \le n$ and $a_{nk} := 0$ otherwise and $u_k := \mu_k^{-1}\sigma_k\epsilon_k$, we obtain the second term in the right-hand side:

$$\sigma_n^{-1} \sum_{k=1}^{n} \sigma_k\epsilon_k = \sum_k a_{nk}u_k \to 0$$

because $\sigma_n \to +\infty$ and $\mu_k^{-1}\sigma_k\epsilon_k \to 0$. Therefore, we have that the

$$\lim_{n\to+\infty} f(\zeta^n) \le f_*.$$

This, together with the fact that $f(\zeta^n) \ge f_*$, implies the desired result.

(b) Suppose that $\zeta^* \in \mathcal{X}$. For any k, we have that $f(\zeta^k) \ge f(\zeta^*)$. From Proposition 5.1(b), it then follows that

$$\mathcal{D}(\zeta^k, \zeta^*) \le \mathcal{D}(\zeta^{k-1}, \zeta^*) + \mu_k\epsilon_k.$$

Since $\sum_{k=1}^{\infty} \mu_k\epsilon_k < +\infty$, using Lemma 3 in the appendix with $v_k := \mathcal{D}(\zeta^k, \zeta^*) \ge 0$ and $\beta_k := \mu_k\epsilon_k \ge 0$ yields that the sequence $\{\mathcal{D}(\zeta^k, \zeta^*)\}$ converges. Thus, by Proposition 5.1(e), the sequence $\{\zeta^k\}$ is bounded and consequently has an accumulation point. Without any loss of generality, let $\widehat{\zeta} \in \mathcal{F}$ be an accumulation point of $\{\zeta^k\}$. Then $\{\zeta^{k_j}\} \to \widehat{\zeta}$ for some $k_j \to +\infty$. Since f is lower semicontinuous, we get $f(\widehat{\zeta}) = \liminf_{k_j\to\infty} f(\zeta^{k_j})$. On the other hand, $f(\zeta^{k_j}) \to f_*$ by part (a). The two sides imply that $f(\widehat{\zeta}) = f_*$. Therefore, $\widehat{\zeta}$ is a solution of the CSOCP. The proof is thus complete. ∎

6. Numerical experiments. In this section, we present some preliminary numerical results for CSOCPs with a specific version of the concept APM, described as Algorithm 6.1 below, and compare the numerical performance of the algorithm

with that of the merit function approach [8]. The purpose of our numerical experiments is to verify the theoretical results obtained in the last section and to illustrate the effectiveness of the proximal-like method proposed.

ALGORITHM 6.1.

Given a sufficiently small $\tau > 0$, a sufficiently large M_0, and constants $\rho > 1$ and $\mu_1 > 0$. Choose a starting point $\zeta^0 \in \mathrm{int}(\mathcal{F})$ and set $k := 1$.

For $k = 1, 2, \cdots$ **until** $\mu_k \geq M_0$ **do**

1. Use an unconstrained minimization method to solve approximately the problem

$$\min_{\zeta \in \mathbb{R}^m} \; F_k(\zeta) := f(\zeta) + \mu_k^{-1} \mathcal{D}(\zeta, \zeta^{k-1}) \tag{50}$$

and obtain a ζ^k such that $\|\nabla f(\zeta^k) + \mu_k^{-1} \nabla_\zeta \mathcal{D}(\zeta^k, \zeta^{k-1})\| \leq \tau$.

2. Set $\mu_{k+1} = \rho \mu_k$ and $k := k + 1$, and then go back to Step 1.

End

Algorithm 6.1 is in fact a special APM with $\epsilon_k = \tau \|\zeta^k - \widehat{\zeta}^k\|$ and $\mu_k = \rho^{k-1}\mu_1$, where $\widehat{\zeta}^k$ is the solution of the subproblem (44), since using the strict convexity of $F_k(\zeta)$, we have that

$$F_k(\widehat{\zeta}^k) \geq F_k(\zeta^k) + \langle \nabla F_k(\zeta^k), \widehat{\zeta}^k - \zeta^k \rangle \geq F_k(\zeta^k) - \tau \|\zeta^k - \widehat{\zeta}^k\|,$$

which implies that

$$\zeta^k \in \epsilon_k - \mathrm{argmin} F_k(\zeta), \text{ with } \epsilon_k = \tau \|\zeta^k - \widehat{\zeta}^k\|.$$

Furthermore, such ϵ_k and μ_k at least satisfy the assumptions of Proposition 5.2(a), since

$$\sigma_n = \sum_{k=1}^{n} \mu_k \to +\infty, \quad \mu_k^{-1}\sigma_k \to \frac{\rho}{\rho - 1}, \quad \text{and} \quad \epsilon_k \to 0.$$

In our experiments, we employed the entropy-like distance functions from Example 3.1 and Example 3.2, respectively, for Algorithm 6.1. For convenience, let

$$\mathcal{D}_1(\zeta, \xi) := D_1(A\zeta + b, A\xi + b) \;\text{ and }\; \mathcal{D}_2(\zeta, \xi) := D_2(A\zeta + b, A\xi + b).$$

All numerical experiments were done on a personal computer with 2.8GHz CPU and 512MB memory. The computer codes were all written in Matlab 6.5. We chose a limited-memory BFGS method with 5 limited-memory vector-updates [4] to solve the minimization subproblem (50). In addition, we adopted a nonmonotone line search described as in [16] to seek a suitable steplength, i.e., we computed the smallest nonnegative integer l such that

$$F_k(\zeta^k + \beta^l d^k) \leq \mathcal{W}_k + \sigma \beta^l \nabla F_k(\zeta^k)^T d^k,$$

where d^k denotes the search direction at the kth iterate, and $\mathcal{W}_k = \max_{j=k-m_k,\ldots,k} F_k(\zeta^j)$ and where, for a given nonnegative integer $\hat{m}$ and s,

$$m_k = \begin{cases} 0 & \text{if } k \leq s, \\ \min\{m_{k-1} + 1, \hat{m}\} & \text{otherwise.} \end{cases}$$

Unless otherwise stated, we chose $\beta = 0.5$, $\sigma = 10^{-4}$, $\hat{m} = 5$, and $s = 5$ for the nonmontone line search, and the following parameters for Algorithm 6.1:

$$\tau = 10^{-5}, \quad M_0 = 1000, \quad \rho = 10, \quad \text{and} \quad \mu_1 = 1.$$

We applied Algorithm 6.1 for the following quadratic convex SOC program:

$$\begin{aligned} &\min \ \frac{1}{2}\zeta^T M\zeta + q^T\zeta \\ &\text{s.t. } \zeta \succeq_{\mathcal{K}^n} 0, \end{aligned} \tag{51}$$

where $M \in \mathbb{R}^{n\times n}$ is a symmetric positive semidefinite matrix and $q \in \mathbb{R}^n$ is a vector. In the experiment, the matrix M and the vector q were generated as follows: elements of q were chosen randomly from the interval $[-1, 1]$, and M was obtained by setting $M = DD^T$, where D was a sparse matrix with approximately *density* $\cdot\, n \cdot n$ nonzero entries, which were chosen from a normal distribution with mean -1 and variance 4. In this procedure, the number of nonzero entries of D is determined so that the nonzero density of M can be approximately estimated. To construct different types of $\mathcal{K}$, we chose n_i and N such that $n_1 + \cdots + n_N = 1000$ and $n_1 = \cdots = n_N = 100$. For each type of $\mathcal{K}$, we have solved 10 test problems with the matrix M of nonzero density 0.5%, 1% and 10%, respectively, and started Algorithm 6.1 from the initial point $\zeta^0 = (\bar{\zeta}^{n_i}, \ldots, \bar{\zeta}^{n_N})$, where $\bar{\zeta}^{n_i} = (2, \omega_i/\|\omega_i\|)$ for $i = 1, 2, \ldots, N$, with $\omega_i \in \mathbb{R}^{n_i-1}$ generated randomly by Matlab's **randn.m**. We also employed the merit function approach [8] to solve these test problems. In other words, we chose the same limited-memory BFGS method as used by Algorithm 6.1 to solve the unconstrained minimization reformulation for the KKT conditions of (51):

$$\min_{\zeta\in\mathbb{R}^n} \Psi_{\rm FB}(\zeta) := \frac{1}{2}\left\|(\zeta^2 + (M\zeta+q)^2)^{1/2} - \zeta - (M\zeta+q)\right\|^2. \tag{52}$$

For the merit function approach, we used the same starting point ζ^0 as Algorithm 6.1 and terminated the iterates once $\sqrt{2\Psi_{\rm FB}(\zeta)} \le 10^{-4}$.

The numerical results were listed in Tables 1–3 (see the appendix), where **Rcond** denotes the condition number of the matrix M computed by Matlab's **rcond.m**, **Gap** means the absolute dual gap, i.e., the value of the function $|\zeta^T(M\zeta+q)|$ at the final iteration, **NF** represents the number of function evaluations for $F_k(\zeta)$ or $\Psi_{\rm FB}(\zeta)$ to solve each problem, which for Algorithm 6.1 is the total sum of the function evaluations used for every subproblem, and **Cpu** represents the CPU time in seconds for solving each problem.

From Tables 1–3, we see that Algorithm 6.1 with $D_1(x, y)$ and $D_2(x, y)$ can solve almost all of the test problems within 10^5 function evaluations, except three test problems in Table 1 for which the merit function approach cannot yield the desired result within 5×10^4 function evaluations, too. Algorithm 6.1 requires more function evaluations than the merit function approach, especially for the problems with the matrix M of nonzero density 0.5% and 1%. This is reasonable since the APM is only a primal algorithm, whereas the merit function approach is a primal-dual one. When comparing Table 1 with Tables 2–3, we find that the condition number of M has a great influence on the numerical performance of Algorithm 6.1 and the merit function approach; for example, the two methods have the worst robustness when the condition number of M equals 0. In addition, from Tables 1–3, it seems that the number of function evaluations of Algorithm 6.1 is not influenced by the nonzero density of

TABLE 1
Numerical results for the matrix M with 0.5% nonzero density.

		$\mathcal{D}_1(\zeta,\xi)$			$\mathcal{D}_2(\zeta,\xi)$			Merit function approach		
N0.	**Rcond**	**Gap**	**Nf**	**Time**	**Gap**	**Nf**	**Time**	**Gap**	**Nf**	**Time**
1	0	2.32e–3	28524	84.8	7.76e–3	37480	127.1	1.72e–4	6034	47.2
2	0	–	$> 10^5$	–	–	$> 10^5$	–	–	>50000	–
3	0	–	$> 10^5$	–	–	$> 10^5$	–	–	>50000	–
4	0	–	$> 10^5$	–	–	$> 10^5$	–	–	>50000	–
5	0	1.62e–3	17436	50.9	4.34e–3	25882	100.9	8.91e–4	957	7.56
6	0	1.36e–3	20516	62.4	3.56e–3	25125	91.2	6.35e–4	883	5.76
7	0	4.79e–3	53555	168.0	1.58e–2	53460	194.3	8.24e–4	1506	12.9
8	0	4.95e–3	79164	236.8	1.64e–2	99687	378.2	1.97e–5	4343	36.0
9	0	1.93e–3	29433	91.2	4.74e–3	31655	112.8	3.98e–4	795	6.43
10	0	6.75e–4	57670	168.0	–	$> 10^5$	–	7.39e–4	1492	11.4

TABLE 2
Numerical results for the matrix M with 1% nonzero density.

		$\mathcal{D}_1(\zeta,\xi)$			$\mathcal{D}_2(\zeta,\xi)$			Merit function approach		
N0.	**Rcond**	**Gap**	**Nf**	**Cpu**	**Gap**	**Nf**	**Cpu**	**Gap**	**Nf**	**Cpu**
1	0	2.65e–3	24966	117.4	5.48e–3	26395	142.4	1.19e–3	1695	23.0
2	1.64e–8	3.44e–3	32540	161.5	6.35e–3	33246	200.4	1.14e–3	1666	23.1
3	8.88e–11	1.94e–3	22785	123.7	3.45e–3	24865	137.9	3.42e–4	1421	18.8
4	2.45e–9	6.89e–3	60638	325.1	1.34e–2	66483	392.9	4.78e–4	2360	35.3
5	6.19e–10	2.97e–3	29612	157.3	5.62e–3	37027	230.3	8.53e–4	1594	24.4
6	5.22e–11	5.05e–3	22620	116.5	1.05e–2	26916	166.1	8.09e–4	1259	20.8
7	7.50e–11	2.09e–3	12419	63.6	3.75e–3	19326	121.7	2.70e–4	1728	21.2
8	4.97e–9	2.63e–3	20661	106.4	5.27e–3	30375	187.9	6.14e–4	1497	19.8
9	5.16e–11	4.32e–3	29157	153.9	7.85e–3	42502	266.1	4.04e–4	1734	24.7
10	1.96e–9	2.48e–3	23804	119.5	4.81e–3	34085	201.2	1.26e–3	1550	22.4

M, but the merit function approach clearly requires more function evaluations as the nonzero density of M increases. Moreover, the merit function approach needs more CPU time at each iteration than Algorithm 6.1 due to an extra multiplication of the matrix M and the vector $\nabla_y \psi_{\rm FB}(\zeta, M\zeta + q)$ involved in the computation of $\nabla \Psi_{\rm FB}(\zeta)$. This accounts for the fact that Algorithm 6.1 is superior to the merit

TABLE 3
Numerical results for the matrix M with 10% nonzero density.

		$\mathcal{D}_1(\zeta,\xi)$			$\mathcal{D}_2(\zeta,\xi)$			Merit function approach		
N0.	**Rcond**	**Gap**	**Nf**	**Cpu**	**Gap**	**Nf**	**Cpu**	**Gap**	**Nf**	**Cpu**
1	4.99e–9	1.12e–3	21486	671.4	1.37e–3	23255	755.6	5.81e–4	15208	1486.4
2	1.05e–8	1.03e–3	27822	896.8	1.24e–3	36318	1225.4	1.19e–3	21823	2054.5
3	2.63e–9	1.08e–3	32345	1031.8	1.33e–3	18571	603.6	1.57e–3	15770	1495.0
4	9.11e–9	1.03e–3	30856	714.6	1.20e–3	38199	890.0	7.05e-4	14742	1372.5
5	4.68e–10	1.11e–3	26957	869.5	1.34e–3	49756	1713.8	1.52e–3	16949	1691.5
6	1.13e–8	1.08e–3	33621	799.1	1.26e–3	38470	904.0	1.10e–3	17071	1591.9
7	1.20e–9	8.33e–4	19452	623.7	9.76e–4	22501	728.7	1.17e–3	23102	2140.7
8	1.21e–8	1.04e–3	24345	763.0	1.21e-3	33600	1185.9	1.46e–3	16011	1492.9
9	1.26e–8	1.04e-3	35613	821.3	1.22e-3	38723	923.2	1.15e–3	27110	2528.3
10	1.16e–8	8.43e-4	13580	410.8	1.00e-3	33869	1204.9	1.61e–3	17883	1655.3

function approach by the CPU time for the problems with the matrix M of nonzero density 10%.

We also applied Algorithm 6.1 for a nonlinear convex SOCP taken from [17].

Example 6.1. Consider the following nonlinear convex SOCP:

$$\begin{aligned} &\min\ \exp(\zeta_1-\zeta_3)+3(2\zeta_1-\zeta_2)^4+\sqrt{1+(3\zeta_2+5\zeta_3)^2}\\ &\text{s.t.}\ \begin{pmatrix} 4 & 6 & 3\\ -1 & 7 & -5\end{pmatrix}\zeta+\begin{pmatrix}-1\\ 2\end{pmatrix}\in\mathcal{K}^2,\quad \zeta\in\mathcal{K}^3. \end{aligned} \tag{53}$$

In order to obtain an initial interior point $\zeta^0\in\operatorname{int}(\mathcal{F})$ for Algorithm 6.1, we constructed the following conic optimization problem:

$$\begin{aligned} &\min\ w\\ &A\zeta+b+w\hat{e}\succeq_{\mathcal{K}}0,\\ &-w+w^*\geq 0,\\ &\mathcal{K}=\mathcal{K}^3\times\mathcal{K}^2, \end{aligned} \tag{54}$$

where $w^*\in\mathbb{R}$ is a constant and $\hat{e}=\begin{pmatrix}e_1\\ e_2\end{pmatrix}, A=\begin{bmatrix}A_1\\ A_2\end{bmatrix}, b=\begin{pmatrix}b_1\\ b_2\end{pmatrix}$, with

$$A_1=\begin{pmatrix}4 & 6 & 3\\ -1 & 7 & 5\end{pmatrix},\ A_2=I,\ b_1=(-1,2)^T,\ b_2=(0,0,0)^T, e_1=(1,0)^T,\ e_2=(1,0,0)^T.$$

It is easy to see that $\zeta=0,\ w=w_0$ belongs to $\operatorname{int}(\mathcal{F})$ only if $w_0>-\lambda_1(b_i),\ i=1,2$ and $w^*>w_0$, and furthermore, when solving (54) with Algorithm 6.1 from

TABLE 4
Numerical results for Example 6.1.

	$\mathcal{D}_1(\zeta,\xi)$			$\mathcal{D}_2(\zeta,\xi)$		
ζ^0	**fopt**	**Nf**	**Time**	**fopt**	**Nf**	**Time**
$(1.8860, -0.1890, -0.4081)^T$	2.597580	89804	79.65	2.597584	65165	62.46
$(4.3425, 0.0875, -0.2332)^T$	2.597591	40402	32.42	2.597610	35820	31.79
$(4.6972, -0.4294, -1.3931)^T$	2.597587	50889	44.36	2.597601	40404	39.59
$(12.3337, -2.6206, -6.2167)^T$	2.597585	67835	50.06	2.597599	64551	66.98
$(3.7282, 0.2875, 0.2737)^T$	2.597591	30763	26.58	2.597611	26485	24.48

$\zeta = 0,\ w = w_0$, if some iterate (ζ^k, w^k) satisfying $w^k < 0$, then the corresponding ζ^k can be used as the starting point to solve (53). This way can also be used to find the starting interior point ζ^0 when applying Algorithm 6.1 for other problems with the form of (1). We have solved the test problem with Algorithm 6.1 from several starting points. The parameters for Algorithm 6.1 were the same as above except $M_0 = 10000$ and $\tau = 10^{-6}$. The numerical results were listed in Table 4, where **fopt** denotes the objective value at the final iteration. We see that the choice of ζ^0 has an influence on the numerical behavior of Algorithm 6.1.

From Tables 1–4, we may draw the following conclusions: the approximate proximal-like algorithm using $D_1(x, y)$ has the better numerical behavior than the one using $D_2(x, y)$ whether by the accuracy of solution or the number of function evaluations required, and the proximal-like algorithm with an appropriate distance measure is superior to the merit function approach by the CPU time for those dense problems.

7. Conclusions. In this paper, we extended the entropy-like proximal algorithm proposed by Eggermont [12] for convex programming subject to nonnegative constraints and proposed a class of interior proximal-like algorithms for solving the CSOCPs. These algorithms are based on a distance-like function generated by a closed proper convex function ϕ satisfying $\mathrm{dom}\phi = [0, +\infty)$ and Conditions (C.1)–(C.4). The given examples illustrated that the conditions required by ϕ are not very stringent. For the proposed proximal-like algorithm, we particularly considered an approximate version which allows inexact minimization steps, and we established the convergence properties under some mild assumptions. Numerical results were also reported for the algorithm with the entropy-like distance functions from Examples 3.1 and 3.2, and we made comparisons with those yielded by the merit function approach [8], which verify the effectiveness of the proposed method.

In our future research works, we will analyze the convergence rate of the proposed algorithms and investigate some practical versions of the algorithms. In addition, we will consider the extension of the class of interior proximal-like algorithms to general convex symmetric cone programming problems. It should be pointed out that the extension is not direct. The main difficulty is how to extend the characterizations of SOC-convexity [7, 9] to the setting of symmetric cones.

Appendix A.

LEMMA 1. *Let $\mathcal{F}$ be the set defined as in* (8). *Then its recession cone $0^+\mathcal{F}$ is given by*

$$0^+\mathcal{F} = \Big\{ d \in \mathbb{R}^m \mid Ad \succeq_{\mathcal{K}^n} 0 \Big\}. \tag{55}$$

Proof. Assume that $d \in \mathbb{R}^m$ such that $Ad \succeq_{\mathcal{K}^n} 0$. Then, for any $\lambda > 0$, $\lambda Ad \succeq_{\mathcal{K}^n} 0$. Considering that $\mathcal{K}^n$ is closed under the "+" operation, we have that, for any $\zeta \in \mathcal{F}$,

$$A(\zeta + \lambda d) + b = (A\zeta + b) + \lambda(Ad) \succeq_{\mathcal{K}^n} 0. \tag{56}$$

By [29, p. 61], this shows that every element in the set of the right-hand side of (55) is a recession direction of $\mathcal{F}$. Consequently, $\{d \in \mathbb{R}^m \mid Ad \succeq_{\mathcal{K}^n} 0\} \subseteq 0^+\mathcal{F}$.

Now take any $d \in 0^+\mathcal{F}$ and $\zeta \in \mathcal{F}$. Then, for any $\lambda > 0$, equation (56) holds. By Property 2.1(d), we then have $\lambda_1\Big[(A\zeta + b) + \lambda Ad\Big] \geq 0$ for any $\lambda > 0$. This implies that $\lambda_1(Ad) \geq 0$, since otherwise letting $\lambda \to +\infty$ and using the fact that

$$\begin{aligned}
\lambda_1\Big[(A\zeta + b) + \lambda Ad\Big] &= (A\zeta + b)_1 + \lambda(Ad)_1 - \|(A\zeta + b)_2 + \lambda(Ad)_2\| \\
&\leq (A\zeta + b)_1 + \lambda(Ad)_1 - \Big(\lambda\|(Ad)_2\| - \|(A\zeta + b)_2\|\Big) \\
&= \lambda\lambda_1(Ad) + \lambda_2(A\zeta + b),
\end{aligned}$$

we obtain that $\lambda_1[(A\zeta + b) + \lambda Ad] \to -\infty$. Thus, we prove that $Ad \succeq_{\mathcal{K}^n} 0$, and consequently $0^+\mathcal{F} \subseteq \{d \in \mathbb{R}^m \mid Ad \succeq_{\mathcal{K}^n} 0\}$. Combining with the above discussions then yields the result. □

LEMMA 2 (see [20, Theorem 2]). *Let $\{a_{nk}\}$ be a sequence of real numbers satisfying*

(i) $a_{nk} \geq 0$ $\forall n = 1, 2, \ldots,\ k = 1, 2, \ldots$.

(ii) $\sum_{k=1}^{\infty} a_{nk} = 1$ $\forall n = 1, 2, \ldots,$ *and* $\lim_{n\to+\infty} \sum_{k=1}^{n} a_{nk}u_k = u$ $\forall k = 1, 2, \ldots$.

If $\{u_k\}$ is a sequence such that $\lim_{k\to+\infty} u_k = u$, then $\lim_{k\to+\infty} a_{nk}u_k = u$.

LEMMA 3 (see [28, Chapter 2]). *Let $\{\upsilon_k\}$ and $\{\beta_k\}$ be nonnegative sequences of real numbers satisfying* (i) $\upsilon_{k+1} \leq \upsilon_k + \beta_k$, (ii) $\sum_{k=1}^{\infty} \beta_k < +\infty$. *Then the sequence $\{\upsilon_k\}$ is convergent.*

Acknowledgments. The authors would like to thank two anonymous referees for their helpful comments and suggestions on the revision of this paper. In particular, the influence of the condition number of M is observed by one referee.

REFERENCES

[1] F. ALIZADEH AND D. GOLDFARB, *Second-order cone programming*, Math. Program., 95 (2003), pp. 3–51.

[2] A. AUSLENDER AND M. HADDOU, *An interior-proximal method for convex linearly constrained problems and its extension to variational inequalities*, Math. Program., 71 (1995), pp. 77–100.

[3] J. BRINKHUIS, Z.-Q. LUO, AND S.-Z. ZHANG, *Matrix Convex Functions with Applications to Weighted Centers for Semidefinite Programming*, preprint; The Chinese University of Hong Kong, Hong Kong, 2005.

[4] R. H. BYRD, P. LU, J. NOCEDAL, AND C. ZHU, *A limited memory algorithm for bound constrained optimization*, SIAM J. Sci. Comput., 16 (1995), pp. 1190–1208.

[5] Y. CENSOR AND S. A. ZENIOS, *The proximal minimization algorithm with D-functions*, J. Optim. Theory Appl., 73 (1992), pp. 451–464.

[6] J.-S. Chen, X. Chen, and P. Tseng, *Analysis of nonsmooth vector-valued functions associated with second-order cones*, Math. Program., 101 (2004), pp. 95–117.
[7] J.-S. Chen, *The convex and monotone functions associated with second-order cone*, Optimization, 55 (2006), pp. 363–385.
[8] J.-S. Chen and P. Tseng, *An unconstrained smooth minimization reformulation of the second-order cone complementarity problem*, Math. Program., 104 (2005), pp. 293–327.
[9] J.-S. Chen, X. Chen, S.-H. Pan, and J.-W. Zhang, *The Characterizations of SOC-monotone and SOC-convex Function*, manuscript, 2007.
[10] X.-D. Chen, D. Sun, and J. Sun, *Complementarity functions and numerical experiments for second-order cone complementarity problems*, Comput. Optim. Appl., 25 (2003), pp. 39–56.
[11] M. Doljansky and M. Teboulle, *An interior proximal algorithm and the exponential multiplier method for semidefinite programming*, SIAM J. Optim., 9 (1998), pp. 1–13.
[12] P. P. B. Eggermont, *Multiplicative iterative algorithms for convex programming*, Linear Algebra Appl., 130 (1990), pp. 25–42.
[13] J. Eckstein, *Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming*, Math. Oper. Res., 18 (1993), pp. 202–226.
[14] J. Faraut and A. Korányi, *Analysis on Symmetrc Cones*, Oxford Mathematical Monographs, Oxford University Press, N. Y., 1994.
[15] M. Fukushima, Z.-Q. Luo, and P. Tseng, *Smoothing functions for second-order cone complementarity problems*, SIAM J. Optim., 12 (2002), pp. 436–460.
[16] L. Grippo, F. Lampariello, and S. Lucidi, *A nonmonotone line search technique for Newton's method*, SIAM J. Numer. Anal., 23 (1986), pp. 707–716.
[17] S. Hayashi, N. Yamashita, and M. Fukushima, *A combined smoothing and regularization method for monotone second-order cone complementarity problems*, SIAM J. Optim., 15 (2005), pp. 593–615.
[18] A. N. Iusem, B. F. Svaiter, and M. Teboulle, *Entropy-like proximal methods in convex programming*, Math. Oper. Res., 19 (1994), pp. 790–814.
[19] A. N. Iusem and M. Teboulle, *Convergence rate analysis of nonquadratic proximal and augmented Lagrangian methods for convex and linear programming*, Math. Oper. Res., 20 (1995), pp. 657–677.
[20] K. Knopp, *Infinite Sequences and Series*, Dover Publications, New York, 1956.
[21] Y.-J. Kuo and H. D. Mittelmann, *Interior point methods for second-order cone programming and OR applications*, Comput. Optim. Appl., 28 (2004), pp. 255–285.
[22] C. Kanzow, I. Ferenczi, and M. Fukushima, *Semismooth methods for linear and nonlinear second-order cone programs*, Technical report, Department of Applied Mathematics and Physics, Kyoto University, Kyoto, Japan, 2006.
[23] M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, *Application of second-order cone programming*, Linear Algebra Appl., 284 (1998), pp. 193–228.
[24] B. Martinet, *Perturbation des methodes d' Optimisation*, Appl. RAIRO Numer. Anal., 12 (1978), pp. 153–171.
[25] J. J. Moreau, *Promimité et Dualité dans un Espace Hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.
[26] S.-H. Pan and J.-S. Chen, *A proximal-like algorithnm using quasi D-function for convex second-order cone programming*, J. Optim. Theory Appl., 138 (2008), pp. 95–113.
[27] J. M. Peng, C. Roos, and T. Terlaky, *Self-regular functions and new search directions for linear and semi-definite optimization*, Math. Program., 93 (2002), pp. 129–171.
[28] B. T. Polyak, *Introduction to Optimization*, Optimization Software, Inc., New York, 1987.
[29] R. T. Rockafella, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[30] R. T. Rockafella, *Augmented Lagrangians and applications of proximial point algorithm in convex programming*, Math. Oper. Res., 1 (1976), pp. 97–116.
[31] R. T. Rockafella, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.
[32] M. Teboulle, *Entropic proximal mappings with applications to nonlinear programming*, Math. Oper. Res., 17 (1992), pp. 670–690.
[33] A. Yoshise, *Interior point trajectories and a homogeneous model for nonlinear complementarity problems over symmetric cones*, SIAM J. Optim., 17 (2006), pp. 1129–1153.

SIAM J. OPTIM.
Vol. 19, No. 2, pp. 911–940

DECENTRALIZED RESOURCE ALLOCATION IN DYNAMIC NETWORKS OF AGENTS*

HARIHARAN LAKSHMANAN† AND DANIELA PUCCI DE FARIAS‡

Abstract. We consider the problem of n agents that share m common resources. The objective is to derive an optimal allocation that maximizes a global objective expressed as a separable concave objective function. We propose a decentralized, asynchronous gradient-descent method that is suitable for implementation in the case where the communication between agents is described in terms of a dynamic network. This communication model accommodates situations such as mobile agents and communication failures. The method is shown to converge provided that the objective function has Lipschitz-continuous gradients. We further consider a randomized version of the same algorithm for the case where the objective function is nondifferentiable but has bounded subgradients. We show that both algorithms converge to near-optimal solutions and derive convergence rates in terms of the magnitude of the gradient of the objective function. We show how to accommodate nonnegativity constraints on the resources using the results derived. Experimental results with the problems of varying dimensions suggest that the algorithms are competitive with centralized approaches and scale well with problem size.

Key words. decentralized algorithms, resource allocation, nondifferentiable optimization

AMS subject classifications. 90C25, 91B32

DOI. 10.1137/060662228

1. Introduction.

We consider the problem of n agents that share m common resources. Agent i has utility function f_i. The optimal allocation of resources for maximizing the average of the utilities among agents is given by the following optimization problem:

$$\max_{\lambda_i \in \Re^m, i=1,\ldots,n} f(\lambda) = \frac{1}{n}\sum_{i=1}^{n} f_i(\lambda_i)$$

$$\text{such that (s.t.)} \sum_{i=1}^{n} \lambda_i = B, \tag{1}$$

where $B \in \Re^m$ corresponds to the total amount of resources.

We propose decentralized, asynchronous algorithms for a solution of (1). The first method applies in the case where $f_i, i = 1, \ldots, n$ are concave and differentiable with Lipschitz-continuous gradients. The second method applies in the case where $f_i, i = 1, \ldots, n$ are concave but not necessarily differentiable. We establish asymptotic convergence and convergence rates of both algorithms under mild conditions for communications among agents.

We assume that agents communicate through a network of dynamic topology in order to solve (1). At each iteration t, communication is represented by an undirected graph $G(t)$, where nodes correspond to agents and edges correspond to communication

*Received by the editors May 21, 2006; accepted for publication (in revised form) January 7, 2008; published electronically September 11, 2008. This research was partly supported by a Draper Laboratory grant and by NSF CAREER Award 0447766. A preliminary version of this work appears in the Proceedings of the American Control Conference 2006 [7].

http://www.siam.org/journals/siopt/19-2/66222.html

†Corresponding author. Department of Civil and Environmental Engineering, MIT, Cambridge, MA 02139 (lhari@mit.edu).

‡Department of Mechanical Engineering, MIT, Cambridge, MA 02139 (pucci@mit.edu).

links. We assume that communication is symmetric, so that if agent i communicates with agent j, then agent j also communicates with agent i. We also assume that the union of the communication graphs is connected over any sufficiently large, bounded period of time. This ensures that resource allocation to every agent is periodically influenced either directly or indirectly by the resource allocation to every other agent. This model of communication accommodates several practical scenarios, arising for instance, if agents are mobile and have limited communication range, or if communication links are subject to failure.

The decentralized algorithm for solving (1) in the case of differentiable utility functions has a simple gradient-descent structure. Starting with an initial feasible resource allocation, agents trade resources with their neighbors at each iteration in proportion to the difference in gradient for the respective utility functions. The algorithm has a natural interpretation. The local gradient computed by each agent can be thought of as the price the agent is willing to pay for additional resources. At each iteration, agents trade resources with their neighbors in proportion to the prices each is willing to pay for the resources.

It can be shown that a large class of separable convex optimization problems with linear constraints can be transformed to equivalent resource allocation problems. However, the functions f_i in the transformed resource allocation problem are usually not differentiable. Motivated by this setting, we consider the case where f_i is no longer differentiable but has bounded subgradients. It is shown in this case that a randomized version of a decentralized subgradient-descent algorithm converges with probability one to a near-optimal solution.

The subgradient-descent algorithm for the case of nondifferentiable utility functions can be interpreted as a stochastic approximation version of the gradient-descent method for differentiable functions applied to a smoothed version of the problem. The particular form of smoothing developed in this paper is motivated by several considerations. Adequate smoothing schemes must lead to a close approximation to the original function. Furthermore, as we build on the results for differentiable problems with a Lipschitz-continuous gradient, the gradient of the resulting smooth function must satisfy the same assumption with an adequate Lipschitz constant. Finally, another consideration in this paper is the computational effort involved in computing the gradient for the smoothed function. With this in mind, we propose a smooth approximation of the form $\hat{f}_i = \mathrm{E}[f_i(\lambda_i + Z_i)]$, where Z_i are vectors of zero-mean normal random variables. We show that, with an appropriate choice for the variance of Z_i, $\hat{f}_i$ is within ϵ of f_i, and its gradient is Lipschitz-continuous, with a Lipschitz constant on the order of $O(\sqrt{\log m}/\epsilon)$ so that it scales gracefully on the dimension m of variable λ_i. In addition, this form of smoothing lends itself to an application of a stochastic approximation scheme for gradient descent which, at each iteration, only requires an evaluation of a subgradient of f_i at a single point λ_i.

A comprehensive treatment of algorithms for various classes of resource allocation problems can be found in [13]. The algorithms introduced and analyzed in [13] are centralized in the sense that a central agent is assumed to have complete information about the problem and computes the optimal solution. In [1] and [5], decentralized resource allocation problems in the context of economics are investigated. The main difference in the approaches of [1, 5] as compared to the one presented here is the presence of a central agent who coordinates the computations performed by individual agents. A setting that is closer to ours is presented in [11], which introduces a completely decentralized algorithm for a resource allocation problem with twice differentiable separable convex objective functions. The algorithm assumes a sym-

metric and fixed communication graph for the agents at all iterations and performs a gradient projection at each iteration onto a subspace related to the communication graph. The same setting is considered in [8], which proposes a decentralized, weighted gradient algorithm for resource allocation problems with objective functions that are twice differentiable with bounded second derivatives. Dynamic communication graphs are considered in [6], which proposes an application-specific decentralized gradient algorithm for the problem of file allocation in distributed computer systems. Asynchronous gradient-descent methods are also considered in [14] for problems of unconstrained optimization with differentiable objective.

Most of the references regarding resource allocation problems in the literature, including the ones mentioned above, contain nonnegativity constraints on the resources (i.e., they require $\lambda_i \geq 0 \ \forall i$), whereas in our formulation resources may be negative. In Section 4, we show how the results in this paper can be applied to problems with nonnegativity constraints. The main motivation for problem (1) is that this formulation arises naturally in problems where a single, generic optimization problem must be solved in a decentralized way; this is the case, for instance, in problems of sequential decision making in teams of mobile agents as considered in [7].

A distributed algorithm for nondifferentiable optimization is presented in [9]. It is shown that a projected subgradient algorithm applied by each agent converges to the optimal solution. An important difference between the work presented in [9] and the work presented here is that the first requires that the long-run frequency of updates performed by each agent to be the same. Smoothing schemes for nondifferentiable optimization can also be found in the literature. [10] proposes a smoothing scheme for functions f_i described as the maximum of differentiable functions. The smoothed function is within ϵ of f_i and has Lipschitz constant on the order of $O(1/\epsilon)$, independent of the dimensions of the problem. A caveat of this approach is that computing the gradient of the smoothed function may require multiple evaluations of the subgradients of the original function. The particular form of smoothing considered here can also be found in the literature (see, e.g., [12]); however, we are unaware of results concerning the Lipschitz constant of the resulting smoothed function, which we develop in this paper.

The paper is organized as follows. In section 2, we describe the structure of communication among agents. In section 3, we introduce and we analyze the decentralized gradient-descent algorithm for problem (1) with differentiable objective functions that have Lipschitz-continuous gradients and its randomized version for problem (1) with nondifferentiable objective functions. In section 4 we describe a method to accommodate the nonnegativity constraints on the variables based on the results developed in section 3. In section 5, we present the results of numerical experiments, which illustrate the practical performance of the developed algorithms. In section 6, we conclude the paper. All proofs can be found in the appendix.

2. Communication between agents. In this section we describe the communication structure between agents. At iteration t, each agent i communicates with a set of agents denoted by $N_i(t)$. We assume that communication is symmetric; i.e., whenever agent i communicates with agent j, agent j also communicates with agent i. The communication between agents at time t can be represented by an undirected graph $G(t) = (N, E(t))$, where $N = \{1, \ldots, n\}$ represents the set of agents and the edge $(i, j) \in E(t)$ if and only if agent i communicates with agent j at time t. Let $E_{k,l} = \cup_{t=k}^{t=l-1} E(t)$. For a decentralized scheme to converge, the update of the variable associated with any agent must be periodically influenced by information from every other agent. This is ensured by the following assumption.

ASSUMPTION 2.1. *There exists a strictly increasing sequence* $\{T_z\}$ *of natural numbers, with* $T_1 = 1$ *such that* $G = (N, E_{T_z, T_{z+1}})$ *is connected for all* z *and* $(T_{z+1} - T_z) \leq \kappa$, *where* κ *is some natural number.*

3. Decentralized resource allocation. We assume that (1) has an optimal solution. Let $\lambda \in \Re^{nm} = (\lambda_1, \lambda_2, \ldots, \lambda_n)$, where $\lambda_i \in \Re^m$ for $i = 1, \ldots, n$.

ASSUMPTION 3.1. *There exists an optimal solution* $\lambda^* = (\lambda_1^*, \lambda_2^*, \ldots, \lambda_n^*)$ *to* (1).

For the rest of the paper, we let $\|\cdot\|$ denote the Euclidean norm.

3.1. The differentiable case. We now develop a decentralized algorithm for the case where f_i is concave and differentiable with a Lipschitz-continuous gradient.

ASSUMPTION 3.2. *There exists a constant* $L > 0$ *such that* $\|\nabla f_i(\lambda_i) - \nabla f_i(\bar{\lambda}_i)\| \leq L\|\lambda_i - \bar{\lambda}_i\|, \ \forall \lambda_i, \bar{\lambda}_i \in \Re^m$.

Recall that $f(\lambda) = \frac{1}{n}\sum_{i=1}^n f_i(\lambda_i)$. Hence

$$\begin{aligned}\|\nabla f(\lambda) - \nabla f(\bar{\lambda})\| &= \frac{1}{n}\sqrt{\sum_{i=1}^n \|\nabla f_i(\lambda_i) - \nabla f_i(\bar{\lambda}_i)\|^2} \\ &\leq \frac{1}{n}\sqrt{\sum_{i=1}^n L^2\|\lambda_i - \bar{\lambda}_i\|^2} \\ &= \frac{L}{n}\|\lambda - \bar{\lambda}\|.\end{aligned}$$

The second equality follows from the fact that $\|\lambda - \bar{\lambda}\| = \sqrt{\sum_{i=1}^n \|\lambda_i - \bar{\lambda}_i\|^2}$. Hence $\frac{L}{n}$ is a Lipschitz constant for the function f. The decentralized algorithm that we develop is based on the following lemma, which characterizes an optimal solution to (1) when functions f_i are all differentiable.

LEMMA 3.1. *A feasible solution* λ^* *of* (1) *is an optimal solution if and only if* $\nabla f_i(\lambda_i^*) = \nabla f_j(\lambda_j^*)$ *for all* i, j.

Let λ_i^t be the value of the variable associated with agent i at iteration t. We consider the following gradient-descent update rule for each agent i:

$$\lambda_i^{t+1} = \lambda_i^t + \gamma \sum_{j \in N_i(t)} \frac{1}{n}(\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)). \tag{2}$$

Here γ is a common constant step size that all of the agents use for updates. It should be noted that, to perform updates at iteration t, agent i uses only the gradient information corresponding to its neighbors $N_i(t)$ for iteration t. Furthermore, each intermediate allocation λ^t generated by the algorithm is a feasible solution of (1).

LEMMA 3.2. *Suppose* λ^1 *is a feasible solution for* (1). *Then* λ^t, *where* λ_i^t *is defined by* (2), *is a feasible solution to* (1) *for all* t.

In order to analyze the convergence properties of the proposed algorithm, it is convenient to define $\tilde{v}(\lambda)$ for any allocation λ as follows:

$$\tilde{v}_i(\lambda) = \sum_{j \in N} \frac{1}{n}(\nabla f_i(\lambda_i) - \nabla f_j(\lambda_j)).$$

Note that $\tilde{v}(\lambda^t)$ is the direction of update when the communication graph $E(t)$ is complete. It can be verified that $\tilde{v}(\lambda^t)$ is also a scaled version of the projection of $\nabla f(\lambda^t)$ onto the subspace $\sum_{i=1}^n \lambda_i^t = B$, hence it represents the centralized update

direction at time t. From Lemma 3.1 it can be seen that a feasible solution λ is optimal if and only if $\|\tilde{v}(\lambda)\| = 0$. We now derive a theorem establishing the convergence of the algorithm based on (2). Under mild conditions on the set of optimal solutions, convergence to optimality is guaranteed. We also derive an upper bound on the rate at which the sequence $\{\|\tilde{v}(\lambda^{T_z})\|\}$ converges to zero. Recall that T_z is a sequence of strictly increasing natural numbers such that the union of the communication graphs between iterations T_z and T_{z+1} is connected. In what follows, let $\tilde{v}^t = \tilde{v}(\lambda^t)$.

THEOREM 3.1. *Suppose that Assumptions* 3.1 *and* 3.2 *hold. With a step size of* $\gamma = \frac{1}{2L}$,

1. *the sequence* $\{f(\lambda^t)\}$ *is monotonically nondecreasing;*
2. *the sequence* $\{\|\tilde{v}^{T_z}\|\}$ *converges to* 0*;*
3. $\min_{z=1,\ldots,p}(\|\tilde{v}^{T_z}\|^2) \leq \frac{3Ln^4\kappa(f(\lambda^*)-f(\lambda^1))}{4p}$ $\forall p$*;*
4. *if the set of optima is bounded,* $\{f(\lambda^t)\}$ *converges to* $f(\lambda^*)$.

3.2. The nondifferentiable case. In this section we consider concave objective functions that are not required to be differentiable at all points. To motivate our interest in such functions we consider the following optimization problem:

$$\max_{x_i, i=1,\ldots,n} \frac{1}{n}\sum_{i=1}^{n} g_i(x_i)$$
$$\text{s.t.} \quad \sum_{i=1}^{n} \mathbf{A_i} x_i \leq B, \tag{3}$$

where $x_i \in \Re^q$, $\mathbf{A_i} \in \Re^{m\times q}, i = 1,\ldots,n$, $B \in \Re^m$, and $g_i(x_i)$ is a concave function, $i = 1,\ldots,n$. Define $f_i(\lambda_i)$ as the optimal value for the following optimization problem:

$$f_i(\lambda_i) = \max_{x_i \in \Re^q} g_i(x_i)$$
$$\text{s.t.} \quad \mathbf{A_i} x_i \leq \lambda_i. \tag{4}$$

With this definition of f_i we see that problem (3) is equivalent to problem (1). Note that, if there are linear constraints that involve only the variables $x_{ij}, j = 1,\ldots,q$ for some i, then these constraints could be included directly in the problem defining f_i. Suppose that $f_i(\lambda_i)$ is well defined and is finite for all λ_i. It can then be shown that $f_i(\lambda_i)$ is a concave function. Thus we can potentially apply the decentralized algorithm developed in the previous section for finding an optimal solution to (3). However, $f_i(\lambda_i)$ is typically nondifferentiable even when $g_i(x_i)$ is. Hence Theorem 3.1 does not immediately apply to (3), as it relies on the assumption that the objective function is differentiable with a Lipschitz-continuous gradient. This motivates us to consider cases where $f_i,\ i = 1,\ldots,n$ are not necessarily differentiable at all points.

In this section, we relax Assumption 3.2 and consider the case where $f_i,\ i = 1,\ldots,n$ are nondifferentiable. We introduce a smooth approximation for f_i that is amenable to optimization via stochastic approximations and propose a randomized version of (2) to solve the smoothed problem. We show that the new scheme converges to a near-optimal solution of the original problem in a tractable number of iterations.

We assume that $f_i,\ i = 1,\ldots,n$ are concave and differentiable outside a set of measure zero. Denote by $\partial f_i(\lambda_i)$ the set of the subgradients of f_i at λ_i. Let $\nabla f_i(\lambda_i)$ be an element chosen arbitrarily from $\partial f_i(\lambda_i)$ for each λ_i. Let $\|\cdot\|_1$ denote the l_1 norm and recall that $\|\cdot\|$ denotes the Euclidean norm. We make the following assumption.

ASSUMPTION 3.3. *For all* i *and* λ_i, $\sup_{i,\lambda_i}\{\|v\|_1 : v \in \partial f_i(\lambda_i)\} \leq L < \infty$.

Note that $\sup_{i,\lambda_i}\{\|v\| : v \in \partial f_i(\lambda_i)\} \leq L < \infty$, since $\|v\| \leq \|v\|_1$ for all v. We now consider approximating f_i by a suitable differentiable function. In particular, let

$$\hat{f}_i(\lambda_i) = \mathrm{E}[f_i(\lambda_i + Z_i)],$$

where each $Z_i = (Z_{ij})_{j=1,\ldots,m}$ is a vector of m independently and identically distributed (i.i.d.) normal random variables [4], with a zero mean and variance equal to

$$\sigma = \frac{\sqrt{2}\epsilon}{\sqrt{\pi \log(m+1)}},$$

where ϵ is a parameter related to the accuracy of the approximation as will be clear from the following lemma. The following lemma shows that $\hat{f}_i$ is a concave and differentiable approximation to f_i and that its gradient $\nabla \hat{f}_i$ can be expressed in terms of ∇f_i.

LEMMA 3.3. *Let f_i and $\hat{f}_i$ be as given above. Then the following hold:*

1. *$\hat{f}_i$ is concave and differentiable, with gradient $\nabla \hat{f}_i(\lambda_i) = \mathrm{E}[\nabla f_i(\lambda_i + Z_i)]$;*
2. *$f_i(\lambda_i) \geq \hat{f}_i(\lambda_i) \geq f_i(\lambda_i) - 2.8\epsilon L$;*
3. *$\|\nabla \hat{f}_i(\lambda_i) - \nabla \hat{f}_i(\bar{\lambda}_i)\| \leq \frac{\sqrt{\log(m+1)}L}{\epsilon}\|\lambda_i - \bar{\lambda}_i\|$.*

Bearing in mind the previous lemma, we consider the problem of maximizing

$$\max_{\lambda} \hat{f}(\lambda) = \sum_{i=1}^{n} \frac{1}{n} \hat{f}_i(\lambda_i) \tag{5}$$
$$\text{s.t.} \sum_{i=1}^{n} \lambda_i = B.$$

Since $\hat{f}_i$ is differentiable with a Lipschitz-continuous gradient, Theorem 3.1 ensures that the update rule (2) leads to convergence. However, note that computing the gradient of $\hat{f}_i$ requires evaluating the expected value $\nabla \hat{f}_i(\lambda_i) = \mathrm{E}[\nabla f_i(\lambda_i + Z_i)]$, which is, in general, computationally expensive. Due to the special form of the smoothing scheme and, in particular, the fact that $\nabla \hat{f}_i$ is expressed as the expected value of the subgradient of f_i, we consider instead of (2) a stochastic approximation version of the update. In particular, we let

$$\lambda_i^{t+1} = \lambda_i^t + \gamma_t \sum_{j \in N_i(t)} \frac{1}{n}(\nabla f_i(\lambda_i^t + Z_i^t) - \nabla f_j(\lambda_j^t + Z_j^t)), \tag{6}$$

where $Z_i^t,\ t = 1, 2, \ldots$ is a sequence of i.i.d. vectors with the same distribution as Z_i.

For each λ, let $\tilde{v}(\lambda)$ be given by

$$\tilde{v}_i(\lambda) = \sum_{j \in N} \frac{1}{n}(\nabla \hat{f}_i(\lambda_i) - \nabla \hat{f}_j(\lambda_j)).$$

Let $\tilde{v}^t = \tilde{v}(\lambda^t)$, and note that $\tilde{v}^t$ corresponds to the expected direction of update when the communication graph is complete. From Lemma 3.1, it is clear that a feasible solution λ is optimal for (5) if and only if $\|\tilde{v}(\lambda)\| = 0$. Furthermore, from Lemma 3.3, if λ is optimal for (5), then it is also near-optimal for (1). The following theorem establishes that, if all agents apply (6), then $\|\tilde{v}^t\|$ converges to zero.

We make the following assumption on the step sizes γ_t.

ASSUMPTION 3.4. *The step sizes* γ_t *satisfy* $\gamma_t = \frac{\epsilon}{(2L\sqrt{\log(m+1)})}\beta_t$, *where* $0 \leq \beta_{t+1} \leq \beta_t \leq 1 \forall t$, $\sum_t \beta_t = \infty$, *and* $\sum_t \beta_t^2 < \infty$.

THEOREM 3.2. *Suppose that Assumptions* 3.3 *and* 3.4 *hold. Then with probability* 1*:*

1. *the sequence* $\{\|\tilde{v}^{T_z}\|\}$ *converges to* 0*;*
2. $\min_{z=1,\ldots,p} \mathrm{E}[\|\tilde{v}^{T_z}\|^2] \leq \frac{\frac{n^4\kappa L\sqrt{\log(m+1)}}{\epsilon}\left[3(f(\lambda^*)-f(\lambda^1)+2.8\epsilon L)+\sum_{t=1}^{t=\kappa p}\frac{4L\beta_t^2\epsilon}{\sqrt{\log(m+1)}}\right]}{4\sum_{z=2}^{p+1}\beta_{\kappa z}}$ $\forall p$*;*
3. *if the set of the optima of* (1) *is bounded, then* $\lim_{t\to\infty} f(\lambda^t) \geq f(\lambda^*) - 2.8\epsilon L$.

It is worth noting some aspects of Theorem 3.2. Unlike in the differentiable case, we cannot guarantee a monotonic increase in the objective function values. Hence the rate of convergence of the sequence $\{\mathrm{E}[\|\tilde{v}^{T_z}\|]\}$ to zero does not have as far-reaching implications as its counterpart in Theorem 3.1. Nevertheless, Theorem 3.2 ensures convergence to a near-optimal solution with probability one. Another substantial difference is on the assumption on step sizes and the corresponding effect on convergence rates. It is easy to see that convergence is ensured if $\beta_t = \frac{1}{t^q}$ for $0.5 < q \leq 1$. The resulting theoretical rate of convergence is clearly dependent on q; when $0.5 < q < 1$, $\frac{1}{x^q}$ is a decreasing function for $x \geq 1$. Hence, for $k \geq 1, \frac{1}{k^q} \geq \int_k^{k+1} \frac{1}{x^q}dx$, and so $\sum_{k=1}^{c} \frac{1}{k^q} \geq \int_1^{c+1} \frac{1}{x^q}dx = \frac{(c+1)^{1-q}-1}{1-q}$. Thus the number of iterations needed for $\mathrm{E}[\|\tilde{v}^{T_z}\|^2] \leq \epsilon$ is polynomial in the problem parameters. Similarly, when $q = 1$, $\frac{1}{x^q}$ is just $\frac{1}{x}$ and is a decreasing function as well for $x \geq 1$. Hence, for $k \geq 1$, $\frac{1}{k+1} \leq \int_k^{k+1} \frac{1}{x}dx$, and so $\sum_{k=1}^{c} \frac{1}{k} \leq 1 + \int_1^c \frac{1}{x}dx = \log(c) + 1$, and so the number of iterations needed for $\mathrm{E}[\|\tilde{v}^{T_z}\|^2] \leq \epsilon$ is exponential in the problem parameters. As is often observed in stochastic approximation methods, the impact of the choice of step sizes on the speed of the convergence of the algorithm is also verified in the numerical experiments.

4. Decentralized resource allocation with nonnegativity constraints. In this section, we use the results developed for (1) to solve the following resource allocation problem with nonnegativity constraints:

$$\begin{aligned} \max_{\lambda_i \in \Re^m, i=1,\ldots,n} \quad & f(\lambda) = \frac{1}{n}\sum_{i=1}^{n} f_i(\lambda_i) \\ \text{s.t.} \quad & \sum_{i=1}^{n} \lambda_i = B, \\ & \lambda_i \geq 0, i = 1, \ldots, n. \end{aligned} \tag{7}$$

We assume that f_i is concave and differentiable outside a set of measure zero. Also let Assumption 3.3 hold for f.

We now define $g_i(\lambda_i)$ as follows:

$$g_i(\lambda_i) = f_i(\lambda_i) + \sum_{j=1}^{m} L_g \min(\lambda_{ij}, 0),$$

where $L_g > 2L$. The following lemma shows that the function $g(\lambda)$ satisfies Assumption 3.3 and is necessary for applying the stochastic approximation version of the gradient-descent algorithm developed in 3.2.

LEMMA 4.1. *Under assumption* 3.3 *for* f,

1. *for all* i, $g_i(\lambda_i)$ *is concave and differentiable outside a set of measure zero;*
2. *for all* i *and* λ_i, $\sup_{i,\lambda_i}\{\|v\|_1 : v \in \partial g_i(\lambda_i)\} \leq L_m < \infty$, *where* $L_m = L + mL_g$.

It can be noted from the definition of g_i that if $\lambda_i \geq 0$, then $g_i(\lambda_i) = f_i(\lambda_i)$. The term $L_g \min(\lambda_{ij}, 0)$ in the above definition can be thought of as a penalty for negative λ_{ij}. This term ensures that solving (1) with g has a nonnegative optimal solution and is equivalent to solving (7) with f.

LEMMA 4.2. *The set of optimal solutions for* (1) *with* g *as the objective function is the same as the set of optimal solutions to* (7) *with* f *as the objective function.*

Since the set of the feasible solutions of (7) is bounded and closed and since f is assumed to be continuous, there exists an optimal solution to (7). Thus any algorithm that finds an optimal solution to (1) with g as the objective function also yields an optimal solution to (7) with f as the objective function.

Lemma 4.1 ensures that we can apply the stochastic approximation version of the gradient-descent algorithm for (1) with g as the objective function. Hence an optimal solution for (7) with f as the objective function can be found by applying the stochastic approximation version of the gradient-descent algorithm developed in 3.2 for (1) with g as the objective function. It should be pointed out that the Lipschitz constant of the smoothed problem, and consequently the convergence rate, is now of the order $O(\frac{m\sqrt{m}}{\epsilon})$ as compared to $O(\frac{\sqrt{m}}{\epsilon})$ for the results of section 3.2.

5. Numerical experiments. In this section, we present the results of numerical experiments, which illustrate the performance of the algorithms presented in the previous sections. We compare the proposed algorithms to centralized algorithms that use direction $\tilde{v}(\lambda)$ as the direction of update. Recall that $\tilde{v}(\lambda)$ is the direction of update if the current resource allocation is λ and the communication graph is complete. Recall also that, when f_i is differentiable, $\tilde{v}(\lambda)$ is the projection of ∇f onto the subspace $\sum_{i=1}^{n} \lambda_i^t = B$. Thus the centralized algorithm reduces to the classic gradient-descent method of nonlinear optimization in this case. We define $p^t = (\frac{f^t - f^0}{f^* - f^0}) \times 100$, where f^t is the objective function value after t iterations and f^* is the objective function value of the optimal solution, and we investigate how p^t converges to 100 in the centralized and decentralized algorithms.

5.1. Problem with differentiable objective function. We first consider a problem studied in [8], which is an instance of (1), with

$$f_i(x_i) = -\left(\frac{1}{2}a_i(x_i - c_i)^2 + \log(1 + e^{b_i(x_i - d_i)})\right), i = 1, \ldots, n.$$

The second derivative $f_i^{''}$ is given by

$$f_i^{''}(x_i) = -\left(a_i + b_i^2 \frac{e^{b_i(x_i - d_i)}}{(1 + e^{b_i(x_i - d_i)})^2}\right), \qquad i = 1, \ldots, n.$$

It can be verified that $f_i^{''}(x_i)$ has a lower bound $-(a_i + \frac{1}{4}b_i^2)$, $i = 1, \ldots, n$. It can be shown that, if a one-dimensional function is differentiable and its gradient is bounded by some constant, then the function is Lipschitz-continuous with the same constant. Since f_i is twice differentiable and $f_i^{''}$ is bounded, it follows that $f_i^{'}$ is Lipschitz-continuous, with constant $(a_i + \frac{1}{4}b_i^2)$, if we assume that $a_i \geq 0$. It follows that $f^{'}$ is Lipschitz-continuous, with constant $\frac{L}{n}$, where $L = \max_i(a_i + \frac{1}{4}b_i^2)$. Thus f satisfies Assumption 3.2.

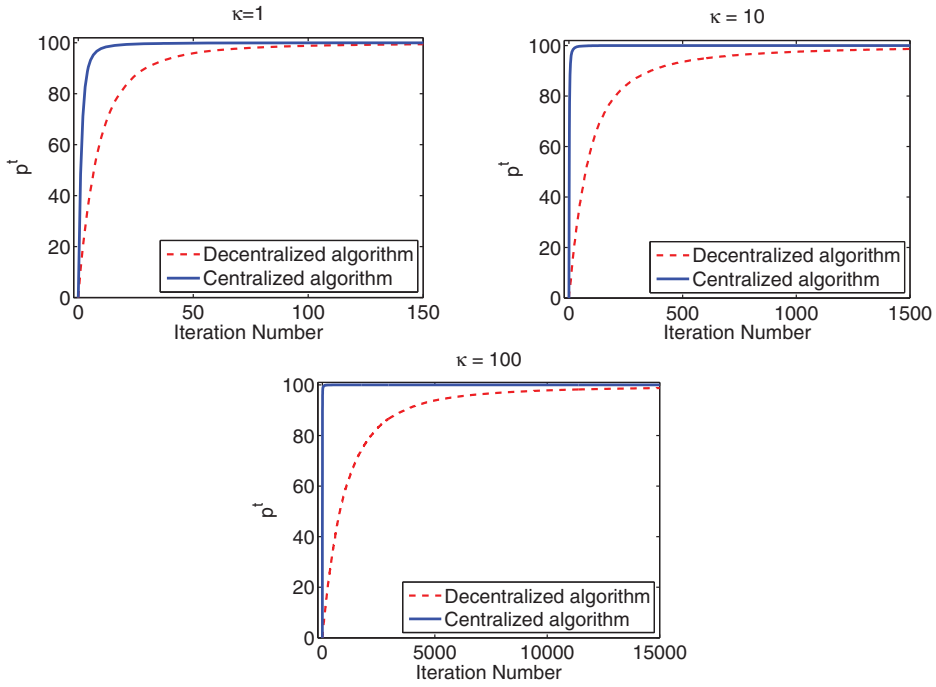


FIG. 1. *A comparison of the convergence behavior of the decentralized and centralized algorithms for various* κ.

We choose problem instances with 20 agents and as in [8]; the coefficients a_i, b_i, c_i, and d_i are generated randomly, with uniform distributions on $[0, 2], [-2, 2], [-10, 10]$, and $[-10, 10]$, respectively. Recall that, for our algorithm to converge, the union of communication graphs should be connected periodically. For a chosen κ, we let the edges $(i, i+1), i = 1, \ldots, n-1$ be a part of the communication graph $E(t)$ for some arbitrarily chosen t such that $m\kappa < t \leq (m+1)\kappa, m = 0, 1, \ldots$. This ensures that $G = (N, E_{m\kappa+1, m(\kappa+1)+1})$ is connected (recall that $E_{k,l} = \cup_{t=k}^{t=l-1} E(t)$). We let every other edge (i, j), with $j \neq i+1$, be a part of at the most one communication graph between iterations $m\kappa + 1$ and $(m+1)\kappa$, with a probability e_p. The parameter e_p controls the density of the graph, $G = (N, E_{m\kappa+1, m(\kappa+1)+1})$. The step size is chosen to be $\frac{1}{2L}$, with L as defined above. Figure 1 shows the convergence behavior of the algorithm for various values of the parameter κ, with $e_p = 0.1$. p^t in the figure represents the average of p^t for 10 randomly chosen problems. It can be seen from the figure that the performance of the decentralized algorithm is comparable to the centralized algorithm for $\kappa = 1$, even though the communication graph is not dense $(e_p = 0.1)$.

Figure 2 shows a comparison of the convergence behavior of the algorithms for problems with a varying number of agents. We fix $\kappa = 1$ in these problems, and $e_p = 0.1$. The other parameters are chosen as described above. We notice from Figure 2 that the scaling of the performance of decentralized algorithms with increasing number of agents is much better than $O(n^4)$ promised by Theorem 3.1.

5.2. Decentralized optimization of linear programming problems. We now consider a decentralized solution of linear programming problems using the randomized version of the decentralized subgradient-descent algorithm developed in this

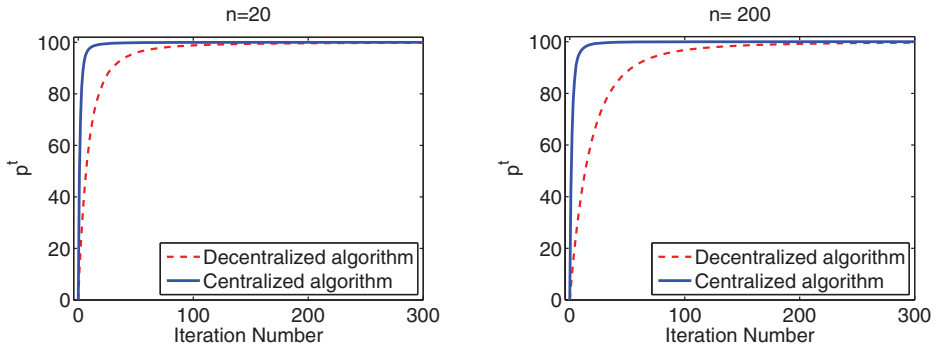


FIG. 2. *A comparison of the convergence behavior of the decentralized and centralized algorithms for various n.*

paper. This problem arises naturally in sequential decision-making problems in teams of mobile agents [7]. Consider the following linear programming problem:

$$\max_{x_i, i=1,\ldots,n} \frac{1}{n} \sum_{i=1}^{n} C_i^T x_i$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \mathbf{A_i} x_i \leq B, \tag{8}$$

where $C_i, x_i \in \Re^q$, $\mathbf{A_i} \in \Re^{m \times q}, i = 1, \ldots, n$, and $B \in \Re^m$. It can be seen that (8) belongs to the class of problems identified by (3). Recall that, for a given $\lambda_i \in \Re^m$, $f_i(\lambda_i)$ is the optimal value of the following optimization problem:

$$\max_{x_i \in \Re^q} C_i^T x_i$$

$$\text{s.t.} \quad \mathbf{A_i} x_i \leq \lambda_i. \tag{9}$$

Suppose that the dual feasible sets defined by $S_i = \{\nu_i | \mathbf{A_i^T} \nu_i = C_i, \nu_i \geq 0\}$ are nonempty and bounded. It is known from linear programming theory that $f_i(\lambda_i) = \min_{p=1,\ldots,P} \lambda_i^T \nu_{ip}$, where ν_{ip} are the extreme points of the polyhedra defined by S_i. Hence $f_i(\lambda_i)$ is nondifferentiable and concave. Further ν_i' is a subgradient of $f_i(\lambda_i)$ at λ_i if and only if it is an optimal solution to the dual problem [3]. Thus if S_i is bounded, it can be seen that Assumption 3.3 is satisfied, and the convergence analysis of section 3.2 holds.

Let the columns of $\mathbf{A_i}$ be denoted as $\mathbf{a_{ij}}$, $j = 1, \ldots, q$. Also let $C_i = [C_{ij}], j = 1, \ldots, q$. Suppose that the column $\mathbf{a_{ik}} > 0$ and $C_{ik} > 0$ for some k such that $1 \leq k \leq q$, and suppose S_i is nonempty. The corresponding dual constraint is $\mathbf{a_{ik}}^T \nu_i = C_{ik}$ showing that S_i is bounded. For the experiments we choose $\mathbf{a_{i1}} = \mathbf{1}$, $i = 1, \ldots, n$, where $\mathbf{1}$ is a vector of ones of the appropriate size. We also choose $C_{i1} = 200$, $i = 1, \ldots, n$. The rest of the constraint matrix and the cost vector are chosen arbitrarily while ensuring that S_i is nonempty.

Although the theoretical results require a randomization of the direction of update, it was observed that both of the decentralized and the centralized versions of the algorithm converge without the required randomization. Unlike the decentralized algorithm for the differentiable case, there is flexibility in choosing step sizes. It was observed in the experiments that the practical performance of both the centralized

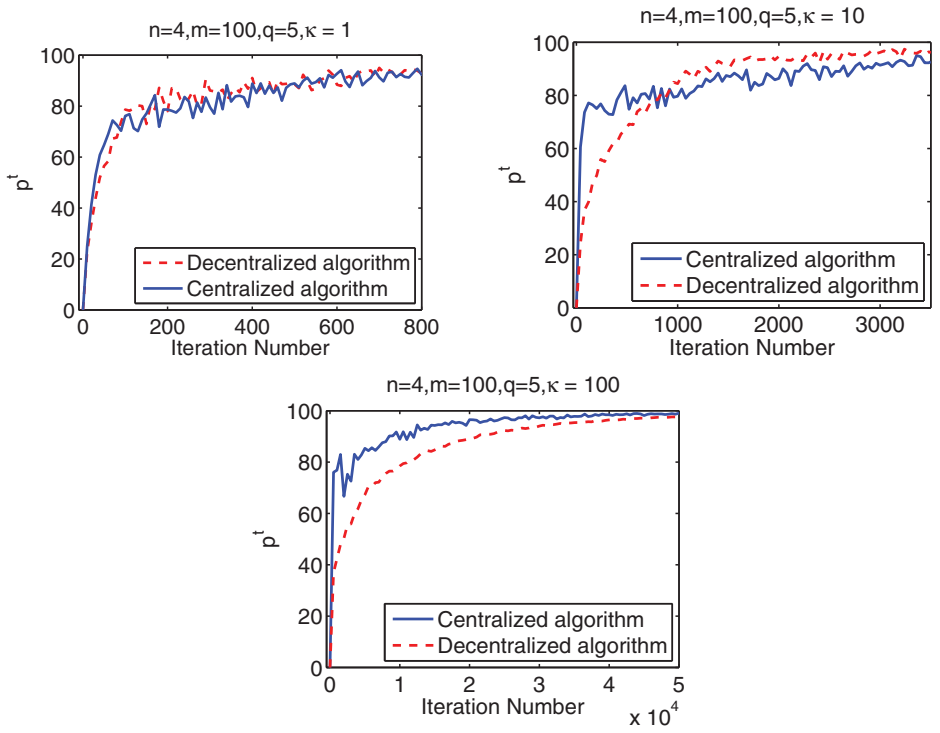


FIG. 3. *A comparison of the convergence behavior of the decentralized and centralized algorithms for various* κ.

algorithm and the decentralized algorithm, with or without the randomization of the direction of update, depends dramatically on the choice of step sizes. We present the results of the experiments where the direction of update was not randomized, as it provides better insight into the convergence behavior of the algorithm. It was observed that convergence was obtained in this case so long as $\sum_t \gamma_t = \infty$ and $\sum_t \gamma_t^2 < \infty$. We choose step sizes of the form $\gamma^t = \frac{\theta(t)}{2L\sqrt{\log m+1}}\beta^t$, where L is the common Lipschitz constant of the functions f_i, $i = 1, \ldots, n$. Since $C_{i1} = 200$ and $\mathbf{a_{i1}} = \mathbf{1}$ for all i, it can be verified from the dual constraint $\mathbf{a_{i1}}^T \nu_i = C_{i1}$ that $L = C_{i1} = 200$. β^t was chosen to be of the form $\frac{1}{1+w(t)t^{0.51}}$. Thus $\theta(t)$ and $w(t)$ control the rate at which γ^t goes to 0. We chose $w(t)$ as a monotonically nondecreasing function bounded above, and $\theta(t)$ as a monotonically nonincreasing function bounded below. This ensures that $\sum_t \gamma_t = \infty$ and $\sum_t \gamma_t^2 < \infty$. For our experiments, we chose $w(0) = 0$ and $w(z\kappa + j) = w(z\kappa)$ for $z = 0, 1, \ldots, j = 1, 2, \ldots, \kappa - 1$, and $w((z+1)\kappa) = \min\{w(z\kappa) + r_w, w_{\max}\}$. For all of the experiments we chose $r_w = 0.0001$ and $w_{\max} = 0.1$. We also chose $\theta(t+1) = \max\{\theta(t) - r_\theta, \theta_{\min}\}$. For these experiments, we chose $\theta(0) = 30$, $\theta_{\min} = 3$, and $r_\theta = 0.1$. We ensured that the union of the communication graphs are connected periodically in the same manner as described in section 5.1. For these experiments, we choose $e_p = 0.5$. Figure 3 presents a comparison of the performance of the decentralized algorithm with the centralized algorithm, for varying κ. In the figures, n represents the number of agents, q represents the number of variables per agent, and m represents the number of constraints.

Figure 4 presents a comparison of the performance of the decentralized algorithm, with the centralized algorithm for varying n. All parameters except $\theta(0)$ were chosen

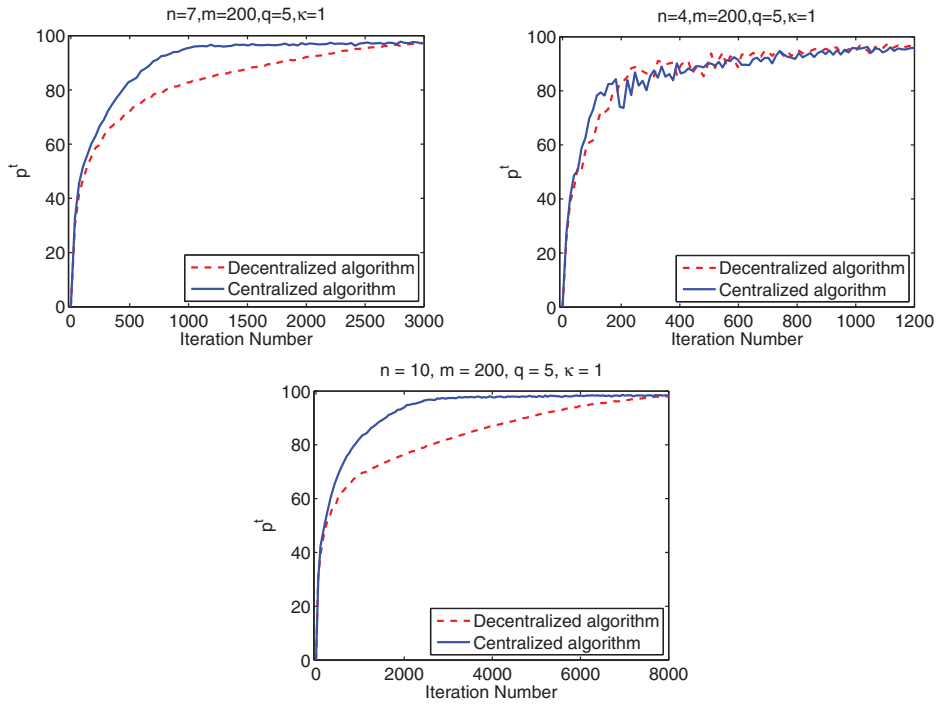


FIG. 4. *A comparison of the convergence behavior of the decentralized and centralized algorithms for various* n.

as described previously. $\theta(0)$ was chosen to be 50 for the experiments of Figure 4. It can be observed that the performance of the decentralized algorithm scales well with increasing n. The numerical experiments suggest that κ has a greater effect on the practical performance of the algorithm than n.

6. Discussion. In this paper, we proposed a decentralized gradient-descent algorithm for a general class of resource allocation problems. We first considered the case where the objective functions have Lipschitz-continuous gradients. We showed that the proposed algorithm converges and established that the rate at which the gradient projection converges to zero as a function of the number of agents in the network. Motivated by the need to develop decentralized algorithms for general convex optimization problems, we proposed a randomized subgradient-descent algorithm for the resource allocation problem with a possibly nondifferentiable objective function. We established an asymptotic convergence of the algorithm to a near-optimal solution and derived a convergence rate. Numerical experiments, both in the differentiable and in the nondifferentiable settings, suggested that the decentralized algorithms are competitive with the centralized versions of gradient and subgradient descent. The experiments also suggested that the performance of the decentralized randomized subgradient-descent algorithm depends dramatically on the choice of step sizes; how to set them up optimally while taking into account the structure of the communication network is a topic for future research.

An appealing feature of the developed algorithms is that the communication topology of the network of agents is allowed to be dynamic provided that the union of communication graphs of the agents is connected within a bounded time. This makes the algorithm particularly suitable in settings involving mobile agents or communi-

cation failures. We finally note that the formulation considered here goes beyond traditional resource allocation problems. In particular, we show in a related paper [7] that this algorithm is particularly suitable for a decentralized solution of a linear programming-based method for approximate dynamic programming for the problems of sequential decision making in the systems of mobile agents.

Appendix A. Proofs.

LEMMA 3.1. *A feasible solution λ^* of* (1) *is an optimal solution if and only if $\nabla f_i(\lambda_i^*) = \nabla f_j(\lambda_j^*)$ for all i, j.*

Proof. First note that we can eliminate one of the variables in (1) to make it unconstrained. For instance, if we let $\lambda_n = B - \sum_{i=1}^{n-1} \lambda_i$, (1) is equivalent to

$$\min_{\lambda} \bar{f}(\lambda) = \frac{1}{n} \sum_{i=1}^{n-1} \left(f_i(\lambda_i) + f_n \left(B - \sum_{i=1}^{n-1} \lambda_i \right) \right).$$

This is an unconstrained convex and differentiable optimization problem; hence a solution λ^* is optimal if and only if $\nabla \bar{f}(\lambda^*) = 0$. Noting that

$$\nabla_{\lambda_i} \bar{f}(\lambda^*) = \frac{1}{n} \left(\nabla f_i(\lambda_i^*) - \nabla f_n \left(B - \sum_{j=1}^{n-1} \lambda_j^* \right) \right),$$

we conclude that λ^* is optimal if and only if

$$\nabla f_i(\lambda_i^*) = \nabla f_j(\lambda_j^*) = \nabla f_n \left(B - \sum_{j=1}^{n-1} \lambda_j^* \right) = \nabla f_n(\lambda_n^*) \ \forall i, j < n. \qquad \square$$

LEMMA 3.2. *Suppose λ^1 is a feasible solution for* (1). *Then λ^t, where λ_i^t is defined by* (2), *is a feasible solution to* (1) *for all t.*

Proof. Suppose λ^t is a feasible solution for (1). Then

$$\begin{aligned} \sum_i \lambda_i^{t+1} &= \sum_i \lambda_i^t - \frac{\gamma}{n} \sum_i \sum_{j \in N_i(t)} (\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)) \\ &= B - \frac{\gamma}{n} \sum_{(i,j) \in E(t)} (\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t) + \nabla f_j(\lambda_j^t) - \nabla f_i(\lambda_i^t)) \\ &= B. \end{aligned}$$

The second equality follows from the assumption that communication is symmetric. Thus λ^{t+1} is a feasible solution for (1), and the lemma follows by induction. $\square$

THEOREM 3.1. *Suppose that Assumptions* 3.1 *and* 3.2 *hold. With a step size of $\gamma = \frac{1}{2L}$,*

1. *the sequence $\{f(\lambda^t)\}$ is monotonically nondecreasing;*
2. *the sequence $\{\|\tilde{v}^{T_z}\|\}$ converges to* 0*;*
3. $\min_{z=1,\ldots,p}(\|\tilde{v}^{T_z}\|^2) \leq \frac{3Ln^4\kappa(f(\lambda^*)-f(\lambda^1))}{4p} \ \forall p$*;*
4. *if the set of optima is bounded, $\{f(\lambda^t)\}$ converges to $f(\lambda^*)$.*

The proof is based on a series of lemmas. Let the direction of update at time t be v^t. It can be seen from (2) that

$$v_i^t = \frac{1}{n} \sum_{j \in N_i(t)} (\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)).$$

We first show that v^t is aligned to the direction of the gradient.

LEMMA A.1. $\nabla f(\lambda^t)^T v^t = \frac{1}{n^2} \sum_{(i,j)\in E(t)} \|\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)\|^2$.

Proof. We have that

$$\begin{aligned}\nabla f(\lambda^t)^T v^t &= \sum_{i\in N} \frac{1}{n} \nabla f_i(\lambda_i^t)^T \left(\frac{1}{n} \sum_{j\in N_i(t)} \nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t) \right) \\ &= \frac{1}{n^2} \sum_{i\in N} \nabla f_i(\lambda_i^t)^T \left(\sum_{j\in N_i(t)} \nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t) \right). \end{aligned} \tag{10}$$

Since communication is symmetric, for every term of the form $\nabla f_i(\lambda_i^t)^T(\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t))$ in the above summation, there is a corresponding term of the form $\nabla f_j(\lambda_j^t)^T (\nabla f_j(\lambda_j^t) - \nabla f_i(\lambda_i^t))$. Hence,

$$\begin{aligned}\nabla f(\lambda^t)^T v^t &= \frac{1}{n^2} \sum_{(i,j)\in E(t)} \nabla f_i(\lambda_i^t)^T (\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)) \\ &\quad + \nabla f_j(\lambda_j^t)^T (\nabla f_j(\lambda_j^t) - \nabla f_i(\lambda_i^t)) \\ &= \frac{1}{n^2} \sum_{(i,j)\in E(t)} \|\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)\|^2. \qquad \square \end{aligned}$$

We now prove a lemma that establishes a relationship between $\|v^t\|$ and $\nabla f(\lambda^t)^T v^t$. We can interpret $\gamma \nabla f(\lambda^t)^T v^t$ as the approximate increase in the objective of (1) when using the direction v^t and a sufficiently small step size γ.

LEMMA A.2. $\|v^t\|^2 \le 2n \nabla f(\lambda^t)^T v^t$.

Proof. Using the Cauchy–Schwarz inequality, $(\sum_{i=1}^k c_i)^2 \le k \sum_{i=1}^k c_i^2$,

$$\begin{aligned}\|v_i^t\|^2 &\le |N_i(t)| \sum_{j\in N_i(t)} \frac{\|\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)\|^2}{n^2} \\ &\le n \sum_{j\in N_i(t)} \frac{\|\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)\|^2}{n^2}, \\ \|v^t\|^2 &= \sum_i \|v_i^t\|^2 \\ &\le n \sum_i \sum_{j\in N_i(t)} \frac{\|\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)\|^2}{n^2} \\ &= 2n \sum_{(i,j)\in E(t)} \frac{\|\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)\|^2}{n^2} \\ &= 2n \nabla f(\lambda^t)^T v^t. \end{aligned}$$

The last equality comes from Lemma A.1. $\square$

We now prove a lemma that establishes a relationship between $\|\tilde{v}^t\|$ and $\nabla f(\lambda^t)^T \tilde{v}^t$.

LEMMA A.3. $\|\tilde{v}^t\|^2 = n \nabla f(\lambda^t)^T \tilde{v}^t$.

Proof. We first have that

$$\begin{aligned}
\|\tilde{v}_i^t\|^2 &= \frac{\|\sum_{j\in N}(\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t))\|^2}{n^2} \\
&= \sum_{j\in N} \frac{\|\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)\|^2}{n^2} \\
&\quad + 2 \sum_{((j,l)\in N^2, j<l)} \frac{(\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t))^T(\nabla f_i(\lambda_i^t) - \nabla f_l(\lambda_l^t))}{n^2} \\
&= \sum_{j\in N} \frac{\|\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)\|^2}{n^2} \\
&\quad + \sum_{((j,l)\in N^2, j<l)} \\
&\quad \frac{\|\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)\|^2 + \|\nabla f_i(\lambda_i^t) - \nabla f_l(\lambda_l^t)\|^2 - \|\nabla f_j(\lambda_j^t) - \nabla f_l(\lambda_l^t)\|^2}{n^2} \\
&= (n-1)\left(\sum_{j\in N} \frac{\|(\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t))\|^2}{n^2}\right) \\
&\quad - \sum_{((j,l)\in N^2, j<l, (j,l\neq i))} \frac{\|\nabla f_j(\lambda_j^t) - \nabla f_l(\lambda_l^t)\|^2}{n^2}, \\
\|\tilde{v}^t\|^2 &= \sum_{i\in N} \|\tilde{v}_i^t\|^2 \\
&= \sum_{i\in N}\left((n-1)\left(\sum_{(j\in N)} \frac{\|(\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t))\|^2}{n^2}\right)\right. \\
&\qquad \left. - \sum_{((j,l)\in N^2, j<l, j,l\neq i)} \frac{\|\nabla f_j(\lambda_j^t) - \nabla f_l(\lambda_l^t)\|^2}{n^2}\right).
\end{aligned}$$

We note that $\|\tilde{v}^t\|^2 = \sum_{((i,j)\in N^2, i<j)} c_{ij}((\|\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)\|^2)/n^2)$. To determine c_{ij}, note that the term $\|\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)\|^2$ appears with a coefficient $(n-1)$ in $\|\tilde{v}_i^t\|^2$, $(n-1)$ in $\|\tilde{v}_j^t\|^2$, and with a coefficient -1 in $\|\tilde{v}_k^t\|^2$ for all $(k \in N, k \neq i, j)$. Hence, $c_{ij} = (n-1) + (n-1) - (n-2) = n$. Therefore,

$$\begin{aligned}
\|\tilde{v}^t\|^2 &= \sum_{((i,j)\in N^2, i<j)} c_{ij}\left(\frac{\|\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)\|^2}{n^2}\right) \\
&= n \sum_{((i,j)\in N^2, i<j)} \frac{\|\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t)\|^2}{n^2} \\
&= n\nabla f(\lambda^t)^T\tilde{v}^t. \qquad \square
\end{aligned}$$

Consider a decentralized direction of update v^t derived from an arbitrary connected graph $G = (N, E(t))$. We now compare the ratio of the approximate increase in the objective of (1) using v^t as the direction of update and for a sufficiently small step size γ to the approximate increase in the objective using $\tilde{v}^t$ as the direction of

update for the same step size. This ratio is given by

$$\frac{(\nabla f(\lambda^t)^T v^t)}{(\nabla f(\lambda^t)^T \tilde{v}^t)} = \frac{\sum_{(i,j)\in E(t)}(\|\nabla f_j(\lambda_j^t) - \nabla f_i(\lambda_i^t)\|^2)}{\sum_{((i,j)\in N^2, i<j)}(\|\nabla f_j(\lambda_j^t) - \nabla f_i(\lambda_i^t)\|^2)}. \tag{11}$$

The following lemma shows that this ratio is bounded away from 0 by a factor that depends only on the number of agents.

LEMMA A.4. *For all connected graphs $G = (N, E)$,*

$$\sum_{(i,j)\in E} \|\nabla f_j(\lambda_j^t) - \nabla f_i(\lambda_i^t)\|^2 \geq \frac{8}{n^3} \sum_{((i,j)\in N^2, i<j)} \|\nabla f_j(\lambda_j^t) - \nabla f_i(\lambda_i^t)\|^2.$$

For any vector X, let $(X)_k$ denote its kth component. We note that $(\sum_{(i,j)\in E} \|\nabla f_j(\lambda_j^t) - \nabla f_i(\lambda_i^t)\|^2)/(\sum_{((i,j)\in N^2, i<j)} \|\nabla f_j(\lambda_j^t) - \nabla f_i(\lambda_i^t)\|^2)$ is of the form $(\sum_{k=1}^m b_k)/(\sum_{k=1}^m c_k)$, where $b_k = \sum_{(i,j)\in E}((\nabla f_j(\lambda_j^t))_k - (\nabla f_i(\lambda_i^t))_k)^2$ and $c_k = \sum_{((i,j)\in N^2, i<j)}((\nabla f_j(\lambda_j^t))_k - (\nabla f_i(\lambda_i^t))_k)^2$, and we recall that $\lambda_j^t \in \Re^m$ for all $j \in N$. Let $r_k = \frac{b_k}{c_k}$. We show that if $c_k > 0$, then $r_k \geq \frac{8}{n^3}$. We define $r(E) = (\sum_{(i,j)\in E}(p_j - p_i)^2)/(\sum_{((i,j)\in N^2, i<j)}(p_j - p_i)^2)$ for arbitrary values of the scalars p_i, $i = 1, \ldots, n$, such that $\sum_{((i,j)\in N^2, i<j)}(p_j - p_i)^2 > 0$. We show that $r(E) \geq \frac{8}{n^3}$, which establishes that when $c_k > 0$, $r_k \geq \frac{8}{n^3}$. This result is based on a series of lemmas. We first establish that, for any fixed value of $p_i, i = 1, \ldots, n$, the worst possible value of r is achieved when G corresponds to a chain whose nodes have monotone values of p_i. Then we compute the worst possible value of r with respect to possible values of p_i.

We also assume that $p_i \neq p_j$ for all $i \neq j$, without loss of generality; since $\sum_{((i,j)\in N^2, i<j)}(p_j - p_i)^2 > 0$ by assumption, for any set of values $p_i, i = 1, \ldots, n$, we can always perturb the values to make them strictly distinct while making $r(E)$ in the resulting graph arbitrarily close to that in the original problem.

LEMMA A.5. *The graph $G = (N, E)$ that minimizes r over all possible sets E, under the constraint that G is a connected graph, is a tree.*

Proof. Take an arbitrary graph (N, E), and suppose that it is not a tree. Then we can convert it into a tree (N, E') by removing some edges from E. It is clear that $r(E') \leq r(E)$, therefore (N, E) cannot be optimal. ∎

LEMMA A.6. *If a certain graph (N, E) contains edges ij and jk such that $p_j < \min(p_i, p_k)$ or $p_j > \max(p_i, p_k)$, then it does not minimize r.*

Proof. Consider the first situation and suppose, without loss of generality, that $p_j < p_i < p_k$. Let $E' = E \backslash \{jk\} \cup \{ik\}$. The difference in the numerator of $r(E)$ and $r(E')$ is equal to $(p_j - p_k)^2 - (p_i - p_k)^2$, which is greater than 0. Therefore (N, E) cannot be optimal. A similar analysis holds when $p_j > \max(p_i, p_k)$. ∎

LEMMA A.7. *If a node j contains more than two neighbors, then it has two neighbors i and k such that $p_j < \min(p_i, p_k)$ or $p_j > \max(p_i, p_k)$.*

Proof. Suppose that i, k, and l are neighbors of j. Then at least two among the three values p_i, p_k, and p_l must be less than or greater than p_j. ∎

LEMMA A.8. *Consider the chain that links nodes $1, \ldots, n$ in increasing order of p_i. Then it minimizes r over all possible connected graphs.*

Proof. From the previous lemmas, we conclude that the optimal graph is a tree. Moreover, each node in the optimal tree must have at most two neighbors. We conclude that the optimal graph is a chain. From Lemma A.6, the nodes in the chain are in increasing or decreasing order of p_i, and the lemma follows. ∎

Proof of Lemma A.4. Without loss of generality, suppose that $p_1 < p_2 < \cdots < p_n$. Let $\Delta_i = p_{i+1} - p_i$. Note that, for all $j > i$, $p_j - p_i = \sum_{k=i}^{j-1} \Delta_k$. In view of the previous

lemmas, we have the following for every connected graph (N, E):

$$\begin{aligned}
r(E) &= \frac{\sum_{(i,j)\in E}(p_i - p_j)^2}{\sum_{((i,j)\in N^2, i<j)}(p_i - p_j)^2} \\
&\geq \frac{\sum_{i=1}^{n-1}(p_{i+1} - p_i)^2}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(p_i - p_j)^2} = \frac{\sum_{i=1}^{n-1}\Delta_i^2}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\sum_{k=i}^{j-1}\Delta_k\right)^2} \\
&\geq \frac{\sum_{i=1}^{n-1}\Delta_i^2}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sum_{k=i}^{j-1}(j-i)\Delta_k^2} \\
&= \frac{\sum_{i=1}^{n-1}\Delta_i^2}{\sum_{i=1}^{n-1}\sum_{k=i}^{n-1}\sum_{j=k+1}^{n}(j-i)\Delta_k^2} \\
&= \frac{\sum_{i=1}^{n-1}\Delta_i^2}{\sum_{k=1}^{n-1}\Delta_k^2\sum_{i=1}^{k}\sum_{j=k+1}^{n}(j-i)} = \frac{\sum_{i=1}^{n-1}\Delta_i^2}{\sum_{k=1}^{n-1}\Delta_k^2\sum_{i=1}^{k}\sum_{j=1}^{n-k}(j+k-i)} \\
&= \frac{\sum_{i=1}^{n-1}\Delta_i^2}{\sum_{k=1}^{n-1}\Delta_k^2\sum_{i=1}^{k}\frac{(n-k)(n-k+1)}{2} + (k-i)(n-k)} \\
&= \frac{\sum_{i=1}^{n-1}\Delta_i^2}{\sum_{k=1}^{n-1}\Delta_k^2\left(\frac{k(n-k)(n-k+1)}{2} + \frac{(n-k)(k)(k-1)}{2}\right)} \\
&= \frac{\sum_{i=1}^{n-1}\Delta_i^2}{\sum_{k=1}^{n-1}\Delta_k^2\frac{k(n-k)(n)}{2}} \\
&\geq \frac{\sum_{i=1}^{n-1}\Delta_i^2}{\sum_{k=1}^{n-1}\Delta_k^2\frac{n^3}{8}} \\
&= \frac{8}{n^3}.
\end{aligned}$$

The second inequality follows from the Cauchy–Schwarz inequality.

We note from the definitions that if $c_k = 0$, then $b_k = 0$. Thus for $k = 1, \ldots, m$, either $b_k = c_k = 0$, or $r_k \geq \frac{8}{n^3}$. The Lemma is trivially true, if for $k = 1, \ldots, m$, $b_k = c_k = 0$. Suppose there exist some $\bar{k} \in (1, \ldots, m)$ such that $c_{\bar{k}} > 0$. Let K be the set of integers from 1 to m such that $c_k > 0$ for $k \in K$. K is not empty since it contains $\bar{k}$.

$$\begin{aligned}
\frac{\sum_{(i,j)\in E}(\|\nabla f_j(\lambda_j^t) - \nabla f_i(\lambda_i^t)\|^2)}{\sum_{((i,j)\in N^2, i<j)}(\|\nabla f_j(\lambda_j^t) - \nabla f_i(\lambda_i^t)\|^2)} &= \frac{\sum_{k=1}^{m} b_k}{\sum_{k=1}^{m} c_k} \\
&= \frac{\sum_{k\in K} b_k}{\sum_{k\in K} c_k} \\
&\geq \frac{\sum_{k\in K}\frac{8}{n^3}c_k}{\sum_{k\in K} c_k} \\
&= \frac{8}{n^3}. \qquad \square
\end{aligned}$$

Let E_{T_z} be a subset of the edge set $E_{T_z, T_{z+1}}$ such that the graph (N, E_{T_z}) is a tree. By Assumption (2.1), the graph $(N, E_{T_z, T_{z+1}})$ is connected, and so E_{T_z} is well defined. Let the decentralized direction of update derived using $G = (N, E_{T_z})$ be denoted by $\bar{v}^{T_z}$. The following lemma shows that the approximate increase in the

objective in period $[T_z, T_{z+1}]$ using the direction of update v^t and a sufficiently small step size γ is comparable to the approximate increase in objective when the direction $\bar{v}^{T_z}$ is used for update at time T_z.

LEMMA A.9. $\nabla f(\lambda^{T_z})^T \bar{v}^{T_z} \leq \frac{3}{2}\kappa \sum_{t=T_z}^{T_{z+1}-1} \nabla f(\lambda^t)^T v^t$.

Proof. We have that

$$\begin{aligned}\|\nabla f(\lambda^{t+1}) - \nabla f(\lambda^t)\|^2 &\leq \frac{L^2}{n^2}\|\gamma v^t\|^2 = \frac{1}{4n^2}\|v^t\|^2 \\ &\leq \frac{1}{4n^2} 2n(\nabla f(\lambda^t)^T v^t) = \frac{1}{2n}\nabla f(\lambda^t)^T v^t. \end{aligned} \tag{12}$$

The first inequality is true because of Assumption 3.2. The first equality is true because $\gamma = \frac{1}{2Ln}$. The second inequality follows from Lemma A.2. Let $t^i_{T_z}$ be the earliest time between time periods T_z and $T_{z+1}-1$ such that there is an edge $(i,j) \in E_{T_z}$ for agent i. It is clear that $T_z \leq t^i_{T_z} \leq T_{z+1} - 1$. Also, by definition, for $l = T_z, T_z + 1, \ldots, (t^i_{T_z} - 1)$, there is no edge $(i,p) \in E(l)$. Thus $\lambda_i^{t^i_{T_z}} = \lambda_i^{T_z}$, and $\nabla f_i(\lambda_i^{t^i_{T_z}}) = \nabla f_i(\lambda_i^{T_z})$. Letting $w_{ij}(t) = \frac{1}{n}(\nabla f_i(\lambda_i^t) - \nabla f_j(\lambda_j^t))$, we have that

$$\begin{aligned}\|w_{ij}(T_z)\| &= \frac{1}{n}\|\nabla f_i(\lambda_i^{t^i_{T_z}}) - \nabla f_j(\lambda_j^{T_z})\| \\ &\leq \frac{1}{n}(\|\nabla f_i(\lambda_i^{t^i_{T_z}}) - \nabla f_j(\lambda_j^{t^i_{T_z}})\| + \|\nabla f_j(\lambda_j^{t^i_{T_z}}) - \nabla f_j(\lambda_j^{T_z})\|) \\ &\leq \frac{1}{n}\left(\|\nabla f_i(\lambda_i^{t^i_{T_z}}) - \nabla f_j(\lambda_j^{t^i_{T_z}})\| + \sum_{t=T_z}^{t^i_{T_z}-1} \|\nabla f_j(\lambda_j^{t+1}) - \nabla f_j(\lambda_j^t)\|\right).\end{aligned}$$

From the Cauchy–Schwarz inequality,

$$\begin{aligned}\|w_{ij}(T_z)\|^2 \leq\ & \frac{(t^i_{T_z} - T_z + 1)}{n^2}\left(\|\nabla f_i(\lambda_i^{t^i_{T_z}}) - \nabla f_j(\lambda_j^{t^i_{T_z}})\|^2 \right. \\ &\left. + \sum_{t=T_z}^{t^i_{T_z}-1} \|\nabla f_j(\lambda_j^{t+1}) - \nabla f_j(\lambda_j^t)\|^2\right) \\ \leq\ & \kappa\left(\frac{\|\nabla f_i(\lambda_i^{t^i_{T_z}}) - \nabla f_j(\lambda_j^{t^i_{T_z}})\|^2}{n^2} + \sum_{t=T_z}^{t^i_{T_z}-1} \frac{\|\nabla f_j(\lambda_j^{t+1}) - \nabla f_j(\lambda_j^t)\|^2}{n^2}\right) \\ \leq\ & \kappa\left(\frac{\|\nabla f_i(\lambda_i^{t^i_{T_z}}) - \nabla f_j(\lambda_j^{t^i_{T_z}})\|^2}{n^2} + \frac{1}{2n}\sum_{t=T_z}^{T_{z+1}-1} \nabla f(\lambda^t)^T v^t\right).\end{aligned}$$

The last inequality comes from (12) and from the fact that $\|\nabla f(\lambda^{t+1}) - \nabla f(\lambda^t)\|^2 = \sum_{i=1}^n \frac{\|\nabla f(\lambda_i^{t+1}) - \nabla f(\lambda_i^t)\|^2}{n^2}$. We finally have that

$$\begin{aligned}
\nabla f(\lambda^{T_z})^T \bar{v}^{T_z} &= \sum_{(i,j)\in E_{T_z}} \|w_{ij}(T_z)\|^2 \\
&\leq \sum_{(i,j)\in E_{T_z}} \kappa \left(\frac{\|\nabla f_i(\lambda_i^{t_{T_z}^i}) - \nabla f_j(\lambda_j^{t_{T_z}^i})\|^2}{n^2} + \frac{1}{2n} \sum_{t=T_z}^{T_{z+1}-1} \nabla f(\lambda^t)^T v^t \right) \\
&\leq \kappa \left(\sum_{t=T_z}^{T_{z+1}-1} \nabla f(\lambda^t)^T v^t + \frac{n-1}{2n} \sum_{t=T_z}^{T_{z+1}-1} \nabla f(\lambda^t)^T v^t \right) \\
&\leq \frac{3}{2}\kappa \left(\sum_{t=T_z}^{T_{z+1}-1} \nabla f(\lambda^t)^T v^t \right).
\end{aligned}$$

The first equality comes from Lemma A.1, with v^t replaced by $\bar{v}^{T_z}$. The second inequality comes from the fact that E_{T_z} is a subset of $E_{T_z,T_{z+1}}$ and from Lemma A.1. It is clear that Lemma A.1 is valid for all decentralized directions of update v derived using some communication graph G, where $v_i = \sum_{j\in N(i)} \frac{1}{n}(\nabla f_i(\lambda_i) - \nabla f_j(\lambda_j))$ and $N(i)$ is the set of neighbors of i in G. Hence Lemma A.1 is valid for $\bar{v}^{T_z}$. The second inequality holds because of Lemma A.1 and because there are exactly $n-1$ edges in the set E_{T_z}, as $G = (N, E_{T_z})$ is a tree. □

Proof of Theorem 3.1.

Proof of 1 *of Theorem* 3.1. First note that

$$\begin{aligned}
f(\lambda^{t+1}) - f(\lambda^t) &\geq \gamma \nabla f(\lambda^t)^T v^t - \frac{L}{2n}\|\gamma v^t\|^2 \\
&\geq \gamma \nabla f(\lambda^t)^T v^t - \frac{\gamma^2 L}{2n} 2n \nabla f(\lambda^t)^T v^t \\
&= \frac{1}{2L}\nabla f(\lambda^t)^T v^t - \frac{1}{4L}\nabla f(\lambda^t)^T v^t = \frac{1}{4L}\nabla f(\lambda^t)^T v^t. \qquad (13)
\end{aligned}$$

The first inequality comes from the descent lemma for differentiable functions [2]. The second inequality comes from Lemma A.2. The first equality comes from the fact that $\gamma = \frac{1}{2L}$. Since $\nabla f(\lambda^t)^T v^t$ is nonnegative, the sequence $\{f(\lambda^t)\}$ is monotonic and nondecreasing establishing the first part of the theorem.

Proof of 2 *of Theorem* 3.1. Since (1) is assumed to have an optimal solution, $f(\lambda^t)$ is bounded from above. We conclude from the first claim that $\{f(\lambda^t)\}$ converges, and $\{\nabla f(\lambda^t)^T v^t\}$ must converge to zero.

We now have that

$$\begin{aligned}
\|\tilde{v}^{T_z}\|^2 &= n \nabla f(\lambda^{T_z})^T \tilde{v}^{T_z} \\
&\leq \frac{n^4}{8} \nabla f(\lambda^{T_z})^T \bar{v}^{T_z} \leq \frac{3n^4}{16}\kappa \left(\sum_{t=T_z}^{T_{z+1}-1} \nabla f(\lambda^t)^T v^t \right),
\end{aligned}$$

where the first equality follows from Lemma A.3, the first inequality follows from Lemma A.4 and Lemma A.1, and the second inequality follows from Lemma A.9. The last inequality and the convergence of $\{\nabla f(\lambda^t)^T v^t\}$ to zero establishes the second part of the theorem.

Proof of 3 *of Theorem* 3.1. Note that

$$\begin{aligned} f(\lambda^{T_{z+1}}) - f(\lambda^{T_z}) &\geq \frac{1}{4L} \sum_{t=T_z}^{t=T_{z+1}-1} \nabla f(\lambda^t)^T v^t \\ &\geq \frac{1}{6L\kappa} \nabla f(\lambda^{T_z})^T \bar{v}^{T_z} \\ &\geq \frac{4}{3Ln^3\kappa} \nabla f(\lambda^{T_z})^T \tilde{v}^{T_z} \\ &= \frac{4}{3Ln^4\kappa} \|\tilde{v}^{T_z}\|^2. \end{aligned}$$

The first inequality comes from (13). The second inequality comes from Lemma A.9. The third inequality comes from Lemma A.4 and Lemma A.1, and the equality comes from Lemma A.3. Thus,

$$\begin{aligned} \sum_{z=1}^{p} f(\lambda^{T_{z+1}}) - f(\lambda^{T_z}) &\geq \frac{4}{3Ln^4\kappa} \sum_{z=1}^{p} \|\tilde{v}^{T_z}\|^2 \\ f(\lambda^{T_{p+1}}) - f(\lambda^1) &\geq \frac{4}{3Ln^4\kappa} \sum_{z=1}^{p} \|\tilde{v}^{T_z}\|^2 \\ f(\lambda^*) - f(\lambda^1) &\geq \frac{4}{3Ln^4\kappa} \sum_{z=1}^{p} \|\tilde{v}^{T_z}\|^2. \end{aligned}$$

The last inequality, together with the fact that $p(\min_{z=1,\ldots,p} \|\tilde{v}^{T_z}\|^2) \leq \sum_{z=1}^{p} \|\tilde{v}^{T_z}\|^2$, proves the third claim.

Proof of 4 *of Theorem* 3.1. If the set of optima of (1) is bounded, $\{\lambda : \|\tilde{v}(\lambda)\| \leq C\}$ is a bounded set for some $C > 0$. We conclude that λ^{T_z} has a converging subsequence $\lambda^{T_{z_k}}$. Let $\bar{\lambda}$ be the limit of $\lambda^{T_{z_k}}$. Since $\|\tilde{v}(\cdot)\|$ is a continuous function and $\|\tilde{v}(\lambda^{T_{z_k}})\|$ converges to zero, we conclude that $\tilde{v}(\bar{\lambda}) = 0$ and $\bar{\lambda}$ is optimal. Since f is continuous, we conclude that $\{f(\lambda^{T_{z_k}})\}$ converges to $f(\bar{\lambda}) = f(\lambda^*)$. Since $\{f(\lambda^t)\}$ converges, we conclude that it must converge to $f(\lambda^*)$. ∎

LEMMA 3.3. *Let f_i and $\hat{f}_i$ be as defined in section* 3.2. *Then the following hold:*

1. *$\hat{f}_i$ is concave and differentiable, with gradient $\nabla \hat{f}_i(\lambda_i) = \mathrm{E}[\nabla f_i(\lambda_i + Z_i)]$;*
2. *$f_i(\lambda_i) \geq \hat{f}_i(\lambda_i) \geq f_i(\lambda_i) - 2.8\epsilon L$;*
3. *$\|\nabla \hat{f}_i(\lambda_i) - \nabla \hat{f}_i(\bar{\lambda}_i)\| \leq \frac{\sqrt{\log(m+1)}L}{\epsilon} \|\lambda_i - \bar{\lambda}_i\|$.*

Proof of 1 *of Lemma* 3.3. For all $a \in [0, 1]$, we have that

$$\begin{aligned} \hat{f}_i(a\lambda_i + (1-a)\bar{\lambda}_i) &= \mathrm{E}[f_i(a\lambda_i + (1-a)\bar{\lambda}_i + Z_i)] \\ &\geq \mathrm{E}[af_i(\lambda_i + Z_i) + (1-a)f_i(\bar{\lambda}_i + Z_i)] = a\hat{f}_i(\lambda_i) + (1-a)\hat{f}_i(\bar{\lambda}_i). \end{aligned}$$

It follows that $\hat{f}_i$ is concave. Since f_i is nondifferentiable only on a set of measure zero, we have that

$$\begin{aligned} (\nabla f_i(\lambda_i + Z_i))_j &= \lim_{\delta \uparrow 0} \frac{f_i(\lambda_i + Z_i + \delta e_j) - f_i(\lambda_i + Z_i)}{\delta} \\ &= \lim_{\delta \downarrow 0} \frac{f_i(\lambda_i + Z_i + \delta e_j) - f_i(\lambda_i + Z_i)}{\delta}, \end{aligned}$$

with probability 1, where e_j is the vector with all entries equal to zero except for the jth entry, which is equal to one. Hence

$$\begin{aligned}\lim_{\delta\uparrow 0}\frac{\mathrm{E}[f_i(\lambda_i+Z_i+\delta e_j)]-\mathrm{E}[f_i(\lambda_i+Z_i)]}{\delta} &= \mathrm{E}\left[\lim_{\delta\uparrow 0}\frac{f_i(\lambda_i+Z_i+\delta e_j)-f_i(\lambda_i+Z_i)}{\delta}\right]\\ &= \mathrm{E}[(\nabla f_i(\lambda_i+Z_i))_j]\\ &= \mathrm{E}\left[\lim_{\delta\downarrow 0}\frac{f_i(\lambda_i+Z_i+\delta e_j)-f_i(\lambda_i+Z_i)}{\delta}\right]\\ &= \lim_{\delta\downarrow 0}\frac{\mathrm{E}[f_i(\lambda_i+Z_i+\delta e_j)]-\mathrm{E}[f_i(\lambda_i+Z_i)]}{\delta}.\end{aligned}$$

Note that $|\frac{f_i(\lambda_i+Z_i+\delta e_j)-f_i(\lambda_i+Z_i)}{\delta}| \le L$. Hence the exchanges between limit and expectation are valid by the bounded convergence theorem. It follows that $\hat{f}_i$ is differentiable, and its gradient is given by

$$\nabla\hat{f}_i(\lambda_i) = \mathrm{E}[\nabla f_i(\lambda_i+Z_i)].$$

Proof of 2 *of Lemma* 3.3. First, we have that

$$\begin{aligned}\hat{f}_i(\lambda_i) &= \mathrm{E}[f_i(\lambda_i+Z_i)]\\ &\le f_i(\lambda_i+EZ_i) = f_i(\lambda_i),\end{aligned}$$

where the inequality follows from the concavity of f_i and Jensen's inequality [4].

For the lower bound on $\hat{f}_i$, we have that

$$\begin{aligned}\hat{f}_i(\lambda_i) &= \mathrm{E}[f_i(\lambda_i+Z_i)]\\ &= \mathrm{E}[f_i(\lambda_i-Z_i)]\\ &\ge \mathrm{E}[f_i(\lambda_i)-Z_i^T\nabla f_i(\lambda_i-Z_i)]\\ &\ge f_i(\lambda_i)-\mathrm{E}[\max_j|Z_{ij}|]L,\end{aligned}\tag{14}$$

where $|Z_{ij}|$ is the modulus function. The first inequality follows from the concavity of f and the fact that $\nabla f_i(\lambda_i-Z_i)$ is a subgradient of f at λ_i-Z_i. The second inequality follows from the fact that $\|\nabla f_i(\lambda_i-Z_i)\|_1 \le L$.

We now show that $\mathrm{E}[\max_j|Z_{ij}|] \le 2.8\epsilon$. Note that this inequality and (14) prove the claim.

We first place a bound on $P(|Z_{ij}|>c)$, for $c>0$. We have that

$$\begin{aligned}P(|Z_{ij}|>c) &= \int_c^\infty \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{z^2}{2\sigma^2}}dz+\int_{-\infty}^{-c}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{z^2}{2\sigma^2}}dz = 2\int_c^\infty\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{z^2}{2\sigma^2}}dz\\ &= 2e^{-\frac{c^2}{2\sigma^2}}\int_0^\infty\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{z^2+2zc}{2\sigma^2}}dz\\ &= 2e^{-\frac{c^2}{2\sigma^2}}\left(\int_0^c\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{z^2+2zc}{2\sigma^2}}dz+\int_c^\infty\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{z^2+2zc}{2\sigma^2}}dz\right)\\ &\le 2e^{-\frac{c^2}{2\sigma^2}}\left(\int_0^c\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{zc}{\sigma^2}}dz+\int_c^\infty\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{z^2}{2\sigma^2}}dz\right)\\ &= 2e^{-\frac{c^2}{2\sigma^2}}\left(\frac{\sigma}{\sqrt{2\pi}c}\left(1-e^{-\frac{c^2}{\sigma^2}}\right)+\frac{1}{2}P(|Z_{ij}|>c)\right).\end{aligned}$$

Hence

$$P(|Z_{ij}| > c) \leq 2\left(\frac{\frac{e^{-\frac{c^2}{2\sigma^2}}\sigma}{\sqrt{2\pi}c}\left(1-e^{-\frac{c^2}{\sigma^2}}\right)}{1-e^{-\frac{c^2}{2\sigma^2}}}\right)$$
$$= \frac{2e^{-\frac{c^2}{2\sigma^2}}\sigma}{\sqrt{2\pi}c}\left(1+e^{-\frac{c^2}{2\sigma^2}}\right) \leq \frac{2e^{-\frac{c^2}{2\sigma^2}}\sqrt{2}\sigma}{\sqrt{\pi}c}.$$

It follows that, for all $c \geq \frac{2\epsilon}{\sqrt{\pi}}$,

$$P(\max_j |Z_{ij}| > c) \leq 2m\frac{e^{-\frac{c^2}{2\sigma^2}}\sqrt{2}\sigma}{\sqrt{\pi}c} \leq 2(m+1)\frac{e^{-\frac{c^2\pi\log(m+1)}{4\epsilon^2}}}{\sqrt{\pi\log(m+1)}}$$
$$= \frac{2e^{-\frac{\left(c^2-\frac{4\epsilon^2}{\pi}\right)\pi\log(m+1)}{4\epsilon^2}}}{\sqrt{\pi\log(m+1)}} \leq \frac{2e^{-\frac{\left(c-\frac{2\epsilon}{\sqrt{\pi}}\right)^2\pi\log(m+1)}{4\epsilon^2}}}{\sqrt{\pi\log(m+1)}}.$$

The first inequality follows from the union bound [4]. The second inequality follows from $c \geq \frac{2\epsilon}{\sqrt{\pi}}$. The last inequality follows from $(c-\frac{2\epsilon}{\sqrt{\pi}})^2 \leq c^2 - \frac{4\epsilon^2}{\pi}$ for all $c > \frac{2\epsilon}{\sqrt{\pi}}$.

Finally,

$$\begin{aligned}
&\mathrm{E}[\max_j |Z_{ij}|] \\
&= \int_0^\infty P(\max_j |Z_{ij}| > z)dz \\
&= \int_0^{\frac{2\epsilon}{\sqrt{\pi}}} P(\max_j |Z_{ij}| > z)dz + \int_{\frac{2\epsilon}{\sqrt{\pi}}}^\infty P(\max_j |Z_{ij}| > z)dz \\
&\leq \frac{2\epsilon}{\sqrt{\pi}} + \int_{\frac{2\epsilon}{\sqrt{\pi}}}^\infty \frac{2e^{-\frac{\left(z-\frac{2\epsilon}{\sqrt{\pi}}\right)^2\pi\log(m+1)}{4\epsilon^2}}}{\sqrt{\pi\log(m+1)}}dz \\
&= \frac{2\epsilon}{\sqrt{\pi}} + \frac{4\epsilon}{\sqrt{\pi\log(m+1)}}\int_{\frac{2\epsilon}{\sqrt{\pi}}}^\infty \frac{\sqrt{\log(m+1)}}{2\epsilon}e^{-\frac{\left(z-\frac{2\epsilon}{\sqrt{\pi}}\right)^2\pi\log(m+1)}{4\epsilon^2}}dz \\
&= \frac{2\epsilon}{\sqrt{\pi}} + \frac{4\epsilon}{\sqrt{\pi\log(m+1)}}\int_{\frac{2\epsilon}{\sqrt{\pi}}}^\infty \frac{1}{\sqrt{2\pi}}\frac{\sqrt{\pi\log(m+1)}}{\sqrt{2}\epsilon}e^{-\frac{\left(\frac{\sqrt{\pi\log(m+1)}\left(z-\frac{2\epsilon}{\sqrt{\pi}}\right)}{\sqrt{2}\epsilon}\right)^2}{2}}dz \\
&= \frac{2\epsilon}{\sqrt{\pi}} + \frac{2\epsilon}{\sqrt{\pi\log(m+1)}} \\
&\leq 2.8\epsilon.
\end{aligned}$$

The last equality comes from the identity $\frac{1}{\sqrt{2\pi}}\int_0^\infty e^{-\frac{t^2}{2}}dt = \frac{1}{2}$.

Proof of 3 *of Lemma* 3.3. Denote by $p(\cdot)$ the probability density function for Z_i; i.e., the joint probability density function for $Z_{i1}, \ldots, Z_{im}$. Then we have that

$$\nabla\hat{f}_i(\lambda_i) = \int_{\Re^m} p(z)\nabla f_i(\lambda_i + z)dz$$
$$= \int_{\Re^m} p(z-\lambda_i)\nabla f_i(z)dz.$$

It follows that

$$\begin{aligned}\|\nabla \hat{f}_i(\lambda_i) - \nabla \hat{f}_i(\bar{\lambda}_i)\| &= \left\| \int_{\Re^m} (p(z-\lambda_i) - p(z-\bar{\lambda}_i)) \nabla f_i(z) dz \right\| \\ &\le \int_{\Re^m} |p(z-\lambda_i) - p(z-\bar{\lambda}_i)| \|\nabla f_i(z)\| dz \\ &\le L \int_{\Re^m} |p(z-\lambda_i) - p(z-\bar{\lambda}_i)| dz. \qquad (15)\end{aligned}$$

Since $p(\cdot)$ is the joint distribution of m i.i.d. zero-mean Gaussian random variables, $p(z)$ is strictly decreasing on $\|z\|$. Hence

$$\begin{aligned}&\int_{\Re^m} |p(z-\lambda_i) - p(z-\bar{\lambda}_i)| dz \\ &\quad = \int_{\{z\in\Re^m : \|z-\lambda_i\| < \|z-\bar{\lambda}_i\|\}} \left(p(z-\lambda_i) - p(z-\bar{\lambda}_i)\right) dz \\ &\qquad + \int_{\{z\in\Re^m : \|z-\lambda_i\| > \|z-\bar{\lambda}_i\|\}} \left(p(z-\bar{\lambda}_i) - p(z-\lambda_i)\right) dz \\ &\quad = 2 \int_{\{z\in\Re^m : \|z-\lambda_i\| < \|z-\bar{\lambda}_i\|\}} \left(p(z-\lambda_i) - p(z-\bar{\lambda}_i)\right) dz \\ &\quad = 2 \int_{\{z\in\Re^m : \|z\| < \|z-(\bar{\lambda}_i-\lambda_i)\|\}} p(z) dz - 2 \int_{\{z\in\Re^m : \|z\| > \|z-(\lambda_i-\bar{\lambda}_i)\|\}} p(z) dz \\ &\quad = 2P(\|Z_i\| < \|Z_i - (\bar{\lambda}_i - \lambda_i)\|) - 2P(\|Z_i\| > \|Z_i - (\lambda_i - \bar{\lambda}_i)\|) \\ &\quad = 2P(2Z_i^T(\bar{\lambda}_i - \lambda_i) < \|\bar{\lambda}_i - \lambda_i\|^2) - 2P(2Z_i^T(\bar{\lambda}_i - \lambda_i) < -\|\bar{\lambda}_i - \lambda_i\|^2) \\ &\quad = 2P(-0.5\|\bar{\lambda}_i - \lambda_i\| < V < 0.5\|\bar{\lambda}_i - \lambda_i\|), \qquad (16)\end{aligned}$$

where

$$V = \frac{Z_i^T(\bar{\lambda}_i - \lambda_i)}{\|\bar{\lambda}_i - \lambda_i\|}.$$

It is easy to verify that V is normal with a zero mean and variance equal to $\sigma = \frac{\sqrt{2}\epsilon}{\sqrt{\pi \log(m+1)}}$. It follows that

$$\begin{aligned}P(-0.5\|\bar{\lambda}_i - \lambda_i\| < V < 0.5\|\bar{\lambda}_i - \lambda_i\|) &\le \frac{1}{\sqrt{2\pi}\sigma} \|\bar{\lambda}_i - \lambda_i\| \\ &= \frac{\sqrt{\log(m+1)}}{2\epsilon} \|\bar{\lambda}_i - \lambda_i\|. \qquad (17)\end{aligned}$$

The claim follows from (15), (16), and (17). □

THEOREM 3.2. *Suppose that Assumptions* 3.3 *and* 3.4 *hold. Then with probability* 1*:*

1. *the sequence* $\{\|\tilde{v}^{T_z}\|\}$ *converges to* 0*;*
2. $\min_{z=1,\ldots,p} \mathrm{E}[\|\tilde{v}^{T_z}\|^2] \le \dfrac{\frac{n^4\kappa L\sqrt{\log(m+1)}}{\epsilon}\left[3(f(\lambda^*) - f(\lambda^1) + 2.8\epsilon L) + \sum_{t=1}^{t=\kappa p} \frac{4L\beta_t^2\epsilon}{\sqrt{\log(m+1)}}\right]}{4\sum_{z=2}^{p+1}\beta_{\kappa z}}$ $\forall p$*;*
3. *if the set of the optima of* (1) *is bounded, then* $\lim_{t\to\infty} f(\lambda^t) \ge f(\lambda^*) - 2.8\epsilon L$.

The proof has the same structure as the proof of Theorem 3.1. Let the expected direction of update at time t be v^t:

$$v_i^t = \frac{1}{n} \sum_{j \in N_i(t)} (\nabla \hat{f}_i(\lambda_i^t) - \nabla \hat{f}_j(\lambda_j^t)).$$

Let δ^t be the random variable denoting the difference between the actual and the expected directions of update:

$$\delta_i^t = \sum_{j \in N_i(t)} \frac{1}{n} (\nabla f_i(\lambda_i^t + Z_i^t) - \nabla f_j(\lambda_j^t + Z_j^t)) - v_i^t.$$

Let $\mathcal{F}_t$ be the sigma-algebra [4] generated by $Z_i^\tau,\ i = 1, \ldots, n, \tau = 1, \ldots, t$. We have the following result about δ_i^t.

LEMMA A.10. *For all t, $\mathrm{E}[\delta^t | \mathcal{F}_{t-1}] = 0$ and $\mathrm{E}[\|\delta^t\|^2 | \mathcal{F}_{t-1}] < 8nL^2$, with probability* 1.

Proof. $\mathrm{E}[\delta_i^t | \mathcal{F}_{t-1}] = 0$ follows from $\nabla \hat{f}_i(\lambda_i) = \mathrm{E}[\nabla f_i(\lambda_i + Z_i^t)]$ for all i. Moreover,

$$\begin{aligned}
&\mathrm{E}[\|\delta_i^t\|^2 | \mathcal{F}_{t-1}] \\
&= \mathrm{E}\left[\left\| \sum_{j \in N_i(t)} \frac{\nabla f_i(\lambda_i^t + Z_i^t) - \nabla \hat{f}_i(\lambda_i^t) - \nabla f_j(\lambda_j^t + Z_j^t) + \nabla \hat{f}_j(\lambda_j^t)}{n} \right\|^2 | \mathcal{F}_{t-1} \right] \\
&= \frac{\mathrm{E}\left[\left\| N_i(t) \left(\nabla f_i(\lambda_i^t + Z_i^t) - \nabla \hat{f}_i(\lambda_i^t) \right) - \sum_{j \in N_i(t)} \left(\nabla f_j(\lambda_j^t + Z_j^t) - \nabla \hat{f}_j(\lambda_j^t) \right) \right\|^2 | \mathcal{F}_{t-1} \right]}{n^2} \\
&\leq \frac{N_i(t)^2 \mathrm{E}[\|\nabla f_i(\lambda_i^t + Z_i^t) - \nabla \hat{f}_i(\lambda_i^t)\|^2 | \mathcal{F}_{t-1}] + \sum_{j \in N_i(t)} \mathrm{E}[\|\nabla f_j(\lambda_j^t + Z_j^t) - \nabla \hat{f}_j(\lambda_j^t)\|^2 | \mathcal{F}_{t-1}]}{n^2} \\
&< 8L^2.
\end{aligned}$$

The last inequality follows from $N_i(t) < n$ and

$$\|\nabla f_j(\lambda_j^t + Z_j^t) - \nabla \hat{f}_j(\lambda_j^t)\| \leq \|\nabla f_j(\lambda_j^t + Z_j^t)\| + \|\nabla \hat{f}_j(\lambda_j^t)\| \leq 2L.$$

Finally,

$$\mathrm{E}[\|\delta^t\|^2 | \mathcal{F}_{t-1}] = \sum_i \mathrm{E}[\|\delta_i^t\|^2 | \mathcal{F}_{t-1}] < 8nL^2. \qquad \square$$

The following results follow immediately from Lemmas A.1–A.4 applied with $\hat{f}_i$ replacing f_i for all i:

$$\nabla \hat{f}(\lambda^t)^T v^t = \frac{1}{n^2} \sum_{(i,j) \in E(t)} \|\nabla \hat{f}_i(\lambda_i^t) - \nabla \hat{f}_j(\lambda_j^t)\|^2 \tag{18}$$

$$\|v^t\|^2 \leq 2n \nabla \hat{f}(\lambda^t)^T v^t \tag{19}$$

$$\|\tilde{v}^t\|^2 = n \nabla \hat{f}(\lambda^t)^T \tilde{v}^t \tag{20}$$

$$\sum_{(i,j) \in E} \|\nabla \hat{f}_j(\lambda_j^t) - \nabla \hat{f}_i(\lambda_i^t)\|^2 \geq \frac{8}{n^3} \sum_{((i,j) \in N^2, i<j)} \|\nabla \hat{f}_j(\lambda_j^t) - \nabla \hat{f}_i(\lambda_i^t)\|^2$$

$$\forall E : (N, E) \text{ is connected.} \tag{21}$$

Let E_{T_z} be a subset of the edge set $E_{T_z,T_{z+1}}$ such that the graph (N, E_{T_z}) is a tree. By Assumption 2.1, the graph $(N, E_{T_z,T_{z+1}})$ is connected and so E_{T_z} is well defined. As before, let the decentralized direction of update derived using $G = (N, E_{T_z})$ be denoted by $\bar{v}^{T_z}$. The following result is the counterpart of Lemma A.9.

LEMMA A.11. *Let* $L_\epsilon = \frac{\sqrt{\log(m+1)}L}{\epsilon}$.
$\nabla \hat{f}(\lambda^{T_z})^T \bar{v}^{T_z} \leq \kappa \sum_{t=T_z}^{t=T_{z+1}-1} [(1 + 2L_\epsilon^2 \gamma_t^2) \mathrm{E}[\nabla \hat{f}(\lambda^t)^T v^t | \mathcal{F}_{T_z-1}] + 8L^2 L_\epsilon^2 \gamma_t^2]$.

Proof. We have that

$$\begin{aligned}
\mathrm{E}[\|\nabla \hat{f}(\lambda^{t+1}) - \nabla \hat{f}(\lambda^t)\|^2 | \mathcal{F}_{t-1}] &\leq \frac{L_\epsilon^2}{n^2} \mathrm{E}[\|\gamma_t (v^t + \delta^t)\|^2 | \mathcal{F}_{t-1}] \\
&= \frac{L_\epsilon^2 \gamma_t^2}{n^2} (\|v^t\|^2 + \mathrm{E}[\|\delta^t\|^2 | \mathcal{F}_{t-1}]) \\
&\leq \frac{L_\epsilon^2 \gamma_t^2}{n^2} (2n \nabla \hat{f}(\lambda^t)^T v^t + 8nL^2). \qquad (22)
\end{aligned}$$

It follows from Lemma 3.3 that L_ϵ is a Lipschitz constant for the functions $\hat{f}_i$, $i = 1, \ldots, n$. Hence $\frac{L_\epsilon}{n}$ is a Lipschitz constant for $\hat{f}$, and the first inequality follows from this. The second inequality follows from (19) and Lemma A.10.

Let $t^i_{T_z}$ be the earliest time between the time periods T_z and $T_{z+1} - 1$ such that there is an edge $(i, j) \in E_{T_z}$ for agent i. It is clear that $T_z \leq t^i_{T_z} \leq T_{z+1} - 1$. Also, by definition, for $l = T_z, T_z + 1, \ldots, (t^i_{T_z} - 1)$, there is no edge $(i, p) \in E(l)$. Thus $\lambda_i^{t^i_{T_z}} = \lambda_i^{T_z}$, and $\nabla \hat{f}_i(\lambda_i^{t^i_{T_z}}) = \nabla \hat{f}_i(\lambda_i^{T_z})$. Let $w_{ij}(t) = \frac{1}{n}(\nabla \hat{f}_i(\lambda_i^t) - \nabla \hat{f}_j(\lambda_j^t))$. Then

$$\begin{aligned}
\|w_{ij}(T_z)\| &= \frac{1}{n} \|\nabla \hat{f}_i(\lambda_i^{t^i_{T_z}}) - \nabla \hat{f}_j(\lambda_j^{T_z})\| \\
&\leq \frac{1}{n} (\|\mathrm{E}[\nabla \hat{f}_i(\lambda_i^{t^i_{T_z}}) - \nabla \hat{f}_j(\lambda_j^{t^i_{T_z}}) | \mathcal{F}_{T_z-1}]\| \\
&\quad + \|\mathrm{E}[\nabla \hat{f}_j(\lambda_j^{t^i_{T_z}}) - \nabla \hat{f}_j(\lambda_j^{T_z}) | \mathcal{F}_{T_z-1}]\|) \\
&\leq \frac{1}{n} \left(\|\mathrm{E}[\nabla \hat{f}_i(\lambda_i^{t^i_{T_z}}) - \nabla \hat{f}_j(\lambda_j^{t^i_{T_z}}) | \mathcal{F}_{T_z-1}]\| \right. \\
&\qquad \left. + \sum_{t=T_z}^{t=t^i_{T_z}-1} \|\mathrm{E}[\nabla \hat{f}_j(\lambda_j^{t+1}) - \nabla \hat{f}_j(\lambda_j^t) | \mathcal{F}_{T_z-1}]\| \right).
\end{aligned}$$

From the Cauchy–Schwarz inequality,

$$\begin{aligned}
\|w_{ij}(T_z)\|^2 &\leq \frac{(t^i_{T_z} - T_z + 1)}{n^2} \left(\|\mathrm{E}[\nabla \hat{f}_i(\lambda_i^{t^i_{T_z}}) - \nabla \hat{f}_j(\lambda_j^{t^i_{T_z}}) | \mathcal{F}_{T_z-1}]\|^2 \right. \\
&\qquad \left. + \sum_{t=T_z}^{t=t^i_{T_z}-1} \|\mathrm{E}[\nabla \hat{f}_j(\lambda_j^{t+1}) - \nabla \hat{f}_j(\lambda_j^t) | \mathcal{F}_{T_z-1}]\|^2 \right) \\
&\leq \frac{\kappa}{n^2} \left(\mathrm{E}[\|\nabla \hat{f}_i(\lambda_i^{t^i_{T_z}}) - \nabla \hat{f}_j(\lambda_j^{t^i_{T_z}})\|^2 | \mathcal{F}_{T_z-1}] \right. \\
&\qquad \left. + \sum_{t=T_z}^{t=t^i_{T_z}-1} \mathrm{E}[\|\nabla \hat{f}_j(\lambda_j^{t+1}) - \nabla \hat{f}_j(\lambda_j^t)\|^2) | \mathcal{F}_{T_z-1}] \right)
\end{aligned}$$

$$\leq \kappa \left(\frac{\mathrm{E}[\|\nabla \hat{f}_i(\lambda_i^{t^i_{T_z}}) - \nabla \hat{f}_j(\lambda_j^{t^i_{T_z}})\|^2 | \mathcal{F}_{T_z-1}]}{n^2} + \frac{L_\epsilon^2}{n^2} \sum_{t=T_z}^{t=T_{z+1}-1} \gamma_t^2 (\mathrm{E}[2n \nabla \hat{f}(\lambda^t)^T v^t | \mathcal{F}_{T_z-1}] + 8nL^2) \right).$$

The last inequality follows from the fact that $\|\nabla \hat{f}(\lambda^{t+1}) - \nabla \hat{f}(\lambda^t)\|^2 = \frac{1}{n^2} \sum_{k=1}^n \|\nabla \hat{f}_k (\lambda_k^{t+1}) - \nabla \hat{f}_k(\lambda_k^t)\|^2 \geq \frac{1}{n^2} \|\nabla \hat{f}_j(\lambda_j^{t+1}) - \nabla \hat{f}_j(\lambda_j^t)\|^2$ and from (22). We finally have that

$$\begin{aligned}
\nabla \hat{f}(\lambda^{T_z})^T \bar{v}^{T_z} &= \sum_{(i,j) \in E_{T_z}} \|w_{ij}(T_z)\|^2 \\
&\leq \sum_{(i,j) \in E_{T_z}} \kappa \left(\frac{\mathrm{E}[\|\nabla \hat{f}_i(\lambda_i^{t^i_{T_z}}) - \nabla \hat{f}_j(\lambda_j^{t^i_{T_z}})\|^2 | \mathcal{F}_{T_z-1}]}{n^2} + \frac{L_\epsilon^2}{n^2} \sum_{t=T_z}^{t=T_{z+1}-1} \gamma_t^2 (\mathrm{E}[2n \nabla \hat{f}(\lambda^t)^T v^t | \mathcal{F}_{T_z-1}] + 8nL^2) \right) \\
&\leq \kappa \sum_{t=T_z}^{t=T_{z+1}-1} \left[(1 + 2L_\epsilon^2 \gamma_t^2) \mathrm{E}[\nabla \hat{f}(\lambda^t)^T v^t | \mathcal{F}_{T_z-1}] + 8L^2 L_\epsilon^2 \gamma_t^2 \right].
\end{aligned}$$

In the last inequality, we have used the fact that

$$\begin{aligned}
&\sum_{(i,j) \in E_{T_z}} \frac{\mathrm{E}[\|\nabla \hat{f}_i(\lambda_i^{t^i_{T_z}}) - \nabla \hat{f}_j(\lambda_j^{t^i_{T_z}})\|^2 | \mathcal{F}_{T_z-1}]}{n^2} \\
&\quad = \sum_{t=T_z}^{t=T_{z+1}-1} \sum_{\left((i,j) \in E_{T_z}, t^i_{T_z} = t\right)} \frac{\mathrm{E}[\|\nabla \hat{f}_i(\lambda_i^{t^i_{T_z}}) - \nabla \hat{f}_j(\lambda_j^{t^i_{T_z}})\|^2 | \mathcal{F}_{T_z-1}]}{n^2} \\
&\quad \leq \sum_{t=T_z}^{t=T_{z+1}-1} \mathrm{E}[\nabla \hat{f}(\lambda^t)^T v^t | \mathcal{F}_{T_z-1}]. \qquad \square
\end{aligned}$$

Proof of Theorem 3.2.

Proof of 1 *of Theorem* 3.2. We first have that

$$\begin{aligned}
\mathrm{E}[\hat{f}(\lambda^{t+1}) | \mathcal{F}_{t-1}] &\geq \hat{f}(\lambda^t) + \gamma_t \nabla \hat{f}(\lambda^t)^T v^t - \frac{L_\epsilon}{2n} \mathrm{E}[\|\gamma_t (v^t + \delta^t)\|^2 | \mathcal{F}_{t-1}] \\
&= \hat{f}(\lambda^t) + \gamma_t \nabla \hat{f}(\lambda^t)^T v^t - \frac{L_\epsilon}{2n} \|\gamma_t v^t\|^2 - \frac{L_\epsilon}{2n} \mathrm{E}[\|\gamma_t \delta^t\|^2 | \mathcal{F}_{t-1}] \\
&\geq \hat{f}(\lambda^t) + (\gamma_t - L_\epsilon \gamma_t^2) \nabla \hat{f}(\lambda^t)^T v^t - 4L_\epsilon L^2 \gamma_t^2. \qquad (23)
\end{aligned}$$

The first inequality comes from the descent lemma for differentiable functions [2]. The equality follows from $\mathrm{E}[\delta | \mathcal{F}_{t-1}] = 0$ from Lemma A.10. The second inequality follows from Lemma A.10 and (19).

Note that $\nabla \hat{f}(\lambda^t)^T v^t \geq 0$. This and Assumption 3.4 imply that the second term in (23) is also greater than or equal to zero. Moreover,

$$\sum_t 4L_\epsilon L^2 \gamma_t^2 < \infty.$$

Since $\hat{f}$ is bounded from above, we conclude by the supermartingale convergence theorem [4] that $\hat{f}(\lambda^t)$ converges with probability 1. Moreover, $\sum_t(\gamma_t - L_\epsilon\gamma_t^2)\nabla\hat{f}(\lambda^t)^T v^t < \infty$ with probability 1 and since $\sum_t \gamma_t = \infty$, we conclude that $\nabla\hat{f}(\lambda^t)^T v^t$ converges to zero with probability 1. Note that

$$\mathrm{E}[\nabla\hat{f}(\lambda^t)^T v^t | \mathcal{F}_{t-1}] = \nabla\hat{f}(\lambda^t)^T v^t$$

with probability 1, and we conclude that $\mathrm{E}[\nabla\hat{f}(\lambda^t)^T v_t | \mathcal{F}_{t-1}]$ also converges to zero with probability 1.

We now have that

$$\begin{aligned}
\|\tilde{v}^{T_z}\|^2 &= n\nabla\hat{f}(\lambda^{T_z})^T\tilde{v}^{T_z} \\
&\leq \frac{n^4}{8}\nabla\hat{f}(\lambda^{T_z})^T\bar{v}^{T_z} \\
&\leq \frac{n^4}{8}\kappa\sum_{t=T_z}^{t=T_{z+1}-1}\left[(1+2L_\epsilon^2\gamma_t^2)\mathrm{E}[\nabla\hat{f}(\lambda^t)^T v^t|\mathcal{F}_{T_z-1}] + 8L^2L_\epsilon^2\gamma_t^2\right] \\
&\leq \frac{n^4}{8}\kappa\sum_{t=T_z}^{t=T_{z+1}-1}\left[1.5\mathrm{E}[\nabla\hat{f}(\lambda^t)^T v^t|\mathcal{F}_{T_z-1}] + 2L^2\beta_t^2\right]. \qquad (24)
\end{aligned}$$

The equality follows from (20). The first inequality follows from (21) and (18). The second inequality follows from Lemma A.11. The third inequality follows from Assumption 3.4 on the step sizes γ_t. We conclude that $\|\tilde{v}^{T_z}\|$ converges to zero with probability 1.

Proof of 2 *of Theorem* 3.2. From (23), we have that

$$\begin{aligned}
\nabla\hat{f}(\lambda^t)^T v^t &\leq \frac{\mathrm{E}[\hat{f}(\lambda^{t+1})|\mathcal{F}_{t-1}] + 4L_\epsilon L^2\gamma_t^2 - \hat{f}(\lambda^t)}{\gamma_t(1-L_\epsilon\gamma_t)} \\
&\leq \frac{2(\mathrm{E}[\hat{f}(\lambda^{t+1})|\mathcal{F}_{t-1}] - \hat{f}(\lambda^t)) + \left(\frac{2L^2\beta_t^2}{L_\epsilon}\right)}{\gamma_t}. \qquad (25)
\end{aligned}$$

In the second inequality we have used $\gamma_t \leq \frac{1}{2L_\epsilon}$ from Assumption 3.4.

Combining (24) and (25), we have that

$$\begin{aligned}
\mathrm{E}[\|\tilde{v}^{T_z}\|^2] &\leq \frac{n^4}{8}\kappa\sum_{t=T_z}^{t=T_{z+1}-1}\left(\frac{3\mathrm{E}[\hat{f}(\lambda^{t+1}) - \hat{f}(\lambda^t)] + \left(\frac{3L^2\beta_t^2}{L_\epsilon}\right)}{\gamma_t} + 2L^2\beta_t^2\right) \\
&\leq \frac{n^4}{8\gamma_{\kappa z}}\kappa\sum_{t=T_z}^{t=T_{z+1}-1}\left[3\mathrm{E}[\hat{f}(\lambda^{t+1}) - \hat{f}(\lambda^t)] + \frac{3L^2\beta_t^2}{L_\epsilon} + \frac{L^2\beta_t^2}{L_\epsilon}\right].
\end{aligned}$$

The last inequality follows from Assumption 3.4 on the step sizes. It follows that

$$\begin{aligned}
\sum_{z=1}^{p}\gamma_{\kappa z}\mathrm{E}[\|\tilde{v}^{T_z}\|^2] &\leq \frac{n^4}{8}\kappa\sum_{t=1}^{t=T_{p+1}-1}\left[3\mathrm{E}[\hat{f}(\lambda^{t+1}) - \hat{f}(\lambda^t)] + \frac{4L^2\beta_t^2}{L_\epsilon}\right] \\
&\leq \frac{n^4}{8}\kappa\left[3(\hat{f}(\hat{\lambda}) - \hat{f}(\lambda^1)) + \sum_{t=1}^{t=\kappa p}\frac{4L^2\beta_t^2}{L_\epsilon}\right],
\end{aligned}$$

where $\hat{\lambda}$ denotes an optimal solution of (5). From Lemma 3.3, we have that $\hat{f}(\lambda^1) \geq f(\lambda^1) - 2.8\epsilon L$. We also have that $\hat{f}(\hat{\lambda}) \leq f(\hat{\lambda}) \leq f(\lambda^*)$. It follows that

$$\min_{z=1,\ldots,p} \mathrm{E}[\|\tilde{v}^{T_z}\|^2] \leq \frac{\frac{n^4\kappa L\sqrt{\log(m+1)}}{\epsilon}\left[3(f(\lambda^*) - f(\lambda^1) + 2.8\epsilon L) + \sum_{t=1}^{t=\kappa p} \frac{4L\beta_t^2\epsilon}{\sqrt{\log(m+1)}}\right]}{4\sum_{z=2}^{p+1}\beta_{\kappa z}} \quad \forall p.$$

Proof of 3 *of Theorem* 3.2. Since $f \geq \hat{f} \geq f - 2.8\epsilon L$, if (1) has a bounded set of optima, so does (5). Recall from the proof of the first claim that $\hat{f}(\lambda^t)$ converges with probability 1. Using the same argument as in the proof of the fourth claim of Theorem 3.1, we conclude that $\hat{f}(\lambda^t)$ converges to $\hat{f}(\hat{\lambda})$ with probability 1. We conclude that

$$\begin{aligned}\limsup_{t\to\infty} f(\lambda^t) &\geq \hat{f}(\hat{\lambda}) \\ &\geq \hat{f}(\lambda^*) \\ &\geq f(\lambda^*) - 2.8\epsilon L.\end{aligned}$$

The first inequality follows from $f(\lambda^t) \geq \hat{f}(\lambda^t)$ for all t from Lemma 3.3. The second inequality follows from the optimality of $\hat{\lambda}$. The third inequality follows from Lemma 3.3. ∎

LEMMA 4.1. *Under assumption* 3.3 *for* f,

1. *for all* i, $g_i(\lambda_i)$ *is concave and differentiable outside a set of measure zero;*
2. *for all* i *and* λ_i, $\sup_{i,\lambda_i}\{\|v\|_1 : v \in \partial g_i(\lambda_i)\} \leq L_m < \infty$, *where* $L_m = L + mL_g$.

Proof of 1 *of Lemma* 4.1. Let $h_j(\lambda_i) = L_g \min(\lambda_{ij}, 0)$. It is clear that h_j is a piecewise linear function. Recall that

$$g_i(\lambda_i) = f_i(\lambda_i) + \sum_{j=1}^{m} h_j(\lambda_i).$$

The concavity of g_i follows from the concavity of f and the functions h_j, $j = 1, \ldots, m$. The points of the nondifferentiability of f_i form a set of measure zero. The other points of nondifferentiability of g_i are points λ_i, where $\lambda_{ij} = 0$ for some j. These points form a set of measure zero. Thus g_i is differentiable outside a set of measure zero.

Proof of 2 *of Lemma* 4.1. Let $\mathbf{e}_j$ be the vector whose jth component is 1 and other components are 0. It is clear that, for λ_i with $\lambda_{ij} \neq 0$, h_j is differentiable and $\nabla h_j(\lambda_i) = L_g\mathbf{e}_j$ if $\lambda_{ij} < 0$ and $\nabla h_j(\lambda_i) = \mathbf{0}$ if $\lambda_{ij} > 0$, where $\mathbf{0}$ is the m-dimensional zero vector. For λ_i with $\lambda_{ij} = 0$, $\partial h_j(\lambda_i)$ consists of vectors of the form $\bar{L}\mathbf{e}_j$, where $0 \leq \bar{L} \leq L_g$. Thus, for all j, $\sup_{\lambda_i}\{\|v\|_1 : v \in \partial h_j(\lambda_i)\} = L_g$. It is known from the theory of convex functions that if $u = \sum_{j=1}^{k} u_j$ where u_j, $j = 1, \ldots, k$ are convex functions, then $\partial u(x) = \sum_{j=1}^{k} \partial u_j(x)$. Thus, if $\sup_x\{\|v\|_1 : v \in \partial u_j(x)\} \leq L_j$, then $\sup_x\{\|v\|_1 : v \in \partial u(x)\} \leq \sum_{j=1}^{k} L_j$. By assumption, $\sup_{\lambda_i}\{\|v\|_1 : v \in \partial f_i(\lambda_i)\} \leq L$. Hence $\sup_{\lambda_i}\{\|v\|_1 : v \in \partial g_i(\lambda_i)\} \leq L + \sum_{i=1}^{m} L_g = L + mL_g$.

LEMMA 4.2. *The set of optimal solutions for* (1) *with* g *as the objective function is the same as the set of optimal solutions to* (7) *with* f *as the objective function.*

Proof. Without loss of generality assume that $B > 0$. Consider some optimal solution λ^* for (7) with f as the objective function. Suppose there exists some feasible solution $\hat{\lambda}$ to (1), with $\hat{\lambda}_{ij} < 0$ for some i, j. We show that $g(\hat{\lambda}) < g(\lambda^*)$. This implies that solving (7) with g as the objective function is equivalent to solving (1) with g as

the objective function. Since $g(\lambda) = f(\lambda)$ when $\lambda \geq 0$, solving (7) with g is equivalent to solving (7) with f. Thus the set of optimal solutions for (1) with g and for (7) with f are the same proving the lemma.

Consider the following problem,

$$\begin{aligned} \max_{\lambda_p \in \Re^m, p=1,\ldots,n} g(\lambda) &= \frac{1}{n} \sum_{p=1}^{n} g_p(\lambda_p) \\ \text{s.t.} \qquad & \sum_{p=1}^{n} \lambda_p = B, \\ & \lambda_p \geq -|\hat{\lambda}_p|, p = 1, \ldots, n. \end{aligned} \tag{26}$$

It can be seen that λ^* and $\hat{\lambda}$ are feasible solutions to (26). We now show that $\hat{\lambda}$ cannot be an optimal solution to (26). Since $B > 0$, there exists some k such that $\hat{\lambda}_{kj} > 0$. Define $\bar{\lambda}$ so that it differs from $\hat{\lambda}$ only in the ij and kj components as follows:

$$\bar{\lambda}_{ij} = \hat{\lambda}_{ij} + \delta,$$

$$\bar{\lambda}_{kj} = \hat{\lambda}_{kj} - \delta.$$

We choose a $\delta > 0$ such that $\bar{\lambda}_{kj} > 0$ and $\bar{\lambda}_{ij} < 0$. It is clear that $\bar{\lambda}$ is a feasible solution to (26). We now have

$$\begin{aligned} g_i(\bar{\lambda}_i) &= f_i(\bar{\lambda}_i) + \sum_{l=1}^{m} h_l(\bar{\lambda}_i) \\ &\geq g_i(\hat{\lambda}_i) + \delta \left(L_g + \left(\nabla f_i(\bar{\lambda}_i) \right)_j \right). \end{aligned}$$

The inequality comes from the concavity of g_i and from the definition of $\bar{\lambda}$. Similarly

$$\begin{aligned} g_k(\bar{\lambda}_k) &= f_k(\bar{\lambda}_k) + \sum_{l=1}^{m} h_l(\bar{\lambda}_k) \\ &\geq g_k(\hat{\lambda}_k) - \delta \left(\nabla f_k(\bar{\lambda}_k) \right)_j. \end{aligned}$$

Hence

$$g_i(\bar{\lambda}_i) + g_k(\bar{\lambda}_k) \geq g_i(\hat{\lambda}_i) + g_k(\hat{\lambda}_k) + \delta \left(L_g + \left(\nabla f_i(\bar{\lambda}_i) \right)_j - \left(\nabla f_k(\bar{\lambda}_k) \right)_j \right).$$

Since $L_g > 2L$, we can conclude from the above that $g(\bar{\lambda}) > g(\hat{\lambda})$, and hence $\hat{\lambda}$ cannot be an optimal solution to (26).

It can be seen from the definition of (26) that its feasible set is bounded. Since g is continuous and since the feasible set of (26) is bounded and closed, it has at least one optimal solution by the extreme value theorem. The above argument, presented to establish the nonoptimality of $\hat{\lambda}$ for (26), holds for any feasible solution for (26) with at least one nonnegative component. Hence all optimal solutions of (26) are nonnegative. Since $g(\lambda) = f(\lambda)$ when $\lambda \geq 0$, solving (26) with g is equivalent to solving (7) with f, and hence λ^* is an optimal solution for (26). Hence $g(\hat{\lambda}) < g(\bar{\lambda}) \leq g(\lambda^*)$. □

Acknowledgments. The second author thanks Dimitri Bertsekas, Rob Freund, and John Tsitsiklis for their helpful comments and suggestions.

REFERENCES

[1] K. Arrow and F. Hahn, *General Competitive Analysis*, Holden Day, San Francisco, 1971.
[2] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1999.
[3] D. Bertsimas and J. N. Tsitsiklis, *Introduction to Linear Optimization*, Athena Scientific, Belmont, MA, 1997.
[4] R. Durrett, *Probability: Theory and Examples*, Duxbury Press, Belmont, CA, 1995.
[5] G. M. Heal, *Planning without prices*, Rev. Econom. Stud., 63 (1969), pp. 343–362.
[6] J. Kurose and R. Simha, *A microeconomic approach to optimal resource allocation computer systems*, IEEE Trans. Comput., 38 (1989).
[7] H. Lakshmanan and D. P. De Farias, *Decentralized approximate dynamic programming for dynamic networks of agents*, in Proceedings of the American Control Conference, Minneapolis, MN, 2006.
[8] L. Xiao and S. Boyd, *Optimal scaling of a gradient method for distributed resource allocation*, J. Optim. Theory Appl., 129 (2006), pp. 469–488.
[9] A. Nedic and D. P. Bertsekas, *Incremental subgradient methods for nondifferentiable optimization*, SIAM J. Optim., 12 (2001), pp. 109–138.
[10] Yu. Nesterov, *Smooth minimization of non-smooth functions*, Math. Program., 103 (2005), pp. 127–152.
[11] L. Servi, Y. C. Ho, and R. Suri, *A class of center-free resource allocation algorithms*, Large Scale Systems, 1 (1980), pp. 51–62.
[12] T.-S. Tang and M. A. Styblinsky, *Yield optimization for nondifferentiable density functions using convolution techniques*, IEEE Trans. Comput. Aided Design, 7 (1988), pp. 1053–1067.
[13] T. Ibaraki and N. Katoh, *Resource Allocation Problems: Algorithmic Approaches*, MIT Press, Cambridge, MA, 1988.
[14] J. N. Tsitsiklis, *Problems in Decentralized Decision Making and Computation*, Ph.D. thesis, MIT, Cambridge, MA, 1984.

SIAM J. OPTIM.
Vol. 19, No. 2, pp. 941–951

GLOBAL OPTIMIZATION OF POLYNOMIALS USING THE TRUNCATED TANGENCY VARIETY AND SUMS OF SQUARES*

HÀ HUY VUI† AND PHẠM TIẾN SƠN‡

Dedicated to Professor Hoàng Tụy on his eightieth birthday

Abstract. This paper proposes a method for finding the global infimum of a multivariate polynomial f via sum of squares (SOS) relaxation over its truncated tangency variety. This variety is truncated of the set of all points $x \in \mathbb{R}^n$ where the level sets of f are tangent to the sphere in $\mathbb{R}^n$ centered in the origin and with radius $\|x\|$. It is demonstrated that:

- The infimum of f on $\mathbb{R}^n$ and on its truncated tangency variety coincide.
- A sums of squares certificate for nonnegativity of f on its truncated tangency variety.

These facts imply that we can find a natural sequence of semidefinite programs whose optimal values converge monotonically, increasing to the infimum of f. This opens up the possibility of solving previously intractable polynomial optimization problems.

Key words. global optimization, polynomials, sum of squares (SOS), semidefinite program (SDP), tangency variety.

AMS subject classifications. Primary, 13J30, 90C26; Secondary, 12Y05, 13P99, 14P10, 90C22

DOI. 10.1137/080719212

1. Introduction.

We consider the global optimization problem

$$f^* := \inf\{f(x) \mid x \in \mathbb{R}^n\} \in \mathbb{R} \cup \{-\infty\},$$

where $x := (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ is a real vector, and f is a multivariate polynomial. As is well known, the above optimization problem is NP-hard even when the degree of f is fixed to be four [16]. Since f^* is the greatest lower bound of f, it is equivalent to compute

$$f^* = \sup\{a \in \mathbb{R} \mid f - a \geq 0 \text{ on } \mathbb{R}^n\} \in \mathbb{R} \cup \{-\infty\}.$$

A lower bound can be computed efficiently using the sum of squares (SOS) relaxation

$$f^{sos} := \sup\{a \in \mathbb{R} \mid f - a \gtrsim_{sos} 0\} \in \mathbb{R} \cup \{-\infty\},$$

where the inequality $g \gtrsim_{sos} 0$ means that the polynomial g is SOS, i.e., a sum of squares of other polynomials. We refer to [10], [11], [18] [19], [20], and [21] for introductions to SOS techniques and their applications. The above SOS relaxation can be expressed as a *semidefinite program* (SDP for short), i.e., as the problem of minimizing (or maximizing) an affine linear function over the intersection of the cone of all positive semidefinite matrices with an affine subspace. In recent years, SDPs (see [27] for an introduction) have become more and more popular for obtaining good lower bounds (or even an optima solution) for global optimization problems with polynomials. For instance, in papers [3], [12], [17], [20], and [25], the authors have proposed

*Received by the editors March 24, 2008; accepted for publication (in revised form) April 24, 2008; published electronically September 11, 2008.
http://www.siam.org/journals/siopt/19-2/71921.html
†Institute of Mathematics, 18, Hoang Quoc Viet Road, Cau Giay District 10307, Hanoi, Vietnam (hhvui@math.ac.vn).
‡Department of Mathematics, University of Dalat, Dalat, Vietnam (pham_ts@yahoo.co.uk).

several methods to find semidefinite relaxations relying on sums of squares certificates and critical point theory. As one could expect, polynomials that do not attain a minimum on $\mathbb{R}^n$ (that are either unbounded from below or have a finite infimum that is not attained) are particularly hard to handle. In [3], this problem (among others) was solved by perturbing the coefficients of the polynomial to guarantee a minimum (in particular, boundedness from below). Though the results in [3] are quite good, we are convinced that one should also look for other methods that avoid perturbations and the danger of numerical ill-conditioning that comes along with them.

By proving SOS representations for polynomials positive on their real gradient variety, it was shown by Nie, Demmel and Sturmfels [17] that an approach without perturbation is possible. Recall that, the *real gradient variety* of a polynomial f, denoted by $V(\nabla f)$, is defined to be the following algebraic set:

$$V(\nabla f) := \{x \in \mathbb{R}^n \mid \nabla f(x) = 0\};$$

the *gradient ideal* of f is the ideal in $\mathbb{R}[x]$ generated by all partial derivatives of f:

$$(\nabla f) := \left\langle \frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n} \right\rangle \subseteq \mathbb{R}[x].$$

Here and in the following, we write $\mathbb{R}[x] := \mathbb{R}[x_1, x_2, \ldots, x_n]$ for the ring of all polynomials in n variables x with real coefficients. It has been shown that if the polynomial $f \in \mathbb{R}[x]$ is nonnegative on $V(\nabla f)$, then f is a SOS modulo its gradient ideal in the case where the ideal is radical. In the general case where the gradient ideal is not necessarily radical, the same thing still holds for polynomials positive on their real gradient variety. The following is essentially [17, Theorems 8 and 9] (confer also with recent works [13] and [25]).

THEOREM 1.1 (Nie, Demmel, and Sturmfels [17]). *For every $f \in \mathbb{R}[x]$ attaining a minimum on $\mathbb{R}^n$, the following are equivalent.*

(i) $f \geq 0$ *on* $\mathbb{R}^n$;
(ii) $f \geq 0$ *on* $V(\nabla f)$;
(iii) *For all $\epsilon > 0$, there exists a SOS s in $\mathbb{R}[x]$ and polynomials $\phi_1, \phi_2, \ldots, \phi_n$ such that*

$$f(x) + \epsilon = s(x) + \phi_1(x)\frac{\partial f}{\partial x_1} + \phi_2(x)\frac{\partial f}{\partial x_2} + \cdots + \phi_n(x)\frac{\partial f}{\partial x_n}.$$

Moreover, (ii) *and* (iii) *are equivalent for all $f \in \mathbb{R}[x]$.*

For each degree restriction $k \in \mathbb{N}$, the problem of computing the supremum over all $a \in \mathbb{R}$ such that

$$f(x) - a = s(x) + \phi_1(x)\frac{\partial f}{\partial x_1} + \phi_2(x)\frac{\partial f}{\partial x_2} + \cdots + \phi_n(x)\frac{\partial f}{\partial x_n}$$

for some SOS s in $\mathbb{R}[x]$ and polynomials $\phi_1, \phi_2, \ldots, \phi_n$ of degree at most k can be expressed as an SDP. Theorem 1.1 shows that the optimal values of the corresponding sequence of SDPs (indexed by k) tend to f^* provided that f attains a minimum on $\mathbb{R}^n$. However, for polynomials that do not attain a minimum, their method yields wrong answers (see, for example, [25]). In [17, section 7], the authors write:

> "This paper proposes a method for minimizing a multivariate polynomial $f(x)$ over its gradient variety. We assume that the infimum f^* is attained. This assumption is nontrivial, and we do not address the (important and difficult) question of how to verify that a given polynomial $f(x)$ has this property."

Shortly thereafter, combining considerable machinery from differential geometry and real algebraic geometry, Schweighofer [25] has shown that part of this limitation can be removed. Let us outline Schweighofer's approach. First, we introduce some definitions.

DEFINITION 1.1. *For any polynomial $f \in \mathbb{R}[x]$ and subset $S \subset \mathbb{R}^n$, the set $R_\infty(f,S)$ of asymptotic values of f on S consists of all $y \in \mathbb{R}$ for which there exists a sequence $\{x^k\}_{k\in\mathbb{N}}$ of points $x^k \in S$ such that $\lim_{k\to\infty} \|x^k\| = +\infty$ and $\lim_{k\to\infty} f(x^k) = y$.*

DEFINITION 1.2. *The preordering generated by polynomials $g_1, g_2, \ldots, g_m \in \mathbb{R}[x]$, denoted by $T(g_1, g_2, \ldots, g_m)$, is defined to be the following set of polynomials:*

$$T(g_1, g_2, \ldots, g_m) := \left\{ \sum_{\delta\in\{0,1\}^m} s_\delta g_1^{\delta_1} g_2^{\delta_2} \ldots g_m^{\delta_m} \mid s_\delta \text{ is a sum of squares in } \mathbb{R}[x] \right\}.$$

By definition, the elements of $T(g_1, g_2, \ldots, g_m)$ have obviously the geometric property that they are nonnegative on the *basic closed semialgebraic set*

$$\{x \in \mathbb{R}^n \mid g_1(x) \geq 0, g_2(x) \geq 0, \ldots, g_m(x) \geq 0\}.$$

The next theorem is a partial converse.

THEOREM 1.2 (see [25, Theorem 9]). *Let $f, g_1, g_2, , \ldots, g_m \in \mathbb{R}[x]$ and set*

$$S := \{x \in \mathbb{R}^n \mid g_1(x) \geq 0, g_2(x) \geq 0, \ldots, g_m(x) \geq 0\}.$$

Suppose that

(i) *f is bounded on S;*
(ii) *$R_\infty(f,S)$ is a finite subset of $\mathbb{R}_{>0} := \{y \in \mathbb{R} \mid y > 0\}$; and*
(iii) *$f > 0$ on S.*

Then $f \in T(g_1, g_2, \ldots, g_m)$.

DEFINITION 1.3. *For a polynomial $f \in \mathbb{R}[x]$, we call*

$$S(\nabla f) := \{x \in \mathbb{R}^n \mid 1 - \|\nabla f(x)\|^2 \|x\|^2 \geq 0\}$$

the principal gradient tentacle of f. Here and in the following, we use the notation $\|x\| := \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$.

Given a polynomial f for which you want to compute f^*, the game will consist in finding a tentacle such that two things will hold at the same time:

- There exist suitable sums of squares certificates for nonnegativity on the tentacle; and
- The infimum of f on $\mathbb{R}^n$ and on the tentacle coincide.

These two properties seem to be hardly compatible. However, there is a sufficient condition to ensure that it is indeed the case.

DEFINITION 1.4 (see [22]). *We say that a polynomial $f \in \mathbb{C}[x]$ has only isolated singularities at infinity if $f \in \mathbb{C}$ (i.e., f is constant) or $d := \deg f \geq 1$ and there are only finitely many $z \in \mathbb{P}^{n-1}(\mathbb{C})$ such that*

$$\frac{\partial f_d}{\partial x_1} = \frac{\partial f_d}{\partial x_2} = \cdots = \frac{\partial f_d}{\partial x_n} = f_{d-1} = 0,$$

where $f = \sum_{i=0}^d f_i$ and each $f_i \in \mathbb{C}[x]$ is zero or homogeneous of degree i.

The above notion appears in the study of the topology of polynomial mappings. One may consult, for example, [1], [7], or [26] for more details.

The following result is a sums of squares certificate for nonnegativity of f on its principal gradient tentacle, which is suitable for optimization purposes. Its proof is based on Theorem 1.2.

THEOREM 1.3 (see [25, Theorem 25]). *Let $f \in \mathbb{R}[x]$ be bounded from below. Furthermore, suppose that f has only isolated singularities at infinity (which is always true in the case $n = 2$) or the principal gradient tentacle $S(\nabla f)$ is compact. Then the following are equivalent.*

(i) $f \geq 0$ *on* $\mathbb{R}^n$;
(ii) $f \geq 0$ *on* $S(\nabla f)$;
(iii) *For every* $\epsilon > 0$, *there are sums of squares of polynomials* s *and* t *in* $\mathbb{R}[x]$ *such that*

$$f(x) + \epsilon = s(x) + t(x)[1 - \|\nabla f(x)\|^2 \|x\|^2].$$

For each $k \in \mathbb{N}$, we define $f_k^* \in \mathbb{R} \cup \{\pm\infty\}$ as the supremum over all $a \in \mathbb{R}$ such that $f - a$ can be written as a sum

$$f(x) - a = s(x) + t(x)[1 - \|\nabla f(x)\|^2 \|x\|^2],$$

where s and t are sums of squares of polynomials with $\deg t \leq 2k$. Here and in the following, we use the convention that the degree of the zero polynomial is $-\infty$ so that $t = 0$ is allowed in the above definition. Then it is easy to see that Theorem 1.3 can be expressed in terms of the sequence $f_0^*, f_1^*, f_2^*, \ldots$ as follows.

THEOREM 1.4 (see [25, Theorem 30]). *Let $f \in \mathbb{R}[x]$ be bounded from below. Suppose that f has only isolated singularities at infinity (e.g., $n = 2$) or the principle gradient tentacle $S(\nabla f)$ is compact. Then the sequence $\{f_k^*\}_{k\in\mathbb{N}}$ converges monotonically increasing to f^*.*

As is well known, the problem of computing the supremum f_k^* can be translated into an SDP as described in [19], [20], and [21]. Theorem 1.4 shows that the optimal values of the corresponding sequence of SDPs (indexed by k) tend to f^* provided that f has only isolated singularities at infinity or the principal gradient tentacle $S(\nabla f)$ is compact. Unfortunately, it is not clear that these technical assumptions are necessary or not. So, Schweighofer presented in [25] a collection of higher gradient tentacles (see [25, Definition 41]) defined by the polynomial inequalities

$$1 - \|\nabla f(x)\|^{2N}(1 + \|x\|^2)^{N+1} \geq 0, \quad N \in \mathbb{N}.$$

Their advantage is that we have a sums of squares representation theorem ([25, Theorem 46]) applicable for *all* $f \in \mathbb{R}[x]$ bounded from below and N is large enough. However, the degree of the defining polynomial inequality for the Nth tentacle in this sequence will be roughly $2N$ times the degree of f. This has the disadvantage that the corresponding SDP relaxations get very big for large N. Also, we have to deal for each N with a sequence of SDPs. All in all, we have, therefore, a double sequence of SDPs. The following was formulated by Schweighofer as an open problem [25, Open Problem 33]:

Problem. *Do Theorems* 1.3 *and* 1.4 *hold without the hypothesis that f has only isolated singularities at infinity or the principal gradient tentacle $S(\nabla f)$ is compact?*

The aim of this paper is to find a solution valid to this problem in the general case, that is, when

- polynomials do not necessarily attain a minimum; and
- polynomials do not necessarily have isolated singularities at infinity.

Although we do not provide a direct answer to the problem, we will replace gradient tentacles by a semialgebraic subset of $\mathbb{R}^n$ for which a new sums of squares representation result of nonnegative polynomials is established. This semialgebraic subset will be called the *truncated tangency variety.* We show that the infimum of any polynomial $f \in \mathbb{R}[x]$ on $\mathbb{R}^n$ will coincide with the infimum on its truncated tangency variety (see Theorem 2.1 and Corollary 3.1). Then, we prove a general sums of squares certificate for nonnegativity of f on its truncated tangency variety which is suitable for optimization purposes. This representation theorem (Theorem 3.1) is of independent interest, and its proof is mainly based on Theorem 1.2.

The paper is organized as follows. The notion about the tangency variety, which plays an important role in the results, is recalled in section 2. The main result and its proof are given in section 3. In section 4 we discuss numerical experiments. Section 5 draws some conclusions.

2. The tangency variety. Throughout this paper let $f \in \mathbb{R}[x]$ be a nonconstant polynomial function. Adapting Durfee's definition [2] (see also [8]), let us define the tangency variety of f by

$$\Gamma(f) := \left\{ x \in \mathbb{R}^n \mid \operatorname{rank} \begin{pmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n} \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \leq 1 \right\}.$$

Geometrically, the tangency variety $\Gamma(f)$ of f consists of all points $x \in \mathbb{R}^n$ where the level sets of f are tangent to the sphere in $\mathbb{R}^n$ centered in the origin and with radius $\|x\|$.

Remark 2.1. It is worth noting that we can replace the tangency variety $\Gamma(f)$ by the following algebraic set

$$\Gamma(a, f) := \left\{ x \in \mathbb{R}^n \mid \operatorname{rank} \begin{pmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n} \\ x_1 - a_1 & x_2 - a_2 & \cdots & x_n - a_n \end{pmatrix} \leq 1 \right\},$$

where $a := (a_1, a_2, \ldots, a_n) \in \mathbb{R}^n$. Then all subsequent results will still hold with obvious modifications. The advantage is that if the center a is general enough, then the semialgebraic set $\Gamma(a, f) \setminus V(\nabla f)$ is a one-dimensional submanifold of $\mathbb{R}^n$. For details the reader may consult [8].

As expressed by the notation $\Gamma(f)$, polynomials f with the same gradient ∇f have the same tangency variety; in other words,

$$\Gamma(f + c) = \Gamma(f) \quad \text{for all} \quad c \in \mathbb{R}.$$

The following is a simple fact about the tangency variety $\Gamma(f)$.

LEMMA 2.1. *The tangency variety* $\Gamma(f)$ *is a nonempty, unbounded, and algebraic set.*

Proof. Obviously, $\Gamma(f)$ is an algebraic set. Let

$$A := \{x \in \mathbb{R}^n \mid f(x) = \min\{f(y) \mid \|y\| = \|x\|,\ y \in \mathbb{R}^n\}\}.$$

Then, the set A is nonempty and unbounded in $\mathbb{R}^n$. Moreover, it follows from Lagrange's multipliers theorem that $A \subset \Gamma(f)$. These facts prove the lemma. ∎

Another property of the tangency variety $\Gamma(f)$ is stated in the following result.

THEOREM 2.1. *We have*

$$\inf\{f(x) \mid x \in \mathbb{R}^n\} = \inf\{f(x) \mid x \in \Gamma(f)\}.$$

Proof. The statement follows immediately from the fact that the set

$$A = \{x \in \mathbb{R}^n \mid f(x) = \min\{f(y) \mid \|y\| = \|x\|,\ y \in \mathbb{R}^n\}\}.$$

is contained in $\Gamma(f)$. □

COROLLARY 2.1. *Let $f \in \mathbb{R}[x]$ be bounded from below. Then f^* is either a critical value of f or f^* is an asymptotic value of f on $\Gamma(f)$.*

Proof. It is an immediate consequence of Theorem 2.1. □

LEMMA 2.2. *$R_\infty(f, \Gamma(f))$ is a finite set.*

Proof. This proof is adapted from [5] (see also [6]). Let us define the *set of asymptotic critical values* of f by

$$K_\infty(f) := \{y \in \mathbb{R} | \text{there exists a sequence } \{x^k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n \text{ such that } \lim_{k\to\infty} \|x^k\| = +\infty,\ \lim_{k\to\infty} f(x^k) = y, \text{ and } \lim_{k\to\infty} \|x^k\|\|\nabla f(x^k)\| = 0\}.$$

As is well known, the set $K_\infty(f)$ is always finite [9]. Hence, it suffices to show that $R_\infty(f, \Gamma(f)) \subset K_\infty(f)$. Assume that $y \in R_\infty(f, \Gamma(f))$; i.e., there exists a sequence $\{x^k\}_{k\in\mathbb{N}}$ of points $x^k \in \Gamma(f)$ such that $\lim_{k\to\infty} \|x^k\| = +\infty$ and $\lim_{k\to\infty} f(x^k) = y$. Then we can write $\nabla f(x^k) = \lambda_k x^k$ for some $\lambda_k \in \mathbb{R}$. By using a version at infinity of the Curve Selection Lemma (see [14], [15]), there exist a meromorphic curve $\varphi(\tau)$ and an analytic function $\lambda(\tau), \tau \in (0, \epsilon)$, such that

(a) $\nabla f[\varphi(\tau)] = \lambda(\tau)\varphi(\tau)$ for $\tau \in (0, \epsilon)$;
(b) $\lim_{\tau\to+0} \|\varphi(\tau)\| = +\infty$; and
(c) $\lim_{\tau\to+0} f[\varphi(\tau)] = y$.

We can write

$$f[\varphi(\tau)] - y = c_1\tau^\nu + \text{ higher order terms in } \tau,$$
$$\|\varphi(\tau)\| = c_2\tau^\rho + \text{ higher order terms in } \tau,$$

where c_1, c_2 are nonzero real numbers and

$$\nu > 0, \qquad \rho < 0.$$

We have

$$\frac{df[\varphi(\tau)]}{d\tau} = \left\langle \nabla f[\varphi(\tau)], \frac{d\varphi(\tau)}{d\tau} \right\rangle = \lambda(\tau) \left\langle \varphi(\tau), \frac{d\varphi(\tau)}{d\tau} \right\rangle.$$

(The second equality follows from Condition (a).) Hence,

$$2\frac{df[\varphi(\tau)]}{d\tau} = \lambda(\tau)\frac{d\|\varphi(\tau)\|^2}{d\tau}.$$

This, together with Condition (a), implies that

$$2\left|\frac{df[\varphi(\tau)]}{d\tau}\right| = \frac{\|\nabla f[\varphi(\tau)]\|}{\|\varphi(\tau)\|}\frac{d\|\varphi(\tau)\|^2}{d\tau}.$$

Then it is easy to see that

$$\|\nabla f[\varphi(\tau)]\|\|\varphi(\tau)\| = c\tau^\nu + \text{ higher order terms in } \tau$$

for some constant $c \neq 0$, which shows that $y \in K_\infty(f)$. The lemma is proved. □

3. Main results. Let $f \in \mathbb{R}[x]$ be a nonconstant polynomial function, and consider the following polynomials

$$g_{ij}(x) := x_j \frac{\partial f}{\partial x_i}(x) - x_i \frac{\partial f}{\partial x_j}(x), \quad 1 \le i < j \le n.$$

In what follows, we shall fix a real number $M \in f(\mathbb{R}^n)$ (for instance, we can set $M := f(0)$). Then by the truncated tangency variety of f we mean the set

$$\Gamma_M(f) := \{x \in \mathbb{R}^n \mid M - f(x) \ge 0,\ g_{ij}(x) = 0,\ 1 \le i < j \le n\}.$$

We begin with the following result of its own interest.

COROLLARY 3.1. *We have*

$$\inf\{f(x) \mid x \in \mathbb{R}^n\} = \inf\{f(x) \mid x \in \Gamma_M(f)\}.$$

Proof. Indeed, if $M = \inf_{x \in \mathbb{R}^n} f(x)$, then the truncated tangency variety $\Gamma_M(f)$ contains a global minimizer of f and there is nothing to prove. Otherwise, the result follows immediately from Theorem 2.1. ∎

LEMMA 3.1. *Let f be a polynomial in n real variables. If*

$$\inf\{f(x) \mid x \in \Gamma_M(f)\} > 0,$$

then f can be written as a sum

$$f(x) = s(x) + t(x)[M - f(x)] + \sum_{1 \le i < j \le n} \phi_{ij}(x) g_{ij}(x),$$

where $s, t, \phi_{ij} \in \mathbb{R}[x], 1 \le i < j \le n$, and s, t are sums of squares in $\mathbb{R}[x]$.

Proof. It is clear from the assumption that f is bounded and strictly positive on the truncated tangency variety $\Gamma_M(f)$. Moreover, the following inclusion holds:

$$R_\infty(f, \Gamma_M(f)) \subset R_\infty(f, \Gamma(f)).$$

This, together with Lemma 2.2, implies that $R_\infty(f, \Gamma_M(f))$ is a finite set of $\mathbb{R}_{>0}$.

On the other hand, we can write

$$\Gamma_M(f) = \{x \in \mathbb{R}^n \mid M - f(x) \ge 0,\ g_{ij}(x) \ge 0,\ -g_{ij}(x) \ge 0,\ 1 \le i < j \le n\}.$$

Therefore $f \in T(M - f, g_{ij}, -g_{ij})$ by Theorem 1.2, which completes the proof. ∎

Here comes one of the main results of this article, which is interesting on its own, but can later be read as a convergence result for a sequence of optimal values of SDPs (Theorem 3.2 below).

THEOREM 3.1. *Let f be a polynomial in n real variables. Then the following conditions are equivalent:*

(i) *$f \ge 0$ on $\mathbb{R}^n$;*
(ii) *$f \ge 0$ on $\Gamma_M(f)$;*
(iii) *For every $\epsilon > 0$, there are sums of squares of polynomials s and t in $\mathbb{R}[x]$ and polynomials $\phi_{ij}, 1 \le i < j \le n$, such that*

$$f(x) + \epsilon = s(x) + t(x)[M - f(x)] + \sum_{1 \le i < j \le n} \phi_{ij}(x) g_{ij}(x).$$

Proof. The implications (i) $\Leftrightarrow$ (ii) and (iii) $\Rightarrow$ (ii) are straightforward. For the implication (ii) $\Rightarrow$ (iii), we only have to apply Lemma 3.1 to $f+\epsilon$ instead of f. □

DEFINITION 3.1. *For all polynomials $f \in \mathbb{R}[x]$ and all $k \in \mathbb{N}$, we define $f_k^* \in \mathbb{R} \cup \{\pm\infty\}$ as the supremum over all $a \in \mathbb{R}$ such that $f-a$ can be written as a sum*

$$f(x) - a = s(x) + t(x)[M - f(x)] + \sum_{1 \le i < j \le n} \phi_{ij}(x) g_{ij}(x),$$

where s, t, ϕ_{ij} are polynomials of degree at most $2k$ and s, t are sums of squares in $\mathbb{R}[x]$.

As is well known, the problem of computing the supremum f_k^* can be reduced to an SDP. Moreover, by Theorem 3.1, the number f_k^* is a lower bound for the infimum f^* of the polynomial f, and this lower bound gets better as k increases

$$\cdots \le f_{k-1}^* \le f_k^* \le f_{k+1}^* \le \cdots \le f^*.$$

We have the following general result concerning the convergence of the lower bounds.

THEOREM 3.2. *Let f be a polynomial in n real variables. Then the sequence $\{f_k^*\}_{k\in\mathbb{N}}$ converges monotonically increasing to the infimum f^*.*

Proof. In fact, if $f^* = -\infty$, then it is easily seen from Theorem 3.1 that for every positive integer k, $f_k^* = -\infty$ and there is nothing to prove. Thus, we may as well assume that $f^* > -\infty$. Let ϵ be any positive constant. The polynomial $f - f^* + \epsilon$ is strictly positive on its truncated tangency variety $\Gamma_M(f - f^* + \epsilon) = \Gamma_M(f)$. By Theorem 3.1, there are sums of squares of polynomials s and t in $\mathbb{R}[x]$ and polynomials $\phi_{ij}, 1 \le i < j \le n$, such that

$$f(x) - f^* + \epsilon = s(x) + t(x)[M - f(x)] + \sum_{1 \le i < j \le n} \phi_{ij}(x) g_{ij}(x).$$

Hence, there exists an integer $k(\epsilon)$ such that

$$f_k^* \ge f^* - \epsilon \quad \text{for all } k \ge k(\epsilon).$$

Since the sequence $\{f_k^*\}_{k\in\mathbb{N}}$ is monotonically increasing, it follows that $\lim_{k\to\infty} f_k^* = f^*$, which completes the proof of Theorem 3.2. □

4. Numerical results. In this section we set $M = f(0)$, and the examples have been computed using the software MATLAB 7 and SOSTOOLS [23]. Most of the computations took a few seconds.

Example 4.1. Let us consider the following polynomial

$$f(x,y) := 2y^4(y+x)^4 + y^2(y+x)^2 + 2y(y+x) + y^2.$$

We have $f^* = -\frac{5}{8} = -0.6250$ and the polynomial f does not attain its infimum value. The computed optimal values of the truncated tangency relaxations are $f_0^* = -0.614$, $f_1^* = -0.59314$, $f_2^* = -0.57259$, and $f_3^* = -0.54373$. By Theorem 3.2, the sequence $f_0^*, f_1^*, f_2^*, \ldots$, converges monotonically to f^*. However, the computed values are larger than f^* so that there are obviously numerical problems. Confer [4] and [5, Example 2.1].

Example 4.2. Let us consider the following polynomial of two real variables:

$$f(x,y) := (xy-1)^2 + (x-1)^2.$$

It is easy to see that $f^* = f(1,1) = 0$. The computed optimal values of the truncated tangency relaxations are $f_0^* = 0.87482 \cdot 10^{-9}$, $f_1^* = 0.67891 \cdot 10^{-8}$, and $f_2^* = 0.87081 \cdot 10^{-8}$. Confer [6, Example 4.2].

Example 4.3. Let us consider the Motzkin polynomial

$$f(x,y) := x^2y^4 + x^4y^2 - 3x^2y^2 + 1,$$

which is nonnegative but *not* a sum of squares: $f^* = 0$ and $f^{sos} = -\infty$. The computed optimal values of the truncated tangency relaxations are $f_0^* = -30394$, $f_1^* = -0.67174$, $f_2^* = -0.15122$, and $f_3^* = 0.62693 \cdot 10^{-8}$. Confer [20, Example 2] and [25, Example 35].

Example 4.4. Let us consider the Berg polynomial $f := x^2y^2(x^2+y^2-1) \in \mathbb{R}[x,y]$. This polynomial is taken from [10]. It has global infimum value $f^* = -1/27 = -0.037037037\ldots$. However $f^{sos} = -33.157325$ is considerably smaller than f^*. If we minimize f over its truncated tangency variety with $k = 3$, then we get $f_3^* = -0.037037$. Confer [3, Example 4], [10, Example 3], [17, Example 3], and [25, Example 37].

Example 4.5. Let $f := (x^2+1)^2 + (y^2+1)^2 - 2(x+y+1)^2 \in \mathbb{R}[x,y]$. By computation, we obtain for all values f_0^*, f_1^*, f_2^* approximately -11.458. Confer [3, Example 3], [10, Example 2], and [25, Example 38].

Example 4.6. Consider the polynomial $f := (x+x^2y+x^4yz)^2 \in \mathbb{R}[x,y,z]$. It was shown by Schweighofer [25, Example 34] that the set $R_\infty(f, S(\nabla f))$ of asymptotic values of f on $S(\nabla f)$ is infinite. This fact implies that f has nonisolated singularities at infinity, and hence, it does not satisfy the assumptions of Theorem 1.3. On the other hand, it is clear that $f^* = 0$. By computation, we get $f_0^* = -0.49293 \cdot 10^{-9}$, $f_1^* = -0.17651 \cdot 10^{-6}$, and $f_2^* = -0.28173 \cdot 10^{-8}$.

Example 4.7. The Motzkin polynomial

$$f := x^2y^2(x^2+y^2-3z^2) + z^6 \in \mathbb{R}[x,y,z]$$

is nonnegative but *not* a sum of squares (see [24], [20]). We have $f^* = 0$ but $f^{sos} = -\infty$. The latter follows from the fact that f is homogeneous and not a sum of squares. By computation, we obtain for all values $f_0^* = -0.30767$, $f_1^* = -0.41528 \cdot 10^{-2}$, $f_2^* = -0.33994 \cdot 10^{-3}$, and $f_3^* = -0.15803 \cdot 10^{-3}$. Confer [17, Example 1] and [25].

Example 4.8. Consider once more the polynomial $f = (1-xy)^2 + y^2$. It does not attain its infimum $f^* = 0$ on $\mathbb{R}^2$ (see also [5], [6], and [25]). Since this polynomial is by definition a sum of squares, we have $f^{sos} = 0 = f^*$ and therefore $f_k^* = 0$ for all $k \in \mathbb{N}$ by definition. By computation, we get $f^{sos} \simeq 1.5142 \cdot 10^{-12}$, which is almost zero, but also $f_0^* = -0.12641 \cdot 10^{-3}$, $f_1^* = 0.12732 \cdot 10^{-1}$, and $f_2^* = 0.49626 \cdot 10^{-1}$, which shows that there are big numerical problems. In [25, Example 40], the author wrote:

> "We have verified that the corresponding SDPs have nevertheless been solved quite accurately. The problem is that small numerical errors in the coefficients of a polynomial can perturb its infimum quite a lot whenever the infimum is not attained (or attained very far from the origin). It should be subject to further research how to fight this problem."

Remark 4.1. One question still unanswered is whether the truncation is necessary. Numerical experiments seem to indicate that the truncation is not needed.

5. Conclusions. This paper proposes a method for computing numerically the infimum of a multivariate polynomial f over its truncated tangency variety. We have shown

- The infimum of f on $\mathbb{R}^n$ and on its truncated tangency variety coincide.
- A sums of squares certificate for nonnegativity of f on its truncated tangency variety.

These facts imply that we can find a natural sequence of SDPs whose optimal values f_k^* converge monotonically increasing to f^*.

Our method has two major problems (see also [25]). First, it turns out that solving semidefinite programs that arise from a polynomial that does not attain a minimum takes sometimes a surprisingly long time. Second, small numerical inaccuracies might lead to big changes in the infimum of a polynomial if the infimum is not attained. All two problems should be subject to further research. Polynomials not attaining a minimum remain hard to handle in practice. On the theoretical side, we have obtained new interesting characterizations of nonnegative polynomials.

Acknowledgments. Both authors wish to thank J. B. Lasserre and M. Schweighofer for many interesting and helpful discussions on the topic. We also thank the referees for their helpful comments.

REFERENCES

[1] A. Durfee, *Five definitions of critical points at infinity,* in Singularities, The Brieskorn Anniversary Volume, Progr. Math., 162 (1998), pp. 345–360.
[2] A. Durfee, *The index of gradf(x, y)*, Topology, 37 (1998), pp. 1339–1361.
[3] D. Jibetean and M. Laurent, *Semidefinite approximations for global unconstrained polynomial optimization,* SIAM J. Optim., 16 (2005), pp. 490–514.
[4] H. V. Hà, *Infimum of Polynomials and Singularity at Infinity,* in From Local to Global Optimization, A. Migdalas, P. M. Pardalos, and P. Värbrand, eds., Kluwer Academic Publishers, (2001), pp. 187–204.
[5] H. V. Hà and T. S. Phạm, *Minimizing Polynomial Functions,* Acta Math. Vietnam., 32 (2007), pp. 71–82.
[6] H. V. Hà and T. S. Phạm, *An estimation of the number of bifurcation values for real polynomials,* Acta Math. Vietnam., 32 (2007), pp. 35–47.
[7] H. V. Hà and T. S. Phạm, *Critical values of singularities at infinity of complex polynomials,* Vietnam J. Math., 36 (2008), pp. 1–38.
[8] H. V. Hà and T. S. Phạm, *On the Łojasiewicz exponent at infinity of real polynomials*, Annales Polonci Mathematici, to appear; also available online from http://publications.ictp.it.
[9] K. Kurdyka, P. Orro, and S. Simon, *Semialgebraic Sard theorem for generalized critical values,* J. Differential Geom., 56 (2000), pp. 67–92.
[10] J. B. Lasserre, *Global optimization with polynomials and the problem of moments,* SIAM J. Optim., 11 (2001), pp. 796–817.
[11] J. B. Lasserre and T. Netzer, *SOS approximations of nonnegative polynomials via simple high degree perturbations,* Math. Z., 256 (2007), pp. 99–112.
[12] M. Marshall, *Optimization of polynomial functions,* Canad. Math. Bull., 46 (2003), pp. 575–587.
[13] M. Marshall, *Representations of non-negative polynomials, degree bounds and applications to optimization,* Canad. J. Math., to appear.
[14] J. Milnor, *Singular points of complex hypersurfaces,* Ann. Math. Stud., 61, Princeton University Press, 1968.
[15] A. Némethi and A. Zaharia, *Milnor fibration at infinity,* Indag. Math., 3 (1992), pp. 323–335.
[16] Y. Nesterov, *Squared functional systems and optimization problems,* High Performance Optimization, H. Frenk et al., eds., Kluwer Academic Publishers, Norwell, MA, (2000), pp. 405–440.
[17] J. Nie, J. Demmel, and B. Sturmfels, *Minimizing polynomials via sum of squares over the gradient ideal,* Math. Prog., Ser. A, 106 (2006), pp. 587–606.

[18] J. Nie and J. Demmel, *Minimum ellipsoid bounds for solutions of polynomial systems via sum of squares,* J. Global Optim., 33 (2005), pp. 511–525.
[19] P. A. Parrilo, *Structured Semidefinite Programs and Semialgebraic Geometry Methods in Robustness and Optimization,* Ph.D. Thesis, California Institute of Technology, 2000.
[20] P. A. Parrilo and B. Sturmfels, *Minimizing Polynomial Functions,* DIMACS Series in Discrete Mathematics and Theoretical Computer Science, AMS, 60 (2003), pp. 83–99.
[21] P. A. Parrilo, *Semidefinite Programming relaxations for semialgebraic problems,* Math. Program. Ser. B, 96 (2003), pp. 293–320.
[22] A. Parusiński, *On the bifurcation set of a complex polynomial with isolated singularities at infinity,* Compos. Math., 97 (1995), pp. 369–384.
[23] S. Prajna, A. Papachristodoulou, and P. Parrilo, *SOSTOOLS User's Guide,* http:// control.ee.ethz.ch/ parrilo/SOSTOOLS/.
[24] B. Reznick, *Some concrete aspects of Hilbert's 17th problem,* Contemp. Math., 253 (2000), pp. 251–272.
[25] M. Schweighofer, *Global optimization of polynomials using gradient tentacles and sums of squares,* SIAM J. Optim., 17 (2006), pp. 920–942.
[26] M. Tibăr, *Polynomials and Vanishing Cycles,* Cambridge University Press, London, 2007.
[27] L. Vandenberghe and S. Boyd, *Semidefinite Programming,* SIAM Rev., 38 (1996), pp. 49–95.

SIAM J. OPTIM.
Vol. 19, No. 2, pp. 952–968

ON STABILITY OF MULTISTAGE STOCHASTIC PROGRAMS*

CHRISTIAN KÜCHLER†

Abstract. We study the quantitative stability of linear multistage stochastic programs under perturbations of the underlying stochastic processes. It is shown that the optimal values behave Lipschitz continuously with respect to an L^p-distance. In order to establish continuity of the recourse function with respect to the current state of the stochastic process, we assume continuity of the conditional distributions in terms of a Fortet–Mourier metric. The main stability result holds for nonanticipative approximations of the underlying process and thus represents a rigorous justification of established discretization techniques.

Key words. stochastic programming, multistage, stability, probability metrics

AMS subject classifications. 90C15, 90C31

DOI. 10.1137/070690365

Introduction. Many stochastic optimization problems of practical interest do not allow for an analytic solution, and numerical approaches require the underlying probability measure to have finite support. Whenever the initial probability measure does not meet these demands, it has to be approximated by an auxiliary measure. Thereby, it is reasonable to choose the approximating measure such that the optimal value and the set of optimal decisions of the auxiliary problem are close to those of the originial problem. Consequently, perturbation and stability analysis of stochastic programs is necessary for the development of reliable techniques for discretization and scenario reduction. While stability properties are well understood for nondynamic chance constrained and two-stage problems, cf. the recent survey by [18], it turned out that the multistage case is more intricate. Recently, the latter situation has been studied by a variety of authors, and thus the following references should not be considered to be exhaustive. Statistical bounds have been provided by [20]. Reference [13] established asymptotic stability of specific approximations for a general class of convex multistage problems in terms of epi-convergence. In doing so, he noticed that such quantitative results, as we discuss in this paper, require stronger assumptions. Indeed, the restriction on models with continuous decisions allowed [12] to establish such a quantitative stability result for their tree approximations. Reference [8] did not require regularity conditions on decisions and underlying processes. Consequently, their quantitative stability result, obtained by considering arbitrary perturbations of the underlying process, incorporates a term measuring the distance of the filtrations induced by the initial and the auxiliary process, respectively. Vanishing in the two-stage case, this filtration distance reflects the relevance of the information structure and of the nonanticipativity constraints for multistage decision problems. We refer also to [2] who studied the role of information in stochastic optimization problems and introduced and reviewed several concepts of distances between filtrations.

The recent approach of [7] aims to incorporate filtration distances into the construction of scenario trees. However, this requires some extra effort and, to the best

*Received by the editors May 2, 2007; accepted for publication (in revised form) May 8, 2008; published electronically October 16, 2008.

http://www.siam.org/journals/siopt/19-2/69036.html

†Humboldt–Universität zu Berlin, Unter den Linden 6, D–10099 Berlin, Germany (ckuechler@math.hu-berlin.de). Support by the Wiener Wissenschafts-, Forschungs- und Technologiefonds (WWTF) and by the Bundesministerium für Bildung und Forschung (BMBF) under the grant 03SF0312E is gratefully acknowledged.

of our understanding, these distances are not taken into account by a variety of established techniques. Thus, the main purpose of this paper is to provide general conditions under which these somewhat delicate terms may be omitted.

One of the main difficulties seems to be that without additional assumptions neither the recourse function nor an optimal decision depend continuously on the current state of the underlying process in general. Reference [16] showed that under weak conditions, the optimal value can be approximated by *continuous decisions*. However, while this allows one to deduce convergence results, such as those due to [13], it does not lead to quantitative estimates. For deriving *continuity of the recourse function* and bounds based on a barycentric approximation scheme, [11] required the underlying processes to be autoregressive. He also indicated that *the key element in any scenario tree construction is the discretization of the conditional probabilities*. In particular, continuous dependency of these probabilities on the current state of the underlying process is necessary for potential continuity of the recourse function and can be seen as *continuity of the available information with respect to the current state*. It is illustrated by Example 2.6 of [8] that the latter property is indispensable in order to omit any filtration distances and to obtain a good approximation of the initial process by usual techniques which are based on stagewise clustering. Thus, we impose Lipschitz continuity of the conditional distributions to verify the same regularity for the recourse function in Theorem 1. With this at hand, we estimate in Theorem 2 the gap between the optimal value and the costs of a decision that is locally *calm*. This leads to our main result, Theorem 3, which provides an upper bound for the perturbation of the optimal value.

Notation and conventions. Random variables are denoted by bold letters, for example $\boldsymbol{\xi}$ or $\boldsymbol{x}$, in contrast to their realizations (i.e., elements of their support) which are denoted by ξ or x, respectively. Given some vectors $\xi_1, \dots, \xi_t$ in $\mathbb{R}^s$ with $s, t \in \mathbb{N}$, the notation ξ^t is used for the vector $(\xi_1, \dots, \xi_t)$. Furthermore, $\|\cdot\|$ denotes the maximum norm on $\mathbb{R}^n$ for the respective value $n \in \mathbb{N}$, and we set $\|\xi^t\| \triangleq \max_{i \le t} \|\xi_i\|$.

1. Problem formulation. On a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ we consider an $\mathbb{R}^s$-valued stochastic process $\boldsymbol{\xi} = (\boldsymbol{\xi}_t)_{t=1}^T$ with time horizon $T \in \mathbb{N}$ and the associated filtration $(\mathcal{F}_t)_{t=1}^T$ defined through $\mathcal{F}_t \triangleq \sigma(\boldsymbol{\xi}^t)$ for $t = 1, \dots, T$. We assume that $\mathcal{F}_1 = \{\Omega, \emptyset\}$, $\boldsymbol{\xi} \in L^p(\Omega, \mathcal{F}, \mathbb{P})$ for every $p \in [1, +\infty)$, and set for $t = 1, \dots, T$,

$$\mathbb{P}_t \triangleq \mathbb{P}\left[\boldsymbol{\xi}_t \in \cdot\,\right], \mathbb{P}^t \triangleq \mathbb{P}\left[\boldsymbol{\xi}^t \in \cdot\,\right] \quad \text{and} \quad \Xi_t \triangleq \operatorname{supp} \mathbb{P}_t \subset \mathbb{R}^s, \Xi^t \triangleq \operatorname{supp} \mathbb{P}^t \subset \mathbb{R}^{s \cdot t}.$$

Furthermore, we consider the *costs* $b_t(\cdot)$, the *technology matrices* $A_{t,1}(\cdot)$, and the *right-hand sides* $h_t(\cdot)$, which all are assumed to depend affinely on $\xi_t \in \Xi_t$ and map into $\mathbb{R}^m$, $\mathbb{R}^{n \cdot m}$, and $\mathbb{R}^n$, respectively, for some $m, n \in \mathbb{N}$ and $t = 1, \dots, T$. Together with the nonrandom *recourse matrices* $A_{t,0} \in \mathbb{R}^{m \cdot n}$ they define the set-valued mappings

$$\begin{aligned} &M_t : X_{t-1} \times \Xi_t \rightrightarrows X_t, \\ &\quad M_t(x_{t-1}, \xi_t) \triangleq \{x_t \in X_t \ : \ A_{t,0} x_t + A_{t,1}(\xi_t) x_{t-1} = h_t(\xi_t)\}, \end{aligned}$$

where $X_t \subset \mathbb{R}^m$ are certain nonempty, closed, and polyhedral sets, with $t = 1, \dots, T$. We assume complete recourse; i.e., $M_t(x_{t-1}, \xi_t)$ is nonempty for every $x_{t-1} \in X_{t-1}$ and every $\xi_t \in \Xi_t$. The objective function is given by

$$\begin{aligned} \varphi &: \mathbb{R}^{m \cdot T} \times \Xi^T \to \mathbb{R}, \\ \varphi\left(x_1, \dots, x_T, \xi^T\right) &\triangleq \sum_{t=1}^T \langle b_t(\xi_t), x_t \rangle . \end{aligned}$$

A tuple $\boldsymbol{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_T)$ of Borel-measurable mappings $\boldsymbol{x}_t : \Xi^t \to X_t$, $t = 1, \dots, T$, is called a *feasible decision* with respect to $\boldsymbol{\xi}$, if the recourse condition

$$\boldsymbol{x}_t\left(\boldsymbol{\xi}^t\right) \in M_t\left(\boldsymbol{x}_{t-1}\left(\boldsymbol{\xi}^{t-1}\right), \boldsymbol{\xi}_t\right) \tag{1}$$

is fulfilled $\mathbb{P}-$a.s. for $t = 1, \dots, T$. The class of feasible decisions $\boldsymbol{x}$ will be denoted by $\mathcal{S}(\boldsymbol{\xi})$ and, for the sake of notational convenience, we set $x_0 = 1$.

We study the following multistage stochastic optimization problem:

$$v(\boldsymbol{\xi}) \triangleq \inf_{\boldsymbol{x} \in \mathcal{S}(\boldsymbol{\xi})} \mathbb{E}\left[\varphi(\boldsymbol{x}(\boldsymbol{\xi}), \boldsymbol{\xi})\right], \tag{2}$$

and aim to establish an upper bound for the perturbation of $v(\boldsymbol{\xi})$ when $\boldsymbol{\xi}$ is replaced by another process $\tilde{\boldsymbol{\xi}}$.

Complete recourse and the polyhedral form of M_t allow one to conclude (see Example 9.35 of [17]) that M_t is Lipschitz continuous on $X_{t-1} \times \Xi_t$ with respect to the Pompeiu–Hausdorff distance $\mathbb{d}$ in the following sense. There exists a constant $M \geq 0$ with

$$\begin{aligned} \mathbb{d}\left(M_t(x_{t-1}, \xi_t), M_t\left(\hat{x}_{t-1}, \xi_t\right)\right) &\leq M \cdot \max\{1, \|\xi_t\|\} \cdot \|\hat{x}_{t-1} - x_{t-1}\| \quad \text{and} \\ \mathbb{d}\left(M_t(x_{t-1}, \xi_t), M_t\left(x_{t-1}, \hat{\xi}_t\right)\right) &\leq M \cdot \max\{1, \|x_{t-1}\|\} \cdot \left\|\hat{\xi}_t - \xi_t\right\|, \end{aligned}$$

for every $(x_{t-1}, \xi_t), (\hat{x}_{t-1}, \hat{\xi}_t) \in X_{t-1} \times \Xi_t$. We recall that the Pompeiu–Hausdorff distance between two sets $A, B \subset \mathbb{R}^m$ is defined by

$$\mathbb{d}(A, B) \triangleq \max\left\{\sup_{a \in A} \operatorname{dist}(a, B), \sup_{b \in B} \operatorname{dist}(b, A)\right\}.$$

Remark 1.1. Throughout this paper, the linearity of M_t is used only to obtain the claimed Lipschitz continuity. Analogously, we assume linear costs $\langle b_t(\xi_t), x_t\rangle$ only to ensure the existence of a constant $B \geq 0$ with

$$\begin{aligned} \|\langle b_t(\xi_t), x_t\rangle - \left\langle b_t\left(\hat{\xi}_t\right), x_t\right\rangle\| &\leq B\left\|\xi_t - \hat{\xi}_t\right\|\|x_t\| \quad \text{and} \\ \|\langle b_t(\xi_t), x_t\rangle - \langle b_t(\xi_t), \hat{x}_t\rangle\| &\leq B \max\{1, \|\xi_t\|\}\|x_t - \hat{x}_t\|. \end{aligned}$$

The integrability condition on $\boldsymbol{\xi}$ is assumed for notational simplicity. Actually, it suffices to have $\boldsymbol{\xi} \in L^p(\Omega, \mathcal{F}, \mathbb{P})$ for a sufficiently large $p \in \mathbb{R}_+$.

Furthermore, all results remain valid if M_t, h_t, and b_t depend on $\boldsymbol{\xi}^t$ instead of $\boldsymbol{\xi}_t$.

2. Continuity of the recourse function. Let $V_t : \Xi^t \times X_{t-1} \to \mathbb{R}$ be the recourse function at time t, which is defined recursively by $V_{T+1} \triangleq 0$ and the dynamic programming equation

$$V_t\left(\xi^t, x_{t-1}\right) \triangleq \inf_{x_t \in M_t(x_{t-1}, \xi_t)} \langle b_t(\xi_t), x_t\rangle + \mathbb{E}\left[V_{t+1}\left(\boldsymbol{\xi}^{t+1}, x_t\right) \middle| \boldsymbol{\xi}^t = \xi^t\right] \quad \text{for } t = T, \dots, 1,$$

where the mapping $(x_t, \xi^t) \mapsto \mathbb{E}[V_{t+1}(\boldsymbol{\xi}^{t+1}, x_t)|\boldsymbol{\xi}^t = \xi^t]$ denotes the *regular conditional expectation* of $V_{t+1}(\cdot, \cdot)$ relative to $\mathcal{F}_t$. The value $V_t(\xi^t, x_{t-1})$ represents the minimal achievable expected future costs after having chosen $\boldsymbol{x}_{t-1} = x_{t-1}$, having observed $\{\boldsymbol{\xi}^t = \xi^t\}$, and before deciding on x_t. In particular, we have the identity $v(\boldsymbol{\xi}) = V_1(\boldsymbol{\xi}_1, x_0)$, and complete recourse implies that $V_t < +\infty$ holds true on $\Xi^t \times X_{t-1}$. It was proved by [6] that V_t is well-defined and measurable under the following assumption.

Assumption 2.1.

(i) There exists an integrable random variable $\boldsymbol{\eta}$ such that $\varphi(x, \boldsymbol{\xi}) \geq \boldsymbol{\eta}$ holds $\mathbb{P}-$a.s. for every $x \in \mathbb{R}^{m \cdot T}$.

(ii) For each $c \in \mathbb{R}$ the random level set $\{x \in \mathbb{R}^{m \cdot T} : \ \varphi(x, \boldsymbol{\xi}) \leq c\}$ is compact $\mathbb{P}-$a.s.

A decision $\boldsymbol{x} \in \mathcal{S}(\boldsymbol{\xi})$ is optimal if and only if the equality

$$V_t(\xi^t, \boldsymbol{x}_{t-1}(\xi^{t-1})) = \big\langle b_t(\xi_t), \boldsymbol{x}_t(\xi^t) \big\rangle + \mathbb{E}\left[V_{t+1}\left(\boldsymbol{\xi}^{t+1}, \boldsymbol{x}_t(\xi^t)\right) \big| \, \boldsymbol{\xi}^t = \xi^t \right] \tag{3}$$

holds for $\mathbb{P}^t-$almost every $\xi^t \in \Xi^t$ and $t = 1, \ldots, T$. Moreover, for every Borel measurable mapping $\boldsymbol{x}_{t-1} : \Xi^{t-1} \to X_{t-1}$, there exists a measurable $\boldsymbol{x}_t : \Xi^t \to X_t$ such that relation (3) holds true for $\mathbb{P}^t-$almost every $\xi^t \in \Xi^t$. Actually, [6] allows one to show that the $\mathbb{P}^t$-null sets on which the latter property does not hold coincide for all measurable $\boldsymbol{x}_{t-1}$. Indeed, the following corollary is an immediate consequence of applying Lemma 4 of [6] within the proof of his Theorem 2.

COROLLARY 2.2. *For $t = 1, \ldots, T$, there is a Borel set $A^{t'} \subset \Xi^t$ with $\mathbb{P}^t[A^{t'}] = 1$ such that the following property holds.*

For every Borel measurable mapping $\boldsymbol{x}_{t-1} : \Xi^{t-1} \to X_{t-1}$ there exists a measurable $\boldsymbol{x}_t : \Xi^t \to X_t$ such that identity (3) *holds true for every $\xi^t \in A^{t'}$.*

We assume the decision $\boldsymbol{x}_t$ can be chosen to fulfill a certain growth condition:

Assumption 2.3. For $t = 1, \ldots, T$, there is a Borel set $A^{t'} \subset \Xi^t$ with $\mathbb{P}^t[A^{t'}] = 1$ and a constant $L \geq 1$ such that the following property holds.

For every Borel measurable mapping $\boldsymbol{x}_{t-1} : \Xi^{t-1} \to X_{t-1}$ there exists a measurable $\boldsymbol{x}_t : \Xi^t \to X_t$ such that identity (3) and the growth condition

$$\|\boldsymbol{x}_t(\xi^t)\| \leq L \cdot \max\left\{1, \|\boldsymbol{x}_{t-1}(\xi^{t-1})\|\right\} \cdot \max\left\{1, \|\xi^t\|\right\} \tag{4}$$

hold true for every $\xi^t \in A^{t'}$.

Remark 2.4. Unfortunately, the existence of decisions which are bounded in the above sense may be hard to verify, in general. However, (4) holds true for *every* $x_t \in M_t(x_{t-1}, \xi_t)$ if X_t is bounded, or, more generally, whenever the projection of X_t onto the kernel of the recourse matrix $A_{t,0}$ is bounded.

Furthermore, the linear growth condition (4) could be relaxed to polynomial growth, and then the growth rate in ξ^t of the Lipschitz constant in Theorem 1 and the subsequent results would change accordingly.

Assumptions 2.1 and 2.3 imply the existence of an optimal decision $\boldsymbol{x}$ satisfying

$$\|\boldsymbol{x}_t(\boldsymbol{\xi}^t)\| \leq L^t \cdot \max\left\{1, \|\boldsymbol{\xi}^t\|\right\}^{t-1} \qquad \mathbb{P}-\text{a.s. for } t = 1, \ldots, T. \tag{5}$$

Indeed, a tuple $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$ of mappings with (1) and (3)–(5) can be constructed by recursion, and from Theorem 14.37 of [17] it follows that every $\boldsymbol{x}_t$ can be chosen to be measurable. Consequently, $\boldsymbol{x}$ is an optimal decision. Decisions fulfilling (4) and (5) will be denoted as *bounded* in the following.

To establish a quantitative stability result, we will study the continuity of V_t. Thereby, regularity properties of the mapping $x_{t-1} \mapsto V_t(\xi^t, x_{t-1})$ are well-known. We refer to [3] and [19] who derived convexity as well as piecewise linearity for the case of finite Ξ^T and to [11] who proved continuity under compactness assumptions on Ξ^T and $X_1, \ldots, X_T$. Thus, the following proposition can be seen as an adaption of these results to our Lipschitz continuous framework.

PROPOSITION 2.5. *The recourse function V_t is Lipschitz continuous with respect to the decision x_{t-1} in the following sense. For $t = 1, \ldots, T$, there exists a constant*

$\bar{M} > 0$ and a Borel set $A^{t''} \subset \Xi^t$ with $\mathbb{P}^t[A^{t''}] = 1$, such that for every $\xi^t \in A^{t''}$ the relation

$$\left|V_t(\xi^t, x_{t-1}) - V_t(\xi^t, \hat{x}_{t-1})\right| \leq [V_t]^x_{Lip}(\xi^t) \cdot \|x_{t-1} - \hat{x}_{t-1}\| \tag{6}$$

holds true for every $x_{t-1}, \hat{x}_{t-1} \in X_{t-1}$ with a (random) Lipschitz constant $[V_t]^x_{Lip}(\xi^t)$ satisfying

$$[V_t]^x_{Lip}(\xi^t) \leq \bar{M} \cdot \mathbb{E}\left[\max\left\{1, \left\|\boldsymbol{\xi}^T\right\|\right\}^{2+T-t} \;\middle|\; \boldsymbol{\xi}^t = \xi^t\right]. \tag{7}$$

Proof. The assertion is true for $V_{T+1} \equiv 0$. Assume it is true also for $s = t+1, \ldots, T$ with Lipschitz constants $[V_s]^x_{Lip}$ and that the difference on the left side of (6) is negative. Then, due to (3), there exists an $\boldsymbol{x}^*_t(\xi^t) \in M_t(x_{t-1}, \xi_t)$, such that the left side of (6) coincides for $\mathbb{P}^t$-a.e. ξ^t with

$$\begin{aligned} &- \left\langle b_t(\xi_t), \boldsymbol{x}^*_t\left(\xi^t\right)\right\rangle - \mathbb{E}\left[V_{t+1}\left(\boldsymbol{\xi}^{t+1}, \boldsymbol{x}^*_t\left(\xi^t\right)\right)\middle|\, \boldsymbol{\xi}^t = \xi^t\right] \\ &\quad + \inf_{\hat{x}_t \in M_t(\hat{x}_{t-1}, \xi_t)} \left\{\langle b_t(\xi_t), \hat{x}_t\rangle + \mathbb{E}\left[V_{t+1}\left(\boldsymbol{\xi}^{t+1}, \hat{x}_t\right)\middle|\, \boldsymbol{\xi}^t = \xi^t\right]\right\}. \end{aligned} \tag{8}$$

Moreover, it follows from Corollary 2.2 that we may assume that the $\mathbb{P}^t(d\xi^t)$-null sets on which this identity does not hold coincide for all $x_{t-1} \in X_{t-1}$. Due to Theorem 14.37 of [17] we can choose a measurable $\hat{\boldsymbol{x}}^*_t$ with

$$\hat{\boldsymbol{x}}^*_t(\xi^t) \in \arg\min_{z \in M_t(\hat{x}_{t-1}, \xi_t)} \left\|z - \boldsymbol{x}^*_t(\xi^t)\right\|$$

to estimate (8) from above by

$$|\left\langle b_t(\xi_t), \boldsymbol{x}^*_t\left(\xi^t\right) - \hat{\boldsymbol{x}}^*_t\left(\xi^t\right)\right\rangle| + |\mathbb{E}\left[V_{t+1}\left(\boldsymbol{\xi}^{t+1}, x^*_t\right) - V_{t+1}\left(\boldsymbol{\xi}^{t+1}, \hat{x}^*_t\right)\middle|\, \boldsymbol{\xi}^t = \xi^t\right]|.$$

From the linear growth of b_t and the Lipschitz continuity of V_{t+1} with respect to x_t, one concludes that this term is not greater than

$$\left(B\max\{1, \|\xi_t\|\} + \mathbb{E}\left[[V_{t+1}]^x_{Lip}\left(\boldsymbol{\xi}^{t+1}\right)\middle|\, \boldsymbol{\xi}^t = \xi^t\right]\right) \cdot \left\|\boldsymbol{x}^*_t\left(\xi^t\right) - \hat{\boldsymbol{x}}^*_t\left(\xi^t\right)\right\|,$$

again $\mathbb{P}^t(d\xi^t)$–a.s. for every $x_{t-1}, \hat{x}_{t-1} \in X_{t-1}$. By definition of $\hat{\boldsymbol{x}}^*_t$ and Lipschitz continuity of M_t, the latter term is bounded from above by

$$\left(MB\max\left\{1, \|\xi_t\|^2\right\} + M\max\{1, \|\xi_t\|\} \cdot \mathbb{E}\left[[V_{t+1}]^x_{Lip}\left(\boldsymbol{\xi}^{t+1}\right)\middle|\, \boldsymbol{\xi}^t = \xi^t\right]\right) \cdot \|x_{t-1} - \hat{x}_{t-1}\|.$$

An analoguous estimate holds whenever the difference on the left side of (6) is positive. Hence, $[V_t]^x_{Lip}(\xi^t)$ is given by the term in parentheses, from which we conclude by recursion that we can put

$$[V_t]^x_{Lip}(\xi^t) \triangleq B\sum_{i=t}^{T} M^{i-t+1}\mathbb{E}\left[\max\{1, \|\boldsymbol{\xi}_i\|^2\} \cdot \prod_{k=t}^{i-1}\max\{1, \|\boldsymbol{\xi}_k\|\} \;\middle|\; \boldsymbol{\xi}^t = \xi^t\right].$$

Thus, the asserted bound for $[V_t]^x_{Lip}$ results from a straightforward estimate. ∎

Establishing continuity of $\xi^t \mapsto V_t(\xi^t, x_{t-1})$ is more subtle since, unlike the decision variable x_{t-1}, the observation ξ^t impacts not only the Lipschitz continuous time coupling constraints at time t, but also the expectations about future realizations of

$\boldsymbol{\xi}$. Therefore, one can hardly expect V_t to be Lipschitz continuous with respect to ξ^t without having that *the conditional distribution of* $(\boldsymbol{\xi}_s)_{s=t+1}^T$ *under* $\{\boldsymbol{\xi}^t = \xi^t\}$ *depends continuously on* ξ^t with respect to some appropriate measure of distance. It is illustrated by Example 2.6 of [8] that without such a continuous dependency stability of optimal values in terms of an L^p-distance does not hold in general. Thus, for establishing recursively the continuity of V_t, we need that continuity of V_{t+1} with respect to ξ^{t+1} is passed down to the mapping $\xi^t \mapsto \mathbb{E}\left[V_{t+1}(\boldsymbol{\xi}^{t+1}, x_t)\middle|\, \boldsymbol{\xi}^t = \xi^t\right]$. To this end, we introduce for $p \geq 1$ and a given Borel set $A^{t+1} \subset \Xi^{t+1}$ with $\mathbb{P}^{t+1}[A^{t+1}] = 1$ the class of functions

$$\mathbb{F}_p^{A^{t+1}}(\Xi^{t+1}) \triangleq \left\{ f : \Xi^{t+1} \to \mathbb{R} : (9) \text{ holds for } \xi^{t+1}, \hat{\xi}^{t+1} \in A^{t+1} \right\}$$

along with the Lipschitz condition

$$\left| f\left(\xi^{t+1}\right) - f\left(\hat{\xi}^{t+1}\right) \right| \leq \max\left\{1, \left\|\xi^{t+1}\right\|, \left\|\hat{\xi}^{t+1}\right\|\right\}^{p-1} \left\|\xi^{t+1} - \hat{\xi}^{t+1}\right\|. \tag{9}$$

We consider the following distance between probability measures P, Q on Ξ^{t+1}:

$$\zeta_p^{A^{t+1}}(P, Q) \triangleq \sup_{f \in \mathbb{F}_p^{A^{t+1}}(\Xi^{t+1})} \left| \int_{\Xi^{t+1}} f\left(\xi^{t+1}\right) P\left(d\xi^{t+1}\right) - \int_{\Xi^{t+1}} f\left(\xi^{t+1}\right) Q\left(d\xi^{t+1}\right) \right|.$$

Recall that, with the exception that we disregard the $\mathbb{P}^{t+1}$-null set $\Xi^{t+1} \setminus A^{t+1}$ within the definition of $\mathbb{F}_p^{A^{t+1}}$, the functional $\zeta_p^{A^{t+1}}$ corresponds to the *p-th order Fortet–Mourier distance*; see [15], [18]. Using this notation, the claimed continuity of the conditional distributions is specified by the following assumption.

Assumption 2.6. There exist constants $W, K > 0$, and $r \geq 0$, such that with

$$m_t \triangleq 1 + (T - t)(1 + r) \quad \text{for } t = 1, \ldots, T, \tag{10}$$

the following conditions are fulfilled.

(i) For every $t = 1, \ldots, T-1$, every Borel set $A^{t+1} \subset \Xi^{t+1}$ with $\mathbb{P}^{t+1}[A^{t+1}] = 1$, and $\mathbb{P}^t$-a.e. $\xi^t, \hat{\xi}^t \in \Xi^t$

$$\begin{aligned} &\zeta_{m_{(t+1)}+1}^{A^{t+1}}\left(\mathbb{P}\left[\boldsymbol{\xi}^{t+1} \in \cdot \,\middle|\, \boldsymbol{\xi}^t = \xi^t\right], \mathbb{P}\left[\boldsymbol{\xi}^{t+1} \in \cdot \,\middle|\, \boldsymbol{\xi}^t = \hat{\xi}^t\right]\right) \\ &\quad \leq K \max\left\{1, \|\xi^t\|, \|\hat{\xi}^t\|\right\}^{m_t - 1} \|\xi^t - \hat{\xi}^t\|. \end{aligned}$$

(ii) For every $t = 1, \ldots, T-1$ and $\mathbb{P}^t$-a.e. $\xi^t \in \Xi^t$

$$\mathbb{E}\left[\max\left\{1, \left\|\boldsymbol{\xi}^T\right\|\right\}^{1+T-t} \;\middle|\; \boldsymbol{\xi}^t = \xi^t \right] \leq W \cdot \max\left\{1, \left\|\xi^t\right\|\right\}^{m_t}.$$

Since the above assumption is crucial for the following continuity and stability results, it is discussed in the following remark.

Remark 2.7. Condition (i) is related to terms usually related to Markov processes, namely the *coefficient of ergodicity* and the *Feller property*; see, e.g., [4], [5], respectively. A similar assumption has been made by [1] to ensure stability of an optimal-stopping problem in a Markovian framework and by [12] for their study of consistency of tree approximations. It is also made implicitly by [11] by focusing on autoregressive processes. The more involved formulation of Assumption 2.6, allowing

for polynomially growing Lipschitz constants, is due to the fact that neither $\langle b_t(\xi_t), x_t\rangle$ nor M_{t+1} are uniformly Lipschitz continuous in ξ_t and x_t, unless both the support Ξ^T and the sets $X_t, t = 1, \dots, T$ are bounded. Indeed, under such a boundedness condition (i) may be significantly simplified; see Remark 2.8 below.

Lemma A.1 in the appendix provides conditions on $\boldsymbol{\xi}$, under which both (i) and (ii) hold true. In particular, this is the case if Ξ^T is finite. In the latter case one sees that ζ_p is the optimal value of a linear optimization problem that can be solved numerically to determine the constants K and r.

The following theorem shows that Assumption 2.6 indeed provides Lipschitz continuity of V_t with respect to ξ^t. We also refer to Proposition 2.7 of [11] which represents a corresponding continuity result.

THEOREM 1. *Suppose the Assumptions* 2.1*,* 2.3*, and* 2.6 *are fulfilled. For every* $t = 1, \dots, T$ *there is a constant* $C_t > 0$ *and a Borel set* $A^t \subset \Xi^t$ *with* $\mathbb{P}^t[A^t] = 1$ *such that*

$$\frac{1}{C_t \max\{1, \|x_{t-1}\|\}} V_t(\,\cdot\,, x_{t-1}) \in \mathbb{F}_{m_t+1}^{A^t}\left(\Xi^t\right)$$

holds true for every $x_{t-1} \in X_{t-1}$.

Proof. The assertion holds true for $V_{T+1} \equiv 0$; we show that it follows recursively for $t \leq T$. To this end, we proceed as in the proof of Proposition 2.5 and choose a measurable $\boldsymbol{x}_t^*$ with $\boldsymbol{x}_t^*(\xi^t) \in M_t(x_{t-1}, \xi_t)$ that fulfills (3) and $\|\boldsymbol{x}_t^*(\xi^t)\| \leq L \cdot \max\{1, \|x_{t-1}\|\} \cdot \max\{1, \|\xi^t\|\}$. Thus, we obtain

$$\begin{aligned}
&\left|V_t\left(\xi^t, x_{t-1}\right) - V_t\left(\hat{\xi}^t, x_{t-1}\right)\right| \\
&= \Bigg| \left\langle b_t\left(\xi_t\right), \boldsymbol{x}_t^*\left(\xi^t\right)\right\rangle + \mathbb{E}\left[V_{t+1}\left(\boldsymbol{\xi}^{t+1}, \boldsymbol{x}_t^*\left(\xi^t\right)\right)\middle|\, \boldsymbol{\xi}^t = \xi^t\right] \\
&\quad - \inf_{\hat{x}_t \in M_t(x_{t-1}, \hat{\xi}_t)} \left\{ \left\langle b_t(\hat{\xi}_t), \hat{x}_t \right\rangle + \mathbb{E}\left[V_{t+1}(\boldsymbol{\xi}^{t+1}, \hat{x}_t) \middle|\, \boldsymbol{\xi}^t = \hat{\xi}^t \right] \right\} \Bigg|, \qquad (11)
\end{aligned}$$

which holds, due to Assumption 2.3, for every $\xi^t, \hat{\xi}^t \in A^{t'}$ with $\mathbb{P}^t[A^{t'}] = 1$ for all $x_{t-1} \in X_{t-1}$. We consider the case when the term under the norm is negative and choose a measurable $\hat{\boldsymbol{x}}_t^*$ with

$$\hat{\boldsymbol{x}}_t^*(\hat{\xi}^t) \in \operatorname{argmin}_{z \in M_t(x_{t-1}, \hat{\xi}_t)} \|z - \boldsymbol{x}_t^*(\xi^t)\|$$

to obtain the following upper bound for (11):

$$\begin{aligned}
&- \left\langle b_t(\xi_t), \boldsymbol{x}_t^*\left(\xi^t\right)\right\rangle - \mathbb{E}\left[V_{t+1}\left(\boldsymbol{\xi}^{t+1}, \boldsymbol{x}_t^*\left(\xi^t\right)\right)\middle|\, \boldsymbol{\xi}^t = \xi^t\right] \\
&\quad + \left\langle b_t\left(\hat{\xi}_t\right), \hat{\boldsymbol{x}}_t^*\left(\hat{\xi}^t\right)\right\rangle + \mathbb{E}\left[V_{t+1}\left(\boldsymbol{\xi}^{t+1}, \hat{\boldsymbol{x}}_t^*\left(\hat{\xi}^t\right)\right)\middle|\, \boldsymbol{\xi}^t = \hat{\xi}^t\right]. \qquad (12)
\end{aligned}$$

Using linearity of b_t and Lipschitz continuity of M_t, the difference of the scalar product terms can be estimated by

$$\begin{aligned}
&\left|\left\langle b_t\left(\xi_t\right), \boldsymbol{x}_t^*\left(\xi^t\right)\right\rangle - \left\langle b_t\left(\hat{\xi}_t\right), \boldsymbol{x}_t^*\left(\xi^t\right)\right\rangle\right| \\
&\quad + \left|\left\langle b_t\left(\hat{\xi}_t\right), \boldsymbol{x}_t^*\left(\xi^t\right)\right\rangle - \left\langle b_t\left(\hat{\xi}_t\right), \hat{\boldsymbol{x}}_t^*\left(\hat{\xi}^t\right)\right\rangle\right| \\
&\leq B\|\xi_t - \hat{\xi}_t\| \cdot L \cdot \max\{1, \|x_{t-1}\|\} \max\left\{1, \|\xi^t\|\right\} \\
&\quad + B \max\left\{1, \left\|\hat{\xi}_t\right\|\right\} \cdot M \max\{1, \|x_{t-1}\|\} \|\xi_t - \hat{\xi}_t\| \\
&\leq B(L+M) \max\{1, \|x_{t-1}\|\} \max\left\{1, \|\xi^t\|, \|\hat{\xi}^t\|\right\} \|\xi_t - \hat{\xi}_t\|. \qquad (13)
\end{aligned}$$

The difference of the conditional expectations in (12) is bounded by

$$
\begin{aligned}
&|\mathbb{E}\left[V_{t+1}\left(\boldsymbol{\xi}^{t+1}, \boldsymbol{x}_t^*\left(\xi^t\right)\right) \mid \boldsymbol{\xi}^t = \xi^t\right] - \mathbb{E}\left[V_{t+1}\left(\boldsymbol{\xi}^{t+1}, \boldsymbol{x}_t^*\left(\xi^t\right)\right) \mid \boldsymbol{\xi}^t = \hat{\xi}^t\right]| \\
&\quad + |\mathbb{E}\left[V_{t+1}\left(\boldsymbol{\xi}^{t+1}, \boldsymbol{x}_t^*\left(\xi^t\right)\right) \mid \boldsymbol{\xi}^t = \hat{\xi}^t\right] - \mathbb{E}\left[V_{t+1}\left(\boldsymbol{\xi}^{t+1}, \hat{\boldsymbol{x}}_t^*\left(\hat{\xi}^t\right)\right) \mid \boldsymbol{\xi}^t = \hat{\xi}^t\right]| \\
&\quad \le C_{t+1} \max\left\{1, \|\boldsymbol{x}_t^*\left(\xi^t\right)\|\right\} \zeta_{m_{(t+1)}+1}^{A^{t+1}}\left(\mathbb{P}\left[\boldsymbol{\xi}^{t+1} \in \cdot \,\middle|\, \boldsymbol{\xi}^t = \xi^t\right], \mathbb{P}\left[\boldsymbol{\xi}^{t+1} \in \cdot \,\middle|\, \boldsymbol{\xi}^t = \hat{\xi}^t\right]\right) \\
&(14) \quad + \mathbb{E}\left[\left[V_{t+1}\right]_{Lip}^x\left(\boldsymbol{\xi}^{t+1}\right) \;\middle|\; \boldsymbol{\xi}^t = \hat{\xi}^t\right] M \max\left\{1, \|x_{t-1}\|\right\} \cdot \left\|\xi^t - \hat{\xi}^t\right\|,
\end{aligned}
$$

whereby the last inequality follows from the assertion for V_{t+1}, Proposition 2.5, and the Lipschitz continuity of M_t. This estimate holds true for every $\xi^t, \hat{\xi}^t \in A^{t''}$ for all $x_{t-1} \in X_{t-1}$, where $A^{t''}$ denotes the sets on which the assertions of Proposition 2.5 hold. Applying now condition (i) of Assumption 2.6 and the estimate (7), we see that the sum (14) does not exceed

$$
\begin{aligned}
&KC_{t+1} \max\left\{1, \left\|\boldsymbol{x}_t^*(\xi^t)\right\|\right\} \max\left\{1, \left\|\xi^t\right\|, \left\|\hat{\xi}^t\right\|\right\}^{m_t - 1} \left\|\xi^t - \hat{\xi}^t\right\| \\
&\quad + \bar{M} \cdot \mathbb{E}\left[\max\left\{1, \left\|\boldsymbol{\xi}^T\right\|\right\}^{1+T-t} \;\middle|\; \boldsymbol{\xi}^t = \xi^t\right] M \max\left\{1, \|x_{t-1}\|\right\} \cdot \left\|\xi^t - \hat{\xi}^t\right\|
\end{aligned}
$$

for every $\xi^t, \hat{\xi}^t \in A^{t'''}$. Thereby, $A^{t'''}$ denotes the set of $\mathbb{P}^t$-probability one on which Assumption 2.6 holds.

From condition (ii) of Assumption 2.6 and the boundedness of $\|\boldsymbol{x}_t^*\|$, we conclude that the latter sum is again bounded from above by

$$
(15) \qquad \left(KC_{t+1}L + \bar{M}WM\right) \max\left\{1, \left\|x_{t-1}\right\|\right\} \max\left\{1, \left\|\xi^t\right\|, \left\|\hat{\xi}^t\right\|\right\}^{m_t} \cdot \|\xi^t - \hat{\xi}^t\|.
$$

The upper bounds (13) and (15) remain valid if the term under the norm in (11) is positive. Piecing this all together, the assertion for V_t follows with $A^t \triangleq A^{t'} \cap A^{t''} \cap A^{t'''}$, and the Lipschitz constant C_t can be chosen by collecting the constants from (13) and (15); i.e.,

$$
C_t \triangleq B(M + L) + KC_{t+1}L + \bar{M}WM. \qquad \square
$$

In the following remark we sketch what can be simplified whenever the sets X_t and Ξ^T are bounded.

Remark 2.8. The constant m_t is chosen to be equal to the growth rate of the term $\max\{1, \|\xi^t\|, \|\hat{\xi}^t\|\}$ within an upper bound of (14). Assuming boundedness of the sets X_t for $t = 1, \ldots, T$ allows one to estimate the term $\max\{1, \|\boldsymbol{x}_t^*(\xi^t)\|\}$ in the first summand of (14) by some constant instead of estimating it by $\max\{1, \|\xi^t\|, \|\hat{\xi}^t\|\}$. Consequently, one can allow the growth rate of the ζ-terms in (14) and in Assumption 2.6 to increase from $m_t - 1$ to m_t.

If the set Ξ^T is bounded as well, then $[V_{t+1}]_{Lip}^x(\boldsymbol{\xi}^{t+1})$ is bounded by a constant, condition (i) of Assumption 2.6 may be simplified to

$$
\zeta_1^{A^{t+1}}\left(\mathbb{P}\left[\boldsymbol{\xi}^{t+1} \in \cdot \,\middle|\, \boldsymbol{\xi}^t = \xi^t\right], \mathbb{P}\left[\boldsymbol{\xi}^{t+1} \in \cdot \,\middle|\, \boldsymbol{\xi}^t = \hat{\xi}^t\right]\right) \le K\|\xi^t - \hat{\xi}^t\|,
$$

condition (ii) of Assumption 2.6 may be omitted, and the assertion of Theorem 1 can be written as $(1/C_t)\, V_t(\cdot, x_{t-1}) \in \mathbb{F}_1^{A^t}(\Xi^t)$, i.e., $(\xi^t, x_{t-1}) \mapsto V_t(\xi^t, x_{t-1})$ is uniformly Lipschitz continuous.

The optimality and boundedness conditions (3)–(5), as well as the continuity properties claimed in Assumption 2.6 and Theorem 1, hold on some Borel set $A \subset \Xi^T$ with $\mathbb{P}^T[A] = 1$. Since an approximation $\tilde{\boldsymbol{\xi}}$ may have its support in the set $\Xi^T \setminus A$, it is reasonable to modify the considered random variables on this $\mathbb{P}^T$-null set to appropriate versions which fulfill the claimed properties for *every* $\xi^T \in \Xi^T$. To this end, we recall that $\mathbb{P}^T[A] = 1$ and $\Xi^T = \operatorname{supp} \mathbb{P}^T$ imply that A is a dense subset of Ξ^T. For every $\hat{\xi}^T \in \Xi^T \setminus A$ we then consider a sequence $(\xi^T_{(n)})_{n\in\mathbb{N}} \subset A$ converging to $\hat{\xi}^T$ as n goes to infinity. The recourse function and the regular conditional distributions are modified in $\hat{\xi}^T$ by setting $V_t(\hat{\xi}^t, x_{t-1}) \triangleq \lim_{n\to\infty} V_t(\xi^t_{(n)}, x_{t-1})$ and

$$\mathbb{E}[g\left(\boldsymbol{\xi}^{t+1}\right) | \boldsymbol{\xi}^t = \hat{\xi}^t] \triangleq \lim_{n\to\infty} \mathbb{E}[g\left(\boldsymbol{\xi}^{t+1}\right) | \boldsymbol{\xi}^t = \xi^t_{(n)}]$$

for every Lipschitz continuous mapping g. A bounded optimal solution $\boldsymbol{x}^*$ can be appropriately modified in $\hat{\xi}^t$ by considering a subsequence $(\xi^T_{(n_k)})_{k\in\mathbb{N}}$ such that $\boldsymbol{x}^*_t(\xi^t_{(n_k)})$ converges toward some

$$z_t\left(\hat{\xi}^t\right) \in X_t$$

for $t = 1, \ldots, T$. Then we put $\boldsymbol{x}^*_t(\hat{\xi}^t) \triangleq z_t(\hat{\xi}^t)$, and we obtain that the above stated conditions and properties indeed hold for every $\xi^T \in \Xi^T$.

3. Approximations. Whenever an auxiliary process $\tilde{\boldsymbol{\xi}}$ is expected to approximate $\boldsymbol{\xi}$ with regard to the optimization problem (2), it is indispensable that $\tilde{\boldsymbol{\xi}}$ is nonanticipative with respect to $\boldsymbol{\xi}$. This is illustrated, for the sake of completeness, by Example A.3 in the appendix. Nonanticipativity is ensured in the following.

DEFINITION 3.1. *A stochastic process $\tilde{\boldsymbol{\xi}}$ on $(\Omega, \mathcal{F}, \mathbb{P})$ is called an* approximation of $\boldsymbol{\xi}$*, if there exist Borel-measurable mappings*

$$f_t : \Xi^t \to \Xi_t \text{ for } t = 1, \ldots, T,$$

fulfilling the following conditions:

(i) $\tilde{\boldsymbol{\xi}}_t = f_t(\boldsymbol{\xi}^t)$ *for* $t = 1, \ldots T$,
(ii) $f^T(\Xi^T) \subset \Xi^T$,
(iii) $f_1(\xi_1) = \xi_1$ *for every* $\xi_1 \in \Xi_1$, *and*
(iv) $f^T(\boldsymbol{\xi}) \in L^p(\Omega, \mathcal{F}, \mathbb{P})$ *for every* $p \in [1, \infty)$.

Thereby, $f^t(\xi^t)$ denotes the vector $(f_i(\xi^i))_{i=1}^t \in \mathbb{R}^{s\cdot t}$ for $t = 1, \ldots, T$. In the following, we use the notation f for the mapping $f^T(\cdot)$.

Remark 3.2. The nonanticipativity condition (i) is equivalent to $\sigma(\boldsymbol{\xi}^t)$-measurability of the random variable $\tilde{\boldsymbol{\xi}}_t$. Condition (ii) ensures that f^T maps onto realizations $\xi^T \in \Xi^T$ of the initial process and thus implies that the restriction of a decision $\boldsymbol{x}(\cdot) \in \mathcal{S}(\boldsymbol{\xi})$ on the set $f^T(\Xi^T)$ is a $\tilde{\boldsymbol{\xi}}$-feasible decision. The integrability condition (iv) is assumed again for the sake of simplicity. For the following results, it suffices that $f^T(\boldsymbol{\xi}) \in L^p(\Omega, \mathcal{F}, \mathbb{P})$ for a constant $p \in \mathbb{R}_+$ that is sufficiently large.

The following proposition relies heavily on the continuity of the recourse function stated in Theorem 1. It is shown that, although an optimal decision $\boldsymbol{x}^*(\cdot)$ is not continuous in general, its expected costs can be approximated by the decision $\boldsymbol{x}^*(f(\cdot))$ (which is piecewise constant whenever $\tilde{\boldsymbol{\xi}}$ has finite support). Although $\boldsymbol{x}^*(f(\cdot))$ may fail to fulfill the time-coupling constraints (1) with respect to $\boldsymbol{\xi}$, it can be used to construct a feasible decision. This will be carried out in the next section.

PROPOSITION 3.3. *Consider an optimal decision $\boldsymbol{x}^*$ which is bounded in the sense of* (5) *and an approximation mapping f according to Definition* 3.1.

Then there exists a constant $D > 0$ such that the following estimate holds:

$$|\varphi(\boldsymbol{x}^*(\boldsymbol{\xi}),\boldsymbol{\xi}) - \varphi(\boldsymbol{x}^*(f(\boldsymbol{\xi})),\boldsymbol{\xi})| \le D\mathbb{E}\left[\max\{1,\|\boldsymbol{\xi}\|,\|f(\boldsymbol{\xi})\|\}^{m_1}\cdot\|\boldsymbol{\xi}-f(\boldsymbol{\xi})\|\right], \tag{16}$$

where the constant m_1 is defined by (10).

Proof. Due to $f_1(\boldsymbol{\xi}_1) = \boldsymbol{\xi}_1$, we have to estimate

$$\left|\mathbb{E}\left[\sum_{t=2}^{T}\left\langle b_t(\boldsymbol{\xi}_t), \boldsymbol{x}_t^*\left(\boldsymbol{\xi}^t\right)\right\rangle\right] - \mathbb{E}\left[\sum_{t=2}^{T}\left\langle b_t(\boldsymbol{\xi}_t), \boldsymbol{x}_t^*\left(f^t\left(\boldsymbol{\xi}^t\right)\right)\right\rangle\right]\right|.$$

By the optimality of $\boldsymbol{x}^*$, the first expectation is equal to $\mathbb{E}\left[V_2(\boldsymbol{\xi}^2, \boldsymbol{x}_1^*)\right]$, and it follows from Theorem 1 and the boundedness of $\boldsymbol{x}_1^*$ (and $\boldsymbol{x}_0^* \triangleq 1$) that

$$\begin{aligned}\mathbb{E}\left[\left|V_2\left(\boldsymbol{\xi}^2,\boldsymbol{x}_1^*\right) - V_2\left(f^2\left(\boldsymbol{\xi}^2\right),\boldsymbol{x}_1^*\right)\right|\right] \\ \le LC_2\mathbb{E}\left[\max\left\{1,\|\boldsymbol{\xi}^2\|,\|f^2\left(\boldsymbol{\xi}^2\right)\|\right\}^{m_2}\|\boldsymbol{\xi}^2 - f^2\left(\boldsymbol{\xi}^2\right)\|\right]\end{aligned} \tag{17}$$

Thus, it remains to estimate

$$\left|\mathbb{E}\left[V_2\left(f^2\left(\boldsymbol{\xi}^2\right),\boldsymbol{x}_1^*\right) - \sum_{t=2}^{T}\left\langle b_t(\boldsymbol{\xi}_t), \boldsymbol{x}_t^*\left(f^t\left(\boldsymbol{\xi}^t\right)\right)\right\rangle\right]\right|. \tag{18}$$

To this end, we consider the following inequality

$$\begin{aligned}\Big|\mathbb{E}\Big[V_2\left(f^2\left(\boldsymbol{\xi}^2\right),\boldsymbol{x}_1^*\right) - \sum_{s=2}^{t-1}\langle b_s(\boldsymbol{\xi}_s), \boldsymbol{x}_s^*(f^s(\boldsymbol{\xi}^s))\rangle \\ - V_t\left(f^t\left(\boldsymbol{\xi}^t\right),\boldsymbol{x}_{t-1}^*\left(f^{t-1}\left(\boldsymbol{\xi}^{t-1}\right)\right)\right)\Big]\Big| \le D_t,\end{aligned} \tag{19}$$

whose left side coincides with (18) for $t = T+1$. It holds trivially for $t = 2$ with $D_2 = 0$, and we assume that it is also true for some $t \in \{2,\dots,T\}$ with a constant $D_t \ge 0$. To prove it recursively for $t+1$, we have to find an upper bound for

$$\begin{aligned}\Big|\mathbb{E}\big[V_t\left(f^t\left(\boldsymbol{\xi}^t\right),\boldsymbol{x}_{t-1}^*\left(f^{t-1}\left(\boldsymbol{\xi}^{t-1}\right)\right)\right) - \left\langle b_t\left(\boldsymbol{\xi}_t\right),\boldsymbol{x}_t^*\left(f^t\left(\xi^t\right)\right)\right\rangle \\ - V_{t+1}\left(f^{t+1}\left(\boldsymbol{\xi}^{t+1}\right),\boldsymbol{x}_t^*\left(f^t\left(\xi^t\right)\right)\right)\big]\Big|.\end{aligned} \tag{20}$$

To this end, we use again $\boldsymbol{x}^*$'s optimality to expand the first summand:

$$\begin{aligned}&\mathbb{E}\left[V_t\left(f^t\left(\boldsymbol{\xi}^t\right),\boldsymbol{x}_{t-1}^*\left(f^{t-1}\left(\boldsymbol{\xi}^{t-1}\right)\right)\right)\right] \\ &= \int\left\langle b_t\left(f_t\left(\xi^t\right)\right),\boldsymbol{x}_t^*\left(f^t\left(\xi^t\right)\right)\right\rangle + \mathbb{E}\left[V_{t+1}\left(\boldsymbol{\xi}^{t+1},\boldsymbol{x}_t^*\left(f^t\left(\xi^t\right)\right)\right)\middle|\,\boldsymbol{\xi}^t = f^t\left(\xi^t\right)\right]\mathbb{P}^t\left(d\xi^t\right).\end{aligned} \tag{21}$$

Now, to estimate (20), we have to replace $b_t(f_t(\xi^t))$ by $b_t(\xi^t)$. The Lipschitz continuity of $b_t(\cdot)$ implies

$$\left|\left\langle b_t\left(f_t\left(\xi^t\right)\right),\boldsymbol{x}_t^*\left(f^t\left(\xi^t\right)\right)\right\rangle - \left\langle b_t\left(\xi_t\right),\boldsymbol{x}_t^*\left(f^t\left(\xi^t\right)\right)\right\rangle\right| \le B\cdot\left\|\boldsymbol{x}_t^*\left(f^t\left(\xi^t\right)\right)\right\|\cdot\left\|\xi^t - f^t\left(\xi^t\right)\right\|.$$

To estimate the difference of the V_{t+1}-terms in (20) and (21), we add and subtract the term $\mathbb{E}\left[V_{t+1}(f^{t+1}(\boldsymbol{\xi}^{t+1}), \boldsymbol{x}_t^*(f^t(\xi^t)))\middle|\, \boldsymbol{\xi}^t = \xi^t\right]$ and use the triangle inequality to estimate

$$\begin{aligned}
&\left|\mathbb{E}\left[V_{t+1}\left(\boldsymbol{\xi}^{t+1}, \boldsymbol{x}_t^*\left(f^t\left(\xi^t\right)\right)\right)\middle|\, \boldsymbol{\xi}^t = f^t\left(\xi^t\right)\right]\right. \\
&\qquad \left. -\mathbb{E}\left[V_{t+1}\left(f^{t+1}\left(\boldsymbol{\xi}^{t+1}\right), \boldsymbol{x}_t^*\left(f^t\left(\xi^t\right)\right)\right)\middle|\, \boldsymbol{\xi}^t = \xi^t\right]\right| \\
&\quad \le \left|\mathbb{E}\left[V_{t+1}\left(\boldsymbol{\xi}^{t+1}, \boldsymbol{x}_t^*\left(f^t\left(\xi^t\right)\right)\right)\middle|\, \boldsymbol{\xi}^t = f^t\left(\xi^t\right)\right]\right. \\
&\qquad \left. -\mathbb{E}\left[V_{t+1}\left(\boldsymbol{\xi}^{t+1}, \boldsymbol{x}_t^*\left(f^t\left(\xi^t\right)\right)\right)\middle|\, \boldsymbol{\xi}^t = \xi^t\right]\right| \\
&\qquad + \left|\mathbb{E}\left[V_{t+1}\left(\boldsymbol{\xi}^{t+1}, \boldsymbol{x}_t^*\left(f^t\left(\xi^t\right)\right)\right)\middle|\, \boldsymbol{\xi}^t = \xi^t\right]\right. \\
&\qquad \left. -\mathbb{E}\left[V_{t+1}\left(f^{t+1}\left(\boldsymbol{\xi}^{t+1}\right), \boldsymbol{x}_t^*\left(f^t\left(\xi^t\right)\right)\right)\middle|\, \boldsymbol{\xi}^t = \xi^t\right]\right|.
\end{aligned}$$

By applying Theorem 1 and Assumption 2.6 we conclude that this term is bounded for $\mathbb{P}^t$-almost every ξ^t by

$$\begin{aligned}
&KC_{t+1}\max\left\{1, \left\|\boldsymbol{x}_t^*\left(f^t\left(\xi^t\right)\right)\right\|\right\}\max\left\{1, \left\|\xi^t\right\|, \left\|f^t\left(\xi^t\right)\right\|\right\}^{m_t-1}\left\|\xi^t - f^t\left(\xi^t\right)\right\| \\
&+ C_{t+1}\max\left\{1, \left\|\boldsymbol{x}_t^*\left(f^t\left(\xi^t\right)\right)\right\|\right\} \\
&\quad \cdot\mathbb{E}\left[\max\left\{1, \left\|\boldsymbol{\xi}^{t+1}\right\|, \left\|f^{t+1}\left(\boldsymbol{\xi}^{t+1}\right)\right\|\right\}^{m_{t+1}}\left\|\boldsymbol{\xi}^{t+1} - f^{t+1}\left(\boldsymbol{\xi}^{t+1}\right)\right\|\middle|\, \boldsymbol{\xi}^t = \xi^t\right] \\
&\le (K+1)C_{t+1}L^t\mathbb{E}\left[\max\left\{1, \left\|\boldsymbol{\xi}^{t+1}\right\|, \left\|f^{t+1}\left(\boldsymbol{\xi}^{t+1}\right)\right\|\right\}^{m_t+t-1}\left\|\boldsymbol{\xi}^{t+1}\right.\right. \\
&\qquad\qquad \left.\left. - f^{t+1}\left(\boldsymbol{\xi}^{t+1}\right)\right\|\middle|\, \boldsymbol{\xi}^t = \xi^t\right],
\end{aligned}$$

where the inequality follows from boundedness of $\boldsymbol{x}_t^*$ and the relation $m_{t+1} \le m_t - 1$. Integration with respect to $\mathbb{P}^t(d\xi^t)$ and combining these estimates with (21) entails that (20) does not exceed

(22)
$$(B+(K+1)C_{t+1})L^t\mathbb{E}\left[\max\left\{1, \left\|\boldsymbol{\xi}^{t+1}\right\|, \left\|f^{t+1}\left(\boldsymbol{\xi}^{t+1}\right)\right\|\right\}^{m_t+t-1}\cdot\left\|\boldsymbol{\xi}^{t+1} - f^{t+1}\left(\boldsymbol{\xi}^{t+1}\right)\right\|\right].$$

Hence, (19) holds for $t+1$ with D_{t+1} being equal to the sum of D_t and (22).

Due to the fact that both $m_t + t - 1$ and m_2 are smaller than m_1, the sum of (17) and (18) does not exceed

$$D\mathbb{E}\left[\max\{1, \|\boldsymbol{\xi}\|, \|f(\boldsymbol{\xi})\|\}^{m_1}\cdot\|\boldsymbol{\xi} - f(\boldsymbol{\xi})\|\right]$$

with $D \triangleq LC_2 + D_{T+1}$. This completes the proof. ∎

4. Calm decisions. One of the main difficulties in establishing the stability of the optimal value $v(\boldsymbol{\xi})$ with respect to perturbations of $\boldsymbol{\xi}$ is that optimal solutions do not depend continuously on the realization of $\boldsymbol{\xi}$, in general. Furthermore, the gap between $v(\boldsymbol{\xi})$ and the minimal expected costs which can be realized by, e.g., Lipschitz continuous solutions may be hard to estimate. In this section we shall introduce specific *calm* decisions and estimate the minimal expected costs realized by those decisions.

We consider an optimal decision $\boldsymbol{x}^*$ which is bounded in the sense of (5). The *calm modification of* $\boldsymbol{x}^*$ is defined by

$$\bar{\boldsymbol{x}}_1^* \triangleq \boldsymbol{x}_1^* \quad \text{and} \quad \bar{\boldsymbol{x}}_t^*\left(\xi^t\right) \in \operatorname{argmin}_{z \in M_t(\bar{\boldsymbol{x}}_{t-1}^*(\xi^{t-1}),\, \xi_t)}\left\|\boldsymbol{x}_t^*\left(f^t\left(\xi^t\right)\right) - z\right\| \text{ for } t = 2, \ldots, T,$$

where, again due to Theorem 14.37 of [17], the latter mappings can be chosen to be measurable. Observe that $\boldsymbol{x}_t^*$ and $\bar{\boldsymbol{x}}_t^*$ coincide on the set $f^t(\Xi^t)$, i.e.,

$$\bar{\boldsymbol{x}}_t^*\left(f^t\left(\xi^t\right)\right) = \boldsymbol{x}_t^*\left(f^t\left(\xi^t\right)\right) \text{ for every } \xi^t \in \Xi^t. \tag{23}$$

Due to the Lipschitz continuity of M_t, the local variability of $\bar{\boldsymbol{x}}_t^*(\cdot)$ in ξ^t can be estimated recursively, and $\bar{\boldsymbol{x}}_t^*(\cdot)$ is indeed *calm locally around* $f^t(\xi^t)$ *for every* $\xi^t \in \Xi^t$ in the following sense.

PROPOSITION 4.1. *For every* $t = 1, \ldots, T$ *and every* $\xi^T \in \Xi^T$ *we have*

$$\left\|\bar{\boldsymbol{x}}_t^*\left(\xi^t\right) - \bar{\boldsymbol{x}}_t^*\left(f^t\left(\xi^t\right)\right)\right\| \le (T-1)M^{T-1}\max\left\{1, \left\|\xi^T\right\|, \left\|f^T\left(\xi^T\right)\right\|\right\}^{T-1}\left\|\xi^T - f^T\left(\xi^T\right)\right\|. \tag{24}$$

Proof. For $t = 1$, the difference on the left side of (24) vanishes. For $t > 1$ we use the identity (23) and the definition of $\bar{\boldsymbol{x}}_t^*(\xi^t)$ to write

$$\left\|\bar{\boldsymbol{x}}_t^*\left(\xi^t\right) - \bar{\boldsymbol{x}}_t^*\left(f^t\left(\xi^t\right)\right)\right\| = \inf_{z \in M_t\left(\bar{\boldsymbol{x}}_{t-1}^*\left(\xi^{t-1}\right), \xi_t\right)}\left\|z - \bar{\boldsymbol{x}}_t^*\left(f^t\left(\xi^t\right)\right)\right\|. \tag{25}$$

Using the inclusion $\bar{\boldsymbol{x}}_t^*(f^t(\xi^t)) \in M_t(\bar{\boldsymbol{x}}_{t-1}^*(f^{t-1}(\xi^{t-1})), f_t(\xi^t))$, we obtain that the right-hand side of (25) is not greater than the Pompeiu–Hausdorff distance of the sets $M_t(\bar{\boldsymbol{x}}_{t-1}^*(\xi^{t-1}), \xi^t)$ and $M_t(\bar{\boldsymbol{x}}_{t-1}^*(f^{t-1}(\xi^{t-1})), f_t(\xi^t))$. We then apply the triangle inequality with respect to this metric and use the Lipschitz continuity of M_t to conclude that the right-hand side of (25) is bounded from above by

$$\begin{aligned} &M\max\left\{1, \left\|\xi^t\right\|\right\}\ \left\|\bar{\boldsymbol{x}}_{t-1}^*\left(\xi^{t-1}\right) - \bar{\boldsymbol{x}}_{t-1}^*\left(f^{t-1}\left(\xi^{t-1}\right)\right)\right\| \\ &+M\max\left\{1, \left\|\bar{\boldsymbol{x}}_{t-1}^*\left(f^{t-1}\left(\xi^{t-1}\right)\right)\right\|\right\}\ \left\|f^t\left(\xi^t\right) - \xi^t\right\|. \end{aligned}$$

By boundedness of $\boldsymbol{x}_{t-1}$, the latter sum does not exceed

$$\begin{aligned} &M\max\{1, \|\xi^t\|\}\ \left\|\bar{\boldsymbol{x}}_{t-1}^*\left(\xi^{t-1}\right) - \bar{\boldsymbol{x}}_{t-1}^*\left(f^{t-1}\left(\xi^{t-1}\right)\right)\right\| \\ &+ML\max\left\{1, \left\|f^{t-1}\left(\xi^{t-1}\right)\right\|\right\}^{t-1}\ \left\|f^t\left(\xi^t\right) - \xi^t\right\|. \end{aligned}$$

Recursively, we obtain that the left side of (24) is bounded by

$$L\sum_{i=2}^{t} M^{t+1-i}\max\left\{1, \left\|f^{i-1}\left(\xi^{i-1}\right)\right\|\right\}^{i-1}\ \max\left\{1, \left\|\xi^t\right\|\right\}^{t-i}\ \left\|\xi^i - f^i\left(\xi^i\right)\right\|.$$

The assertion follows by a straightforward estimate. ∎

By combining Propositions 3.3 and 4.1, one immediately concludes the following theorem, which shows that the difference of the expected costs generated by $\boldsymbol{x}^*$ and $\bar{\boldsymbol{x}}^*$ can be estimated in terms of the deviation between $\boldsymbol{\xi}$ and $f(\boldsymbol{\xi})$.

THEOREM 2. *Suppose Assumptions* 2.1, 2.3, *and* 2.6 *are fulfilled. Consider an optimal decision* $\boldsymbol{x}^*$ *which is bounded in the sense of* (5) *and its calm modification* $\bar{\boldsymbol{x}}^*$.

Then there exists a constant $C > 0$ *such that the following estimate holds:*

$$\left|\mathbb{E}\varphi(\boldsymbol{x}^*(\boldsymbol{\xi}), \boldsymbol{\xi}) - \mathbb{E}\varphi(\bar{\boldsymbol{x}}^*(\boldsymbol{\xi}), \boldsymbol{\xi})\right| \le C\,\mathbb{E}\left[\max\{1, \|\boldsymbol{\xi}\|, \|f(\boldsymbol{\xi})\|\}^{m_1} \cdot \|\boldsymbol{\xi} - f(\boldsymbol{\xi})\|\right],$$

where the constant m_1 *is defined by* (10).

Proof. To prove the assertion, we have to estimate the following term:

$$\left|\mathbb{E}\left[\sum_{t=2}^{T}\left\langle b_t(\boldsymbol{\xi}_t), \boldsymbol{x}_t^*\left(\boldsymbol{\xi}^t\right)\right\rangle - \sum_{t=2}^{T}\left\langle b_t(\boldsymbol{\xi}_t), \bar{\boldsymbol{x}}_t^*\left(\boldsymbol{\xi}^t\right)\right\rangle\right]\right|. \tag{26}$$

Recall that, by Proposition 3.3, $\boldsymbol{x}^*(\boldsymbol{\xi})$ and $\boldsymbol{x}^*(f(\boldsymbol{\xi}))$ produce comparable costs. On the other hand, the difference between $\boldsymbol{x}^*(f(\boldsymbol{\xi}))$ and $\bar{\boldsymbol{x}}^*(\boldsymbol{\xi})$ can be estimated due to the calmness of the latter decision. Thus, we add and subtract the term

$$\mathbb{E}\left[\sum_{t=2}^{T}\left\langle b_t(\boldsymbol{\xi}_t), \boldsymbol{x}_t^*\left(f^t\left(\boldsymbol{\xi}^t\right)\right)\right\rangle\right]$$

to the expectation within (26). Using then the triangle inequality as well as Proposition 3.3, we conclude that (26) is not greater than the sum of

$$\left|\mathbb{E}\left[\sum_{t=2}^{T}\left\langle b_t(\boldsymbol{\xi}_t), \boldsymbol{x}_t^*\left(f^t\left(\boldsymbol{\xi}^t\right)\right)\right\rangle - \sum_{t=2}^{T}\left\langle b_t(\boldsymbol{\xi}_t), \bar{\boldsymbol{x}}_t^*\left(\boldsymbol{\xi}^t\right)\right\rangle\right]\right| \tag{27}$$

and the right-hand side of (16). It thus remains to estimate (27). By applying identity (23) as well as the calmness property of $\bar{\boldsymbol{x}}^*$ from Proposition 4.1, we obtain the following upper bound:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=2}^{T} B\max\left\{1,\left\|\boldsymbol{\xi}^t\right\|\right\}(T-1)M^{T-1}\max\left\{1,\left\|\boldsymbol{\xi}^T\right\|,\left\|f^T\left(\boldsymbol{\xi}^T\right)\right\|\right\}^{T-1}\left\|\boldsymbol{\xi}^T - f^T\left(\boldsymbol{\xi}^T\right)\right\|\right]\\
&\leq (T-1)^2 BM^{T-1}\mathbb{E}\left[\max\left\{1,\left\|\boldsymbol{\xi}^T\right\|,\left\|f^T\left(\boldsymbol{\xi}^T\right)\right\|\right\}^{T}\left\|\boldsymbol{\xi}^T - f^T\left(\boldsymbol{\xi}^T\right)\right\|\right].
\end{aligned}$$

Finally, the sum of the latter term and the right-hand side of (16) is smaller than

$$C\,\mathbb{E}\left[\max\left\{1,\left\|\boldsymbol{\xi}^T\right\|,\left\|f^T\left(\boldsymbol{\xi}^T\right)\right\|\right\}^{m_1}\cdot\left\|\boldsymbol{\xi}^T - f^T\left(\boldsymbol{\xi}^T\right)\right\|\right],$$

with a constant $C \triangleq D + (T-1)^2\, B\, M^{T-1}$. □

5. Stability. In order to address the question of stability, we have to consider the following issue. Although we assume the existence of bounded optimal solutions to the initial problem (2), the perturbed problem may be unbounded. This is illustrated, for the sake of completeness, by Example A.4 in the Appendix. Heitsch, Römisch, and Strugarek [8] avoid such unfavorable cases by their Assumption (A2) of level-boundedness of the objective, *locally around* $\boldsymbol{\xi}$. We proceed by assuming that $\tilde{\boldsymbol{\xi}}$ fulfills Assumption 2.3 too; i.e., the perturbed problem $v(\tilde{\boldsymbol{\xi}})$ admits a bounded optimal solution. We now state our main result.

THEOREM 3. *Suppose Assumptions* 2.1*,* 2.3*, and* 2.6 *are fulfilled. Let* $\tilde{\boldsymbol{\xi}}$ *be an approximation of* $\boldsymbol{\xi}$ *according to Definition* 3.1*, which fulfills Assumption* 2.3*, too, and consider the constant* m_1 *defined by* (10)*.*

Then there exists a constant $\gamma > 0$*, such that*

$$\left|v(\boldsymbol{\xi}) - v\left(\tilde{\boldsymbol{\xi}}\right)\right| \leq \gamma\,\mathbb{E}\left[\max\left\{1,\left\|\boldsymbol{\xi}\right\|,\left\|\tilde{\boldsymbol{\xi}}\right\|\right\}^{m_1}\cdot\left\|\boldsymbol{\xi} - \tilde{\boldsymbol{\xi}}\right\|\right]$$

holds.

Proof. We denote the approximation mapping corresponding to $\tilde{\boldsymbol{\xi}}$ by f and consider a bounded optimal decision $\boldsymbol{x}^* \in \mathcal{S}(\boldsymbol{\xi})$ and the corresponding calm modification $\bar{\boldsymbol{x}}^*$ from section 4.

Applying Theorem 2 yields the following inequality:

$$\begin{aligned}
v\left(\tilde{\boldsymbol{\xi}}\right) - v(\boldsymbol{\xi}) &= v\left(\tilde{\boldsymbol{\xi}}\right) - \mathbb{E}\varphi(\boldsymbol{x}^*(\boldsymbol{\xi}), \boldsymbol{\xi})\\
&\leq v\left(\tilde{\boldsymbol{\xi}}\right) - \mathbb{E}\varphi(\bar{\boldsymbol{x}}^*(\boldsymbol{\xi}), \boldsymbol{\xi}) + C\mathbb{E}\left[\max\left\{1,\left\|\boldsymbol{\xi}\right\|,\left\|\tilde{\boldsymbol{\xi}}\right\|\right\}^{m_1}\cdot\left\|\boldsymbol{\xi} - \tilde{\boldsymbol{\xi}}\right\|\right].
\end{aligned} \tag{28}$$

Since the restriction of $\bar{\boldsymbol{x}}^*$ on $f(\Xi)$ is contained in $\mathcal{S}(\tilde{\boldsymbol{\xi}})$, we can write

$$\begin{aligned} v\left(\tilde{\boldsymbol{\xi}}\right) - \mathbb{E}\varphi(\bar{\boldsymbol{x}}^*(\boldsymbol{\xi}), \boldsymbol{\xi}) \\ &\leq \mathbb{E}\varphi\left(\bar{\boldsymbol{x}}^*\left(\tilde{\boldsymbol{\xi}}\right), \tilde{\boldsymbol{\xi}}\right) - \mathbb{E}\varphi(\bar{\boldsymbol{x}}^*(\boldsymbol{\xi}), \boldsymbol{\xi}) \\ &= \sum_{t=2}^{T} \mathbb{E}\left[\left\langle b_t\left(\tilde{\boldsymbol{\xi}}_t\right) - b_t(\boldsymbol{\xi}_t), \bar{\boldsymbol{x}}_t^*\left(\tilde{\boldsymbol{\xi}}^t\right)\right\rangle + \left\langle b_t(\boldsymbol{\xi_t}), \bar{\boldsymbol{x}}_t^*\left(\tilde{\boldsymbol{\xi}}^t\right) - \bar{\boldsymbol{x}}_t^*\left(\boldsymbol{\xi}^t\right)\right\rangle\right] \\ &\leq B\sum_{t=2}^{T} \mathbb{E}\left[\|\tilde{\boldsymbol{\xi}}_t - \boldsymbol{\xi}_t\| \left\|\bar{\boldsymbol{x}}_t^*\left(\tilde{\boldsymbol{\xi}}^t\right)\right\| + \max\left\{1, \|\boldsymbol{\xi}_t\|\right\} \left\|\bar{\boldsymbol{x}}_t^*\left(\tilde{\boldsymbol{\xi}}^t\right) - \bar{\boldsymbol{x}}_t^*\left(\boldsymbol{\xi}^t\right)\right\|\right]. \end{aligned} \tag{29}$$

Due to the fact that $\bar{\boldsymbol{x}}^*$ and $\boldsymbol{x}^*$ coincide on the set $f(\Xi^T)$, see (23), we obtain that $\bar{\boldsymbol{x}}^*$ fulfills the boundedness condition (5) on $f(\Xi^T)$. Using this boundedness as well as the calmness of $\bar{\boldsymbol{x}}^*$, each of the $T-1$ summands in (29) can be estimated. Thus, (29) is bounded from above by

$$H\mathbb{E}\left[\max\left\{1, \|\boldsymbol{\xi}\|, \|\tilde{\boldsymbol{\xi}}\|\right\}^T \cdot \|\boldsymbol{\xi} - \tilde{\boldsymbol{\xi}}\|\right], \tag{30}$$

with an appropriate constant $H > 0$, and we can use the relation $T \leq m_1$ to obtain

$$v\left(\tilde{\boldsymbol{\xi}}\right) - v(\boldsymbol{\xi}) \leq (C+H)\mathbb{E}\left[\max\left\{1, \|\boldsymbol{\xi}\|, \|\tilde{\boldsymbol{\xi}}\|\right\}^{m_1} \cdot \|\boldsymbol{\xi} - \tilde{\boldsymbol{\xi}}\|\right].$$

Now, we consider a bounded optimal decision $\tilde{\boldsymbol{x}}^*$ of $v(\tilde{\boldsymbol{\xi}})$. Following exactly the construction of section 4, we obtain a decision $\bar{\tilde{\boldsymbol{x}}}^* \in \mathcal{S}(\boldsymbol{\xi})$ that is calm in the sense of Proposition 4.1 and whose restriction on $f(\Xi^T)$ is optimal for $v(\tilde{\boldsymbol{\xi}})$. As in (29), it follows that

$$\begin{aligned} v(\boldsymbol{\xi}) - v\left(\tilde{\boldsymbol{\xi}}\right) &\leq \mathbb{E}\varphi\left(\bar{\tilde{\boldsymbol{x}}}^*(\boldsymbol{\xi}), \boldsymbol{\xi}\right) - \mathbb{E}\varphi\left(\bar{\tilde{\boldsymbol{x}}}^*\left(\tilde{\boldsymbol{\xi}}\right), \tilde{\boldsymbol{\xi}}\right) \\ &\leq B\sum_{t=2}^{T} \mathbb{E}\left[\max\left\{1, \|\boldsymbol{\xi}_t\|\right\} \left\|\bar{\tilde{\boldsymbol{x}}}_t^*\left(\tilde{\boldsymbol{\xi}}^t\right) - \bar{\tilde{\boldsymbol{x}}}_t^*\left(\boldsymbol{\xi}^t\right)\right\| + \|\tilde{\boldsymbol{\xi}}_t - \boldsymbol{\xi_t}\| \left\|\bar{\tilde{\boldsymbol{x}}}_t^*\left(\tilde{\boldsymbol{\xi}}^t\right)\right\|\right] \\ &\leq H\mathbb{E}\left[\max\left\{1, \|\boldsymbol{\xi}^T\|, \left\|\tilde{\boldsymbol{\xi}}^T\right\|\right\}^T \cdot \left\|\boldsymbol{\xi}^T - \tilde{\boldsymbol{\xi}}^T\right\|\right]. \end{aligned}$$

Applying again $T \leq m_1$ and setting $\gamma \triangleq C + H$ completes the proof. □

Remark 5.1. Since the purpose of this paper is to establish a stability result rather than the development of new approximation techniques, we restrict ourselves to refer to existing approaches based on conditional or unconditional clustering, which can be used to control the upper bound on $|v(\boldsymbol{\xi}) - v(\tilde{\boldsymbol{\xi}})|$ from Theorem 3. We mention here the recent approaches of [7], [1], [9], and [14].

Other approximation techniques, e.g., those proposed by [10], [12], are not based on projections of the initial process $\boldsymbol{\xi}$. Consequently, neither the joint distribution of $\boldsymbol{\xi}$ and $\tilde{\boldsymbol{\xi}}$ nor the underlying probability spaces are necessarily specified. However, under some weak regularity conditions our results may be applied to such cases as well. Indeed, one may choose the sample space $(\Xi^T \times \Xi^T, \mathcal{B}(\Xi^T) \otimes \mathcal{B}(\Xi^T), \mathbb{P}^T \otimes \tilde{\mathbb{P}}^T)$ as underlying probability space for both processes and define a nonanticipative coupling mapping $f : \Xi^T \to \Xi^T$ by, e.g., using successive projections. Thereby, $\mathcal{B}(\Xi^T)$ denotes the Borel sets of Ξ^T, and $\tilde{\mathbb{P}}^T$ denotes the distribution of the approximating process $\tilde{\boldsymbol{\xi}}$. Determining such a mapping f is closely related to *L^p-minimal metrics* and to *mass transportation problems*, see also Remark 2.3 of [8].

Appendix. The following lemma provides conditions under which the conditions of Assumption 2.6 hold true.

LEMMA A.1. *Assume the dynamics of the process* $\boldsymbol{\xi}$ *are given by the following scheme:*

$$\boldsymbol{\xi}_{t+1} = g_t\left(\boldsymbol{\xi}^t, \boldsymbol{\varepsilon}_{t+1}\right), \tag{31}$$

where $\boldsymbol{\varepsilon}_{t+1}$ *is a* $\mathbb{R}^n$*-valued random variable that is independent of* $\boldsymbol{\xi}^t$*, and* g_t *are measurable mappings from* $\mathbb{R}^{s\cdot t} \times \mathbb{R}^n$ *to* $\mathbb{R}^s$ *which satisfy the following Lipschitz and linear growth conditions:*

(i) $\|g_t(\xi^t, \varepsilon) - g_t(\hat{\xi}^t, \varepsilon)\| \le \max\{1, \|\xi^t\|, \|\hat{\xi}^t\|\}^r \|\xi^t - \hat{\xi}^t\| \, h(\|\varepsilon\|),$,

(ii) $\|g_t(\xi^t, \varepsilon)\| \le \max\{1, \|\xi^t\|\} \, k(\|\varepsilon\|),$

for all $\varepsilon \in \mathbb{R}^n$ *and* $\xi^t, \hat{\xi}^t \in \mathbb{R}^{s\cdot t}$*, some constant* $r \ge 1$ *and Borel-measurable mappings* $h, k \ge 1$*, such that* $h(\|\boldsymbol{\varepsilon}_{t+1}\|)$ *and* $k(\|\boldsymbol{\varepsilon}_{t+1}\|)$ *are in* L^p *for every* $p \in [1, +\infty)$*. Then* $\boldsymbol{\xi}$ *fulfills both conditions of Assumption* 2.6 *with the constants*

$$K \triangleq \mathbb{E}\left[k(\|\boldsymbol{\varepsilon}_{t+1}\|)^{m_1} h(\|\boldsymbol{\varepsilon}_{t+1}\|)\right] \quad \textit{and} \quad W \triangleq \mathbb{E}\left[\prod_{i=t+1}^{T} k(\|\boldsymbol{\varepsilon}_i\|)^{1+T-t}\right].$$

Proof. Consider $f \in \mathbb{F}_{1+m_{t+1}}(\Xi^{t+1})$. Then we obtain

$$\begin{aligned}
&\left|\mathbb{E}\left[f\left(\boldsymbol{\xi}^{t+1}\right) \middle| \boldsymbol{\xi}^t = \xi^t\right] - \mathbb{E}\left[f\left(\boldsymbol{\xi}^{t+1}\right) \middle| \boldsymbol{\xi}^t = \hat{\xi}^t\right]\right| \\
&= \left|\mathbb{E}\left[f\left(g_t\left(\xi^t, \boldsymbol{\varepsilon}_{t+1}\right)\right)\right] - \mathbb{E}\left[f\big(g_t\big(\hat{\xi}^t, \boldsymbol{\varepsilon}_{t+1}\big)\big)\right]\right| \\
&\le \mathbb{E}\left[\max\left\{1, \left\|g_t\left(\xi^t, \boldsymbol{\varepsilon}_{t+1}\right)\right\|, \left\|g_t\big(\hat{\xi}^t, \boldsymbol{\varepsilon}_{t+1}\big)\right\|\right\}^{m_{t+1}} \left\|g_t\big(\xi^t, \boldsymbol{\varepsilon}_{t+1}\big) - g_t\big(\hat{\xi}^t, \boldsymbol{\varepsilon}_{t+1}\big)\right\|\right] \\
&\le \mathbb{E}\left[\max\left\{1, \left\|g_t\left(\xi^t, \boldsymbol{\varepsilon}_{t+1}\right)\right\|, \left\|g_t\big(\hat{\xi}^t, \boldsymbol{\varepsilon}_{t+1}\big)\right\|\right\}^{m_{t+1}} h(\|\boldsymbol{\varepsilon}_{t+1}\|)\right] \\
&\quad \cdot \max\left\{1, \|\xi^t\|, \|\hat{\xi}^t\|\right\}^r \|\xi^t - \hat{\xi}^t\| \\
&\le \mathbb{E}\left[\max\left\{1, \|\xi^t\|, \|\hat{\xi}^t\|\right\}^{m_{t+1}} k(\|\boldsymbol{\varepsilon}_{t+1}\|)^{m_{t+1}} h(\|\boldsymbol{\varepsilon}_{t+1}\|)\right] \max\left\{1, \|\xi^t\|, \|\hat{\xi}^t\|\right\}^r \|\xi^t - \hat{\xi}^t\| \\
&= \mathbb{E}\left[k(\|\boldsymbol{\varepsilon}_{t+1}\|)^{m_{t+1}} h(\|\boldsymbol{\varepsilon}_{t+1}\|)\right] \cdot \max\left\{1, \|\xi^t\|, \|\hat{\xi}^t\|\right\}^{r+m_{t+1}} \|\xi^t - \hat{\xi}^t\|.
\end{aligned}$$

Due to the identity $r + m_{t+1} = m_t - 1$, this entails condition (i) of Assumption 2.6. The asserted form of K follows from $m_1 \ge m_t$ for $t = 1, \ldots, T$.

Furthermore, we apply (31) recursively to obtain the following estimate:

$$\left\|\boldsymbol{\xi}^T\right\| \le \max\left\{1, \left\|\boldsymbol{\xi}^t\right\|\right\} \prod_{i=t+1}^{T} k(\|\boldsymbol{\varepsilon}_i\|),$$

Raising both sides to the power of $1 + (T - t)$ and taking conditional expectations, $\mathbb{E}[\ \cdot\ |\boldsymbol{\xi}^t = \xi^t]$ verifies condition (ii) of Assumption 2.6. □

The conditions of Lemma A.1 are fulfilled, e.g., by a variety of time-series models. We provide the following simple example.

Example A.2. Let $\boldsymbol{\xi}$ be a GARCH process defined by the following difference equations:

$$\begin{aligned}
&\boldsymbol{\xi}_t = (\boldsymbol{w}_t, \boldsymbol{v}_t, \boldsymbol{\varepsilon}_t) \text{ with} \\
&\boldsymbol{v}_{t+1} \triangleq \sum_{i=0}^{s} (\beta_i v_{t-i} + \gamma_i \boldsymbol{\varepsilon}_{t-i}) \quad \text{and} \quad \boldsymbol{w}_{t+1} \triangleq \sum_{i=0}^{s} \alpha_i \boldsymbol{w}_{t-i} + \boldsymbol{v}_{t+1} \cdot \boldsymbol{\varepsilon}_{t+1}
\end{aligned}$$

for certain parameters $\alpha_i, \beta_i, \gamma_i \in \mathbb{R}$. Thereby, $\boldsymbol{v}$ represents the stochastic volatility process of $\boldsymbol{w}$ and $(\boldsymbol{\varepsilon}_t)_{t\geq 0}$ is a sequence of i.i.d. random variables, following a standard normal distribution. It is easy to see that $\boldsymbol{\xi}$ fulfills the conditions of Lemma A.1 with $r = 1$ and $h(\cdot), k(\cdot)$ being affine functions.

The following example shows that nonanticipativity with respect to the initial process is indispensable for an approximating process.

Example A.3. Consider $T = 3$ and the process $\boldsymbol{\xi}$ that is given by $\boldsymbol{\xi}_1 \equiv 0$ and the two independent random variables $\boldsymbol{\xi}_2$ and $\boldsymbol{\xi}_3$, both uniformly distributed on $[0, 1]$. For $n \in \mathbb{N}$ and $0 < \varepsilon < 1$ we introduce the grids $A^{(n)} \triangleq \{\frac{i}{n} : i = 1, \dots, n\}$ and the associated (right-continuous) projections $\pi_{A^{(n)}} : [0, 1] \to A^{(n)}$, defined by

$$\pi_{A^{(n)}}(z) \triangleq \max\left\{\frac{i}{n} \in A^{(n)} : \left|z - \frac{i}{n}\right| \leq \left|z - \frac{j}{n}\right| \text{ for all } \frac{j}{n} \in A^{(n)}\right\}.$$

Furthermore, we define processes $\boldsymbol{\xi}^{(n)}, n \in \mathbb{N}$, given by $\boldsymbol{\xi}_1^{(n)} \equiv 0$, $\boldsymbol{\xi}_3^{(n)} \triangleq \pi_{A^{(n)}}\boldsymbol{\xi}_3$, and

$$\boldsymbol{\xi}_2^{(n)} \triangleq \begin{cases} \pi_{A^{(n)}}\boldsymbol{\xi}_2 & \text{if } \boldsymbol{\xi}_3 \leq 1/2, \\ (\pi_{A^{(n)}}\boldsymbol{\xi}_2) + \frac{\varepsilon}{n} & \text{if } \boldsymbol{\xi}_3 > 1/2. \end{cases}$$

We remark that by observing $\boldsymbol{\xi}_2^{(n)}$ one knows whether $\boldsymbol{\xi}_3 > 1/2$ or not. Furthermore, the sequence $\boldsymbol{\xi}^{(n)}$ can be seen as an approximation of $\boldsymbol{\xi}$, since $\mathbb{E}[\|\boldsymbol{\xi} - \boldsymbol{\xi}^{(n)}\|] \leq \frac{1+2\varepsilon}{2n}$ holds.

We consider the following optimization problem

$$v(\boldsymbol{\xi}) \triangleq \min\left\{\mathbb{E}\left[\boldsymbol{x}_2 \cdot \boldsymbol{\xi}_2 + \boldsymbol{x}_3 \cdot \boldsymbol{\xi}_3\right] : \ \boldsymbol{x}_t \geq 0,\, \boldsymbol{x}_t \in \sigma(\boldsymbol{\xi}^t),\, t = 2, 3,\, \boldsymbol{x}_2 + \boldsymbol{x}_3 = 1 \text{ a.s.}\right\},$$

which is solved by $\boldsymbol{x}_2^* = \mathbb{1}_{\{\boldsymbol{\xi}_2 \leq 1/2\}}$ and $\boldsymbol{x}_3^* = 1 - \boldsymbol{x}_2^*$ with optimal value $v(\boldsymbol{\xi}) = 12/32$. When replacing $\boldsymbol{\xi}$ by $\boldsymbol{\xi}^{(n)}$, we use the decisions

$$\boldsymbol{x}_2^{(n)} = \mathbb{1}_{\{\boldsymbol{\xi}_2^{(n)} \leq 1/4\}} + \mathbb{1}_{\{\boldsymbol{\xi}_2^{(n)} \in\,]1/4,\, 3/4[\, \setminus A^{(n)}\}} \quad \text{and} \quad \boldsymbol{x}_3^{(n)} = 1 - \boldsymbol{x}_2^{(n)}$$

to obtain $\limsup_{n\to\infty} v(\boldsymbol{\xi}^{(n)}) \leq 11/32$. Obviously, convergence of $v(\boldsymbol{\xi}^{(n)})$ to $v(\boldsymbol{\xi})$ does not hold since the processes $\boldsymbol{\xi}^{(n)}$ do not fulfill the nonanticipativity condition (i) of Definition 3.1. With regard to the results of [8] and [12], convergence fails since the conditional distributions

$$\mathbb{P}[\boldsymbol{\xi}_3^{(n)} \in \cdot\, | \boldsymbol{\xi}_2^{(n)} = z]$$

do not converge toward $\mathbb{P}[\boldsymbol{\xi}_3 \in \cdot\, | \boldsymbol{\xi}_2 = z]$, and the filtration distance between $\boldsymbol{\xi}^{(n)}$ and $\boldsymbol{\xi}$ does not converge toward 0, respectively.

The following example shows that the perturbed problem $v(\tilde{\boldsymbol{\xi}})$ may be unbounded, even if the initial problem $v(\boldsymbol{\xi})$ admits a bounded optimal solution.

Example A.4. Consider some $\varepsilon \in (0, \frac{1}{4})$, $T = 2$, and the *stock prices* $\boldsymbol{\xi}_1 \equiv \frac{1}{2} + \varepsilon$ and $\boldsymbol{\xi}_2$, where the latter is uniformly distributed on $[0, 1]$. The optimal investment problem $v(\boldsymbol{\xi}) = \min_{x\geq 0}\, x \cdot \boldsymbol{\xi}_1 - \mathbb{E}[x \cdot \boldsymbol{\xi}_2] = \min_{x\geq 0} x \cdot \varepsilon$ is solved by $x = 0$. The process $\tilde{\boldsymbol{\xi}}$, defined by $\tilde{\boldsymbol{\xi}}_1 \triangleq \boldsymbol{\xi}_1$ and

$$\tilde{\boldsymbol{\xi}}_2 \triangleq \begin{cases} 1 & \text{if } \boldsymbol{\xi}_2 \geq \frac{1}{2} - 2\,\varepsilon, \\ 0 & \text{else} \end{cases},$$

is an approximation of $\boldsymbol{\xi}$ according to Definition 3.1. However, we see that $\mathbb{E}[\tilde{\boldsymbol{\xi}}_2] = \frac{1}{2} + 2\,\varepsilon$ and, consequently, $v(\tilde{\boldsymbol{\xi}}) = \min_{x\geq 0} -x \cdot \varepsilon = -\infty$.

Acknowledgments. The author is grateful to Professor Werner Römisch for his help and encouragement, to Stefan Vigerske for being perpetually willing for helpful discussions, and to Thomas Surowiec for proofreading the manuscript. The author also would like to thank the anonymous referees whose careful comments helped to clarify the presentation.

REFERENCES

[1] V. Bally, G. Pagès, and J. Printems, *A quantization tree method for pricing and hedging multidimensional American options*, Math. Finance, 15 (2005), pp. 119–168.
[2] K. Barty, *Contributions à la discrétisation des contraintes de mesurabilité pour les problèmes d'optimisation stochastique*, Ph.D. thesis, École Nationale des Ponts et Chaussées, Paris, 2004.
[3] J. Birge and F. Louveaux, *Introduction to Stochastic Programming*, Springer Series in Operations Research, Springer-Verlag, Berlin, 1997.
[4] R. Dobrushin, *Central limit theorem for non-stationary Markov chains* I., Teor. Veroyatnost. i Primenen., 1 (1956), pp. 72–89.
[5] E. Dynkin, *Markov Processes*, Springer, Berlin, 1965.
[6] I. Evstigneev, *Measurable selection and dynamic programming*, Math. Oper. Res., 1 (1976), pp. 267–272.
[7] H. Heitsch and W. Römisch, *Scenario tree modeling for multistage stochastic programs*, Math. Program., 2008, to appear.
[8] H. Heitsch, W. Römisch, and C. Strugarek, *Stability of multistage stochastic programs*, SIAM J. Optim., 17 (2006), pp. 511–525.
[9] R. Hochreiter and G. Pflug, *Financial scenario generation for stochastic multi-stage decision processes as facility location problems*, Ann. Oper. Res., 152 (2007), pp. 257–272.
[10] K. Høyland and S. Wallace, *Generating scenario trees for multistage decision problems*, Management Science, 47 (2001), pp. 295–307.
[11] D. Kuhn, *Generalized bounds for convex multistage stochastic programs*, Lecture Notes in Econ. Math. Systems 548, Springer, Berlin, 2005.
[12] R. Mirkov and G. Pflug, *Tree approximations of dynamic stochastic programs*, SIAM J. Optim., 18 (2007), pp. 1082–1105.
[13] T. Pennanen, *Epi-convergent discretizations of multistage stochastic programs*, Math. Oper. Res., 30 (2005), pp. 245–256.
[14] T. Pennanen, *Epi-convergent discretizations of multistage stochastic programs via integration quadratures*, Math. Program., 116 (2009), pp. 461–479.
[15] S. Rachev, *Probability Metrics and the Stability of Stochastic Models*, Wiley, Chichester, 1991.
[16] R. Rockafellar and R. J.-B. Wets, *Continuous Versus Measurable Recourse in N-Stage Stochastic Programming*, J. Math. Anal. Appl., 48 (1974), pp. 836–859.
[17] R. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[18] W. Römisch, *Stability of Stochastic Programming Problems*, Chap. 8 (2003), pp. 483–554, Vol. 10 of [19].
[19] A. Ruszczyński and A. Shapiro, eds., *Stochastic programming*, Handbooks Oper. Res. Management Sci. 10, Elsevier, Amsterdam, 2003.
[20] A. Shapiro, *Inference of statistical bounds for multistage stochastic programming problems*, Math. Methods Oper. Res., 58 (2003), pp. 57–68.

SIAM J. OPTIM.
Vol. 19, No. 2, pp. 969–1001

CONVERGENCE OF A CLASS OF SEMI-IMPLICIT TIME-STEPPING SCHEMES FOR NONSMOOTH RIGID MULTIBODY DYNAMICS*

BOGDAN I. GAVREA†, MIHAI ANITESCU‡, AND FLORIAN A. POTRA§

Abstract. In this work we present a framework for the convergence analysis in a measure differential inclusion sense of a class of time-stepping schemes for multibody dynamics with contacts, joints, and friction. This class of methods solves one linear complementarity problem per step and contains the semi-implicit Euler method, as well as trapezoidal-like methods for which second-order convergence was recently proved under certain conditions. By using the concept of a reduced friction cone, the analysis includes, for the first time, a convergence result for the case that includes joints. An unexpected intermediary result is that we are able to define a discrete velocity function of bounded variation, although the natural discrete velocity function produced by our algorithm may have unbounded variation.

Key words. rigid body, contact dynamics, friction, measure differential inclusion, complementarity problems

AMS subject classifications. 65K10, 90C33

DOI. 10.1137/060675745

1. Introduction.

The dynamic rigid multibody contact problem is concerned with predicting the motion of several rigid bodies in contact, and it is one of the fundamental paradigms in modern computational science. It appears in the description of fuel motion in the pebble bed reactor [21], in the compaction of nanopowders [27, 9], and in the study of biological membranes [41, 26, 52, 22]. Such simulations are also used extensively in structural engineering [16], pedestrian evacuation dynamics [24], granular matter [40], robotics simulation and design [17], and virtual reality [1].

The problem of multibody rigid systems involving contact and friction is a *differential complementarity problem* (DCP). The DCP is part of a broader class of problems known as *differential variational inequalities*, which were recently intro-

*Received by the editors November 22, 2006; accepted for publication (in revised form) May 27, 2008; published electronically October 16, 2008.

http://www.siam.org/journals/siopt/19-2/67574.html

†Technical University of Cluj-Napoca, Faculty of Automation and Computer Science, 26-28 Gh. Baritiu St., RO-400027 Cluj-Napoca, Romania, (Bogdan.GAVREA@math.utcluj.ro). Parts of the work were completed while the author was with the GRASP Laboratory, Department of Mechanical Engineering, University of Pennsylvania and with the Department of Mathematics and Statistics, University of Maryland-Baltimore County. This author was supported at the University of Maryland-Baltimore County by NSF grant DMS-0139701 and at the University of Pennsylvania by NSF grant IIS-0413138 (principal investigator: Vijay Kumar).

‡Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439 (anitescu@mcs.anl.gov). This author was supported by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy, under Contract DE-AC02-06CH11357.

§Department of Mathematics and Statistics, University of Maryland-Baltimore County, Baltimore, MD 21250 (potra@math.umbc.edu). This author was supported at the University of Maryland-Baltimore County by NSF grant DMS-0139701.

duced in [36, 37]. Approaches used in the past for the numerical approximation of rigid multibody dynamics with contact and friction include piecewise DAE approaches [23], nonsmooth contact dynamics method [31, 25, 40, 11, 10], acceleration-force linear complementarity problem (LCP) approaches [8, 38, 50, 18, 19], penalty approaches [15, 43, 44, 35], and velocity-impulse LCP-based time-stepping methods [45, 48, 4, 3].

The DCP gives rise to event-driven time-stepping schemes that are solved in an acceleration-force framework. These types of schemes will detect the discontinuity events; and, if these events are isolated, they will treat the dynamics as differential algebraic equations (DAEs) on each smooth piece. For the corresponding DAEs, numerical schemes of high accuracy may be used. This approach is natural and appealing because it leads to high-order time-stepping schemes. The major weakness of such an approach is that it excludes the presence of impulsive forces in the absence of an impact. One fairly simple example where such forces occur was pointed out in 1895 by Painlevé [34], who argued that the equations of classical rigid body dynamics are incompatible with the Coulomb friction model. Recently it was shown that a solution of Painlevé's example exists in the sense of *measure differential inclusions* (MDIs) [46].

Time-stepping schemes that are not vulnerable to Painlevé-type examples integrate the dynamics at a velocity-impulse level, thereby allowing for impulsive forces at any time instant. Most of the time-stepping schemes that build a discrete model at a velocity-impulse level are based on Euler's method for solving ordinary differential equations (ODEs). In this context, the methods of Anitescu and Potra [4] and Stewart and Trinkle [45] are based on a semi-implicit Euler scheme, while the model of [5] is based on a linearly implicit Euler scheme. All three formulations require the solution of one LCP at each time step (see [14] for an extensive analysis of LCPs).

Recently, we proposed a new time-stepping scheme based on the trapezoidal method for solving ODEs [39]. The scheme solves one LCP at each noncollisional integration step. We have shown that the numerical velocity is uniformly bounded as the integration step approaches zero, and that the scheme has global second-order convergence under additional restrictions. In order to globally achieve this convergence, events such as collision, take-off (contact deletion), and stick-slip transitions have to be detected with sufficient accuracy. To do so, we have proposed detection strategies that use information only at the position-velocity level, thereby remaining consistent with the solution concept of MDIs. We have demonstrated that the scheme behaves well (its energy stays bounded for arbitrary large stiffness parameters) for stiffness introduced by springs and dampers attached to pairs of bodies in the system. The scheme was implemented in the simulation framework UMBRA [20] and has proved successful in industrial-scale applications.

1.1. Our contribution. The treatment of joints in time-stepping schemes is not new [4, 2]. What is novel in this work is that we prove that the solution produced by a class of time-stepping schemes that solves one LCP per step and that includes the methods presented in [4] and [39] converges to a solution of an MDI, in a sense to be defined in section 6, *even when joint constraints are present.* The main conceptual novelty is the reduced friction cone which allows us to reduce the treatment of bilateral constraints to one of unilateral constraints, without altering the pointedness property. It is conceivable that a proof of the convergence of linearized backward Euler schemes can be obtained from the one in the jointless case [47] for configurations with joints if the system is represented in coordinates which eliminate the joint constraints [23]. Nonetheless, from a practical perspective, this is inadvisable for the following reasons:

- The representation of the system using only independent coordinates can rarely be done explicitly, and a difficult nonlinear system needs to be solved at each time step [23, section 7.2]
- The equations governing the dynamics of the system where joints are (locally) eliminated by the use of independent coordinates involve second-order derivatives of the joint functions [7, section 9.2.1] which may be costly to obtain.

We note that the same reasons have prompted the search for methods that are specific to differential algebraic, as opposed to ordinary differential, equations [7].

In addition, we prove for the first time that certain trapezoidal-like methods [39] converge in the sense of MDIs. In doing that, we are able to define a discrete velocity function of bounded variation, although the natural discrete velocity function produced by our algorithm may have unbounded variation. Here, the *natural discrete velocity function* is defined as the unweighted discrete velocity obtained by the algorithm in section 4 (see (4.11) for the exact definition).

2. Notation and model. In our analysis, we use the notation and framework from [47, 4]. We assume that the state of the system of rigid bodies can be described by a generalized position vector $q \in \mathbb{R}^s$ and a generalized velocity vector $v \in \mathbb{R}^s$. We assume that the system is subject to equality, nonpenetration, contact, and Coulomb friction.

The *equality constraints* that we consider in this work are described by equations

$$\begin{aligned} \Theta^{(i)}(q) &= 0, \quad i = 1, 2, \ldots, m_J, \\ \nu^{(i)}(q)^T v &= 0, \quad i = m_J + 1, m_J + 2, \ldots, m, \end{aligned} \tag{2.1}$$

where $\Theta^{(i)}$, $i = 1, 2, \ldots, m_J$ and $\nu^{(i)}(q)$, $i = m_J + 1, m_J + 2, \ldots, m$ are sufficiently smooth functions. The first m_J equality constraints are holonomic and usually originate from the transcription of joint constraints [23], whereas the last $m - m_J$ constraints are Pfaffian constraints [32]. A holonomic constraint can be represented as a Pfaffian constraint, but the reverse is not necessarily true [32].

The force exerted by constraint (i), $i = 1, 2, \ldots, m_J$ on the system is $c_\nu^{(i)} \nu^{(i)}(q)$, where $\nu^{(i)}(q) = \nabla_q \Theta^{(i)}(q)$ is the gradient of $\Theta^{(i)}(q)$ and $c_\nu^{(i)}$ is the appropriate Lagrange multiplier [23]. To simplify our discussion we may refer to the constraints (2.1) as joint constraints, even if only the first m_J are technically such constraints.

The *nonpenetration constraints* are generated by the rigid body hypothesis according to which the bodies constituting the system cannot penetrate each other. We assume that, for any pair of bodies, we can define a signed distance function $\Phi^{(j)}(q)$ so that the noninterpenetration constraints can be written as

$$\Phi^{(j)}(q) \geq 0, \quad j = 1, 2, \ldots, p\,, \tag{2.2}$$

where p is the number of pairs of bodies of the system that could get in contact, which in most applications is substantially smaller than the number of all possible pairs of bodies. Details of how the functions $\Phi^{(j)}$ can be defined and calculated are presented in [1].

The contact and frictional constraints may be introduced by means of the active set and the friction cone. If $\Phi^{(j)}(q) > 0$, then the jth constraint doesn't contribute to the dynamics of the system. When $\Phi^{(j)}(q) = 0$, however, the contact impulse generated by the jth noninterpenetration constraint must lie inside the *contact friction*

cone:

$$FC^{(j)}(q) = \left\{ z_c = n^{(j)} {c_n}^{(j)} + \overline{D}^{(j)} \beta^{(j)} \,\middle|\, {c_n}^{(j)} \geq 0,\ \|\beta^{(j)}\|_2 \leq \mu^{(j)} c_n^{(j)} \right\}, \tag{2.3}$$

where we used the simplified notation $n^{(j)} := n^{(j)}(q) = \nabla_q \Phi^{(j)}(q)$ and $\overline{D}^{(j)} := \overline{D}^{(j)}(q)$. Here the columns of $\overline{D}^{(j)} \in \mathbb{R}^{s\times 2}$ span the friction space, and $\beta^{(j)} \in \mathbb{R}^2$ is the corresponding tangential impulse due to friction. The parameter $\mu^{(j)} \geq 0$, which may be different for each contact, is the friction coefficient, and the second inequality that involves it in (2.3) is the first part of the Coulomb law. By including the joint forces in the above multivalued map, we obtain what we call the *constraint friction cone* $\mathcal{FC}^{(j)}(q)$ corresponding to the jth contact. More precisely, we have

$$\mathcal{FC}^{(j)}(q) = \left\{ z = \widetilde{\nu}\widetilde{c}_\nu + n^{(j)} {c_n}^{(j)} + \overline{D}^{(j)} \beta^{(j)} \,\middle|\, {c_n}^{(j)} \geq 0,\ \|\beta^{(j)}\|_2 \leq \mu^{(j)} c_n^{(j)} \right\}, \tag{2.4}$$

where

$$\widetilde{\nu} = \left[\nu^{(1)}, \nu^{(2)}, \ldots, \nu^{(m)} \right], \quad \widetilde{c}_\nu = \left[c_\nu^{(1)}, c_\nu^{(2)}, \ldots, c_\nu^{(m)} \right], \tag{2.5}$$

with $\nu^{(i)} := \nu^{(i)}(q)$. Here we have used $[\cdot, \cdot, \ldots, \cdot]$ to denote a block matrix with the same number of rows as its blocks and $\lceil \cdot, \cdot, \ldots, \cdot \rfloor$ to denote a block matrix with the same number of columns as its blocks. The total friction cones are then defined by

$$FC(q) = \sum_{\Phi^{(j)}(q)=0} FC^{(j)}(q) \tag{2.6}$$

for the *total contact friction cone* and by

$$\mathcal{FC}(q) = \sum_{\Phi^{(j)}(q)=0} \mathcal{FC}^{(j)}(q) \tag{2.7}$$

for the *total constraint friction cone*. To simplify terminology, we will refer, unless specified in advance, to the cone in (2.7) as the *total friction cone*. Note that the definition above implies that the set of active contact constraints $\mathcal{A}$ is determined by

$$\mathcal{A} = \left\{ j \in \{1, \ldots, p\} \ :\ \Phi^{(j)}(q) = 0 \right\}$$

for given position q. The total friction cone can be approximated by a polyhedral cone [45]. That is, $\mathcal{FC}^{(j)}(q)$ is replaced by

$$\widehat{\mathcal{FC}}^{(j)}(q) = \left\{ z = \widetilde{\nu}\widetilde{c}_\nu + n^{(j)} {c_n}^{(j)} + D^{(j)} \beta^{(j)} \,\middle|\, {c_n}^{(j)} \geq 0,\ \beta^{(j)} \geq 0,\ \|\beta^{(j)}\|_1 \leq \mu^{(j)} c_n^{(j)} \right\}, \tag{2.8}$$

where $D^{(j)} := D^{(j)}(q) \in R^{s\times m_C}$ is a balanced matrix in the sense that if ${d_i}^{(j)}$ is a column of $D^{(j)}$, then there is another index k such that ${d_i}^{(j)} = -{d_k}^{(j)}$. In this way we can represent the frictional impulses by using a nonnegative vector of multipliers $\beta^{(j)} = (\beta^{(j)})_i \geq 0$, with the 2-norm being replaced by the 1-norm. Here the nonnegative integer m_C represents the number of edges used in the approximation of the full cone. The polyhedral approximation of the friction cone is then given by

$$\widehat{\mathcal{FC}}(q) = \sum_{j\in\mathcal{A}} \widehat{\mathcal{FC}}^{(j)}(q) = \left\{ z = \widetilde{\nu}\widetilde{c}_\nu + \widetilde{n}\widetilde{c}_n + \widetilde{D}\widetilde{\beta} \,\middle|\, \widetilde{c}_n \geq 0,\ \widetilde{\beta} \geq 0,\ \|\beta^{(j)}\|_1 \leq \mu^{(j)} c_n^{(j)}\ \forall j \in \mathcal{A} \right\}, \tag{2.9}$$

where the block matrices $\widetilde{n}$, $\widetilde{c}_n$, $\widetilde{D}$, and $\widetilde{\beta}$ are defined by

$$\begin{aligned} \widetilde{n} &= \left[n^{(j_1)}, n^{(j_1)}, \ldots, n^{(j_a)}\right], \qquad \widetilde{c}_n = \left[c_n^{(j_1)}, c_n^{(j_2)}, \ldots, c_n^{(j_a)}\right], \\ \widetilde{D} &= \left[D^{(j_1)}, D^{(j_2)}, \ldots, D^{(j_a)}\right], \quad \widetilde{\beta} = \left[\beta^{(j_1)}, \beta^{(j_2)}, \ldots, \beta^{(j_a)}\right] \end{aligned} \tag{2.10}$$

for an active set $\mathcal{A} = \{j_1, j_2, \ldots, j_a\}$.

3. Total friction cone and regularity assumptions. A regularity assumption, the pointedness of the friction cone, is used to obtain convergence results in the contact-only case [6, 47]. We present here an extension of the pointedness assumption to the case including bilateral constraints.

DEFINITION 3.1. *We say that*

(3.1)
$$\begin{aligned} &\mathcal{FC}(q) \text{ is } \textbf{pointed} \Leftrightarrow \forall \left(\tilde{c}_\nu, \tilde{c}_n \geq 0, \tilde{\beta}\right) \neq 0 \text{ such that } \left\|\beta^{(j)}\right\|_2 \leq \mu^{(j)} c_n^{(j)} \ \forall j \in \mathcal{A} \\ &\qquad \text{we must have that } \widetilde{\nu}\widetilde{c}_\nu + \widetilde{n}\widetilde{c}_n + \widetilde{\widetilde{D}}\widetilde{\beta} \neq 0. \end{aligned}$$

(3.2)
$$\begin{aligned} &\widehat{\mathcal{FC}}(q) \text{ is } \textbf{pointed} \Leftrightarrow \forall \left(\widetilde{c}_\nu, \widetilde{c}_n \geq 0, \widetilde{\beta} \geq 0\right) \neq 0 \text{ such that } \left\|\beta^{(j)}\right\|_1 \leq \mu^{(j)} c_n^{(j)} \ \forall j \in \mathcal{A} \\ &\qquad \text{we must have that } \widetilde{\nu}\widetilde{c}_\nu + \widetilde{n}\widetilde{c}_n + \widetilde{D}\widetilde{\beta} \neq 0. \end{aligned}$$

This definition clearly implies that the joint-constraint matrix $\widetilde{\nu}$ is of full rank. Moreover, the pointed friction cone assumption is weaker than the linear independence of the columns of the matrix $(\widetilde{\nu}^T,\ \widetilde{n}^T,\ \widetilde{D}^T)^T$. Its name originates in the fact that, when there are no joint constraints, the condition is equivalent to the cone's not containing any proper linear subspace and thus being "pointed." An equivalent definition of the pointed friction cone assumption is given by the following condition [6]:

$$\begin{aligned} &\mathcal{FC}(q) \text{ is pointed} \Leftrightarrow \text{there exists } c_{\mathcal{FC}} > 0, \text{ such that } \left\|\left(\widetilde{c}_\nu, \widetilde{c}_n, \widetilde{\beta}\right)\right\| \leq c_{\mathcal{FC}}\|z\| \\ &\qquad \text{with } z = \widetilde{\nu}\widetilde{c}_\nu + \widetilde{n}\widetilde{c}_n + \widetilde{\widetilde{D}}\widetilde{\beta} \in \mathcal{FC}(q). \end{aligned} \tag{3.3}$$

$$\begin{aligned} &\widehat{\mathcal{FC}}(q) \text{ is pointed} \Leftrightarrow \text{there exists } c_{\widehat{\mathcal{FC}}} > 0, \text{ such that } \left\|\left(\widetilde{c}_\nu, \widetilde{c}_n, \widetilde{\beta}\right)\right\| \leq c_{\widehat{\mathcal{FC}}}\|z\| \\ &\qquad \text{with } z = \widetilde{\nu}\widetilde{c}_\nu + \widetilde{n}\widetilde{c}_n + \widetilde{\widetilde{D}}\widetilde{\beta} \in \widehat{\mathcal{FC}}(q). \end{aligned} \tag{3.4}$$

We say that the total friction cone $\mathcal{FC}(q)$ ($\widehat{\mathcal{FC}}(q)$) is **uniformly pointed** if the constant $c_{\mathcal{FC}}$ ($c_{\widehat{\mathcal{FC}}}$) can be taken the same for all possible configurations q. As noted in [2], the pointedness assumption is equivalent (in the frictionless case) to the existence of a force that will dissasemble all contacts without breaking the joints.

LEMMA 3.2. *Assume that $\mathcal{FC}(q)$ is pointed. Let*

$$z = \widetilde{n}\widetilde{c}_n + \widetilde{\widetilde{D}}\widetilde{\beta}, \ \text{ where } \ \widetilde{c}_n \geq 0, \ \widetilde{c}_n \neq 0, \ \ \left\|\beta^{(j)}\right\|_2 \leq \mu^{(j)} c_n^{(j)} \ \forall j \in \mathcal{A}.$$

That is, z is an element of the (full) friction cone obtained by excluding the bilateral constraints, with the normal impulses not all equal to 0. *Then,*

$$z \neq 0 \ \text{ and } \ z \notin \text{Range}(\widetilde{\nu}).$$

Proof. As suggested by the claim above, consider the set

$$FC(q) = \left\{ z_c = \widetilde{n}\widetilde{c}_n + \widetilde{\widetilde{D}}\widetilde{\beta} \,\middle|\, \widetilde{c}_n \geq 0, \ \left\|\beta^{(j)}\right\|_2 \leq \mu^{(j)} c_n^{(j)} \ \forall j \in \mathcal{A} \right\} \tag{3.5}$$

(note the difference in notation: caligraphic characters denote the friction cone that includes all constraint impulses, while roman letters denote the cone that containins only the contact impulses). Clearly the pointedness of $\mathcal{FC}(q)$ implies the pointedness of $FC(q)$. Therefore, by taking $z \in FC(q)$ with the normal impulses not all zero ($\widetilde{c}_n \neq 0$), we obtain $z \neq 0$. If $z \in \text{Range}(\widetilde{\nu})$, then $z = \widetilde{\nu}u$, $u \neq 0$, and therefore by taking $\overline{z} = -\widetilde{\nu}u + z$, we have $\overline{z} \in \mathcal{FC}(q)$, $\overline{z} = 0$ with a nonzero constraint impulse, a contradiction to the pointedness of $\mathcal{FC}(q)$. ∎

We will use this lemma to analyze the properties of the set

$$\widetilde{\nu}_\perp^T \mathcal{FC}(q) = \left\{ \widetilde{\nu}_\perp^T z \ : \ z \in \mathcal{FC}(q) \right\}.$$

Here $\widetilde{\nu}_\perp$ denotes the orthogonal complement of $\widetilde{\nu} \in \mathbb{R}^{s\times m}$. More precisely, $\widetilde{\nu}_\perp \in \mathbb{R}^{s\times(s-m)}$ such that $\widetilde{\nu}_\perp^T \widetilde{\nu} = 0$ and $\widetilde{\nu}_\perp^T \widetilde{\nu}_\perp = I$. It follows that, for any $x \in \mathbb{R}^s$, there exist unique vectors $u \in \mathbb{R}^m$ and $w \in \mathbb{R}^{s-m}$ such that the decomposition

$$x = \widetilde{\nu}u + \widetilde{\nu}_\perp w \tag{3.6}$$

holds. We have the following simple results.

LEMMA 3.3. *Assume that $\mathcal{FC}(q)$ is pointed. Then for all $j \in \mathcal{A}$, we have*

$$\widetilde{\nu}_\perp^T n^{(j)} \neq 0.$$

Proof. The proof follows immediately from Lemma 3.2. More precisely, assume that $\widetilde{\nu}_\perp^T n^{(j)} = 0$ for some $j \in \mathcal{A}$. Take $z = c_n^{(j)} n^{(j)}$, with ${c_n}^{(j)} > 0$ (note that $n^{(j)} \neq 0$). It follows that $\widetilde{\nu}_\perp^T z = 0$. Therefore, by the decomposition above, we must have $z \in \text{Range}(\widetilde{\nu})$, which contradicts the conclusion of Lemma 3.2. ∎

Remark 3.4. *We cannot say the same thing about $\widetilde{\nu}_\perp^T {\overline{d}_i}^{(j)}$, where ${\overline{d}_i}^{(j)}$ is a column of $\overline{D}^{(j)}$. Actually it is possible to have $\widetilde{\nu}_\perp^T {\overline{d}_i}^{(j)} = 0$, which shows once again that the pointedness assumption is weaker than the linear independence of the active set (active set here includes all constraints).*

Let us define

$${W_n}^{(j)} = \widetilde{\nu}_\perp^T n^{(j)} \ \text{ and } \ {\overline{W}_D}^{(j)} = \widetilde{\nu}_\perp^T \overline{D}^{(j)}.$$

As discussed above, all of the ${W_n}^{(j)}$ are nonzero vectors; thus, by adjoining all of them, we obtain a matrix that we denote by $\widetilde{W}_n$. By taking only those ${\overline{W}_D}^{(j)}$ that are nonzero and adjoining we obtain, in a similar fashion, a matrix denoted by $\widetilde{\overline{W}}_D$. Now let us define the *full reduced friction cone* $FC_r(q)$ by

$$\begin{aligned} FC_r(q) = \Big\{ z_r = \widetilde{W}_n \widetilde{\overline{c}}_n + \widetilde{\overline{W}}_D \widetilde{\overline{\beta}} \Big| \widetilde{\overline{c}}_n \geq 0, \ \left\| \overline{\beta}^{(j)} \right\|_2 \\ \leq \mu^{(j)} \overline{c}_n^{(j)}, \ \Big[\forall j \in \mathcal{A} \ \text{ such that (s.t.) } \ \left(\widetilde{\nu}_\perp(q)\right)^T \overline{D}^{(j)}(q) \neq 0 \Big] \Big\}, \end{aligned} \tag{3.7}$$

where the matrix $\widetilde{\overline{W}}_D$ is assumed to have only nonzero columns. In a similar fashion we introduce the *polyhedral reduced friction cone* $\widehat{FC}_r(q)$:

$$\begin{aligned} \widehat{FC}_r(q) = \Big\{ z_r = \widetilde{W}_n \widetilde{\overline{c}}_n + \widetilde{W}_D \widetilde{\overline{\beta}} \Big| \widetilde{\overline{c}}_n \geq 0, \ \widetilde{\overline{\beta}} \geq 0, \ \left\| \overline{\beta}^{(j)} \right\|_1 \\ \leq \mu^{(j)} \overline{c}_n^{(j)}, \ \Big[\forall j \in \mathcal{A} \text{ s.t. } \left(\widetilde{\nu}_\perp(q)\right)^T D^{(j)}(q) \neq 0 \Big] \Big\}. \end{aligned} \tag{3.8}$$

The active set used for the reduced friction cone is the same as the one used for the nonreduced one. However, the number of frictional contacts in the reduced cone may be smaller than the number in the nonreduced cone. We have the following result.

LEMMA 3.5. *If $\mathcal{FC}(q)$ ($\widehat{\mathcal{FC}}(q)$) is pointed for all q, then the full (polyhedral) reduced friction cone $FC_r(q)$ ($\widehat{FC}_r(q)$) is pointed for all q. Here the pointedness of $FC_r(q)$ ($\widehat{FC}_r(q)$) is to be understood in the sense of Definition* 3.1 *where the joint constraints are omitted. This is the same with applying this definition to the representations* (3.7) *and* (3.8).

Proof. Let q be any possible system configuration, and let z_r be an arbitrary element of $FC_r(q)$. Then z_r can be written as

$$z_r = \widetilde{\nu}_\perp^T \widetilde{n}\widetilde{c}_n + \widetilde{\nu}_\perp^T \widetilde{\overline{D}}\widetilde{\beta},$$

where we have already eliminated those columns of $\widetilde{\overline{D}}$ that are in the range of $\widetilde{\nu}$. This new matrix is denoted by $\widetilde{\overline{\overline{D}}}$, and the corresponding frictional impulses are given by the vector $\widetilde{\overline{\beta}}$. Note that, as shown above, all of the columns of $\widetilde{n}$ have nonzero components outside the range of $\widetilde{\nu}$. The normal and tangential impulses $(\widetilde{c}_n, \widetilde{\overline{\beta}})$ satisfy $c_n^{(j)} \geq 0$ for all $j \in \mathcal{A}$ and $\|\overline{\beta}^{(j)}\|_2 \leq \mu^{(j)} c_n^{(j)}$ for all $j \in \mathcal{A}$ such that $(\widetilde{\nu}_\perp(q))^T \overline{D}^{(j)}(q) \neq 0$.

Assume now that $z_r = 0$, with $\widetilde{c}_n \neq 0$ (a necessary condition for $(\widetilde{c}_n, \widetilde{\overline{\beta}}) \neq 0$). We want to reach a contradiction to the pointedness of the nonreduced cone. This immediately follows from the fact that

$$z_r = 0, \widetilde{c}_n \neq 0 \Rightarrow z_c = \widetilde{n}\widetilde{c}_n + \widetilde{\overline{D}}\widetilde{\beta} \in FC(q) \text{ satisfies } z_c = \widetilde{\nu}u,\ u \neq 0.$$

Here $\widetilde{\beta}$ is obtained from $\widetilde{\overline{\beta}}$ by adding zeros to the columns of $\widetilde{\overline{D}}$ missing in $\widetilde{\overline{\overline{D}}}$. By taking $\widetilde{c}_\nu = -u$, we obtain that $z := -\widetilde{\nu}u + z_c \in \mathcal{FC}(q)$ is zero, but $(\widetilde{c}_\nu, \widetilde{c}_n, \widetilde{\beta}) \neq 0$, which contradicts the pointedness of $\mathcal{FC}(q)$. Given that the argument can be carried out for any q for which $\mathcal{FC}(q)$ is pointed, we obtained the pointedness for $FC_r(q)$. Following the same argument, one proves the pointedness of the polyhedral reduced friction cone $\widehat{FC}_r(q)$. ∎

Remark 3.6. The pointedness of the reduced cones is equivalent to the usual notion of pointedness, that is, "a cone K is pointed $\Leftrightarrow K \cap (-K) = \{0\}$."

Now let $z = \widetilde{\nu}\widetilde{c}_\nu + \widetilde{n}\widetilde{c}_n + \widetilde{\overline{D}}\widetilde{\beta} \in \mathcal{FC}(q)$. From the pointedness of the reduced cone there exists [47] a unitary vector $u_0 := u_0(q)$ and the constants $C_2 := C_2(q) > 0$, $C_3 := C_3(q) > 0$ such that, for any $z = \widetilde{\nu}\widetilde{c}_\nu + \widetilde{n}\widetilde{c}_n + \widetilde{\overline{D}}\widetilde{\beta} \in \mathcal{FC}(q)$, we have

$$u_0{}^T \widetilde{\nu}_\perp^T z \geq C_2 \|z_r\| \geq C_3 \|\widetilde{c}_n\|, \tag{3.9}$$

where $z_r = \widetilde{\nu}_\perp^T z$. This estimate is one of the main ingredients that will be later used in proving the uniform bound on the variation of the velocities.

We can visualize the friction cones $\mathcal{FC}(q)$ and $\widehat{\mathcal{FC}}(q)$ as mappings from $\mathbb{R}^s$ to the subsets of $\mathbb{R}^s$, that is, $\mathcal{FC}(q), \widehat{\mathcal{FC}}(q) : \mathbb{R}^s \to \mathcal{P}(\mathbb{R}^s)$. The graph of $\mathcal{FC}(\cdot)$ is defined by

$$graph(\mathcal{FC}) = \{(q, z(q)) \mid z(q) \in \mathcal{FC}(q)\}, \tag{3.10}$$

and similarly for $\widehat{\mathcal{FC}}(q)$. Clearly, from the constructions above, $\mathcal{FC}(q)$ and its approximation $\widehat{\mathcal{FC}}(q)$ are convex sets for each fixed q. Under the uniform pointedness assumption we obtain that these mappings have closed graphs.

LEMMA 3.7 (closed graph property of the friction cones). *Assume that* $\mathcal{FC}(q)$ $(\widehat{\mathcal{FC}}(q))$ *is uniformly pointed. Then the graph of* $\mathcal{FC}(\cdot)$ $(\widehat{\mathcal{FC}}(\cdot))$ *is closed.*

Proof. We will prove the result for $\mathcal{FC}(\cdot)$. A similar argument is used for the polyhedral approximation $\widehat{\mathcal{FC}}(\cdot)$. Consider a sequence $(q^n, z^n) \in graph(\mathcal{FC}(q^n))$, where z^n has the form $z^n = \widetilde{\nu}(q^n)c_\nu^n + \widetilde{n}(q^n)\widetilde{c}_n^n + \widetilde{\overline{D}}(q^n)\widetilde{\beta}^n$. Assume that $q^n \to q$ and $z^n \to z$ as $n \to \infty$. We want to show that $z \in \mathcal{FC}(q)$. From the uniform pointedness of $\mathcal{FC}(\cdot)$ we obtain that

$$\left\|\left(\widetilde{c}_\nu^n, \widetilde{c}_n^n, \widetilde{\beta}^n\right)\right\| \le C_{\mathcal{FC}}\|z^n\|,$$

where $C_{\mathcal{FC}}$ is independent of q^n. Given that $z^n \to z$, it follows from the above inequality that $\|(\widetilde{c}_\nu^n, \widetilde{c}_n^n, \widetilde{\beta}^n)\|$ is bounded, and therefore we can extract convergent subsequences $\widetilde{c}_\nu^{n_k} \to \widetilde{c}_\nu^*$, $\widetilde{c}_n^{n_k} \to \widetilde{c}_n^*$, and $\widetilde{\beta}^{n_k} \to \widetilde{\beta}^*$, where $\widetilde{c}_n^* \ge 0$ and $\|\beta^{(j),*}\|_2 \le \mu^{(j)}\|c_n^{(j),*}\|$ due to the similar inequalities satisfied by the corresponding subsequence. Using the fact that $\widetilde{\nu}(\cdot)$, $\widetilde{n}(\cdot)$, and $\widetilde{\overline{D}}(\cdot)$ are continuous, we have that the subsequence z^{n_k} converges to $z^* = \widetilde{\nu}(q)c_\nu^* + \widetilde{n}(q)\widetilde{c}_n^* + \widetilde{\overline{D}}(q)\widetilde{\beta}^* \in \mathcal{FC}(q)$. Given that z^{n_k}, with $z^{n_k} \to z^*$, is a subsequence of the convergent sequence z^n, with $z^n \to z$, we conclude that $z = z^* \in \mathcal{FC}(q)$, which proves our claim. ∎

Note that the closed graph property implies that the values of the multivalued mappings are closed. This can be easily seen by taking $q^n = q$ and $z^n \in \mathcal{FC}(q)$. An immediate consequence of the above lemma is the following corollary.

COROLLARY 3.8. *Assume that* $\mathcal{FC}(q)$ $(\widehat{\mathcal{FC}}(q))$ *is uniformly pointed. Then the graph of* $FC_r(\cdot)$ $(\widehat{FC}_r(\cdot))$ *is closed.*

Proof. Consider a sequence (q^n, z_r^n) such that $z_r^n \in FC_r(q^n)$ and $q^n \to q$, $z_r^n \to z_r$, as $n \to \infty$. By definition, $z_r^n = (\widetilde{\nu}_\perp(q^n))^T z^n$ for some $z^n \in \mathcal{FC}(q^n)$. Taking the limit, as $n \to \infty$, we conclude that $z^n \to z$. The fact that $\mathcal{FC}(\cdot)$ has a closed graph implies that $z \in \mathcal{FC}(q)$, which immediately leads to $z_r = (\widetilde{\nu}_\perp(q))^T z \in FC_r(q)$. ∎

4. The time-stepping scheme. We are interested in convergence properties for a family of linearly implicit time-stepping schemes that accommodate methods based on semi-implicit Euler methods [4, 46] as well as various instances of the trapezoidal method from [39]. The time-stepping scheme solves at each integration step an LCP. We will assume that only inelastic collisions are solved. In terms of the collision rule given in [4] which involves a Poisson rule with a compression phase followed by a decompression phase, for inelastic collisions only the former LCP needs to be solved, and therefore the algorithm will solve only one LCP per time step. The main difference between a noncollisional and a collisional integration step is that the latter uses a zero time step to get out of the compression phase.

Some of the results that we obtain, such as the uniform boundedness of the velocity sequence, will hold even in the case of partially elastic collisions, since we will assume that the number of collisions is bounded above with respect to the time step. Such a proof has been obtained for a different, though related, time-stepping scheme [4]. The problem, however, is that the Poisson rule is very difficult to prove (or, indeed, to state) in the framework of MDIs from section 6, where the notion of instantaneous reaction impulse does not truly appear. A possibility may be to consider the case of collisions with a Newton rule [19, 51], though the weak convergence concept we are using results in only almost everywhere convergence of the velocity. In turn, this results in difficulties when dealing with statements involving instantaneous velocity as is the case of Newton rules. We defer convergence issues involving partially elastic collisions to future research.

To write the integration step as a *mixed linear complementarity problem* (MLCP), we use the following approximations. The joint constraints are written at the velocity level and approximated by

$$\left(\nu^{(i)}\left(q^l\right)\right)^T\left(\alpha v^{l+1}+(1-\alpha)v^l\right)=0,\quad i=1,\dots,m,$$

where α is a scalar parameter $\alpha\in(0,1]$.

The nonpenetration and frictional constraints are approximated in the same fashion. We can write these as the following complementarity conditions:

$$\begin{aligned}
0\le\rho^{(j),l+1}&:=\left(n^{(j)}\left(q^l\right)\right)^T\left(\alpha v^{l+1}+(1-\alpha)v^l\right)\perp c_n^{(j),l+1}\ge 0,\quad j\in\mathcal{A},\\
0\le\sigma^{(j),l+1}&:=\lambda^{(j),l+1}e^{(j)}+\left(D^{(j)}\left(q^l\right)\right)^T\left(\alpha v^{l+1}+(1-\alpha)v^l\right)\perp\beta^{(j),l+1}\ge 0,\quad j\in\mathcal{A},\\
0\le\zeta^{(j),l+1}&:=\mu^{(j)}c_n^{(j),l+1}-e^{(j)^T}\beta^{(j),l+1}\perp\lambda^{(j),l+1}\ge 0,\quad j\in\mathcal{A}.
\end{aligned}$$

Here $e^{(j)}$ is a vector of dimension $m_C^{(j)}$ whose every entry is 1. The equations of motion in implicit form can be written as

$$M\left(v^{l+1}-v^l\right)-z^{l+1}=hk\left(t_{l+1},q^{l+1},v^{l+1}\right). \tag{4.1}$$

Here M is the mass matrix, which is assumed to be a constant symmetric positive definite matrix, z^{l+1} represent the contact and joint impulses, and $k(t_{l+1},q^{l+1},v^{l+1})$ are the inertial and applied forces acting at time t_{l+1}. Since the goal is to formulate the integration step as an LCP, we will linearize (4.1) as follows. The term

$$z^{l+1}=\widetilde{\nu}\left(q^{l+1}\right)\widetilde{c}_\nu^{l+1}+\widetilde{n}\left(q^{l+1}\right)\widetilde{c}_n^{l+1}+\widetilde{D}\left(q^{l+1}\right)\widetilde{\beta}^{l+1}$$

is replaced by

$$z^{l+1}=\widetilde{\nu}^l\widetilde{c}_\nu^{l+1}+\widetilde{n}^l\widetilde{c}_n^{l+1}+\widetilde{D}^l\widetilde{\beta}^{l+1},$$

where $\widetilde{\nu}^l=\widetilde{\nu}(q^l)$, $\widetilde{n}^l=\widetilde{n}(q^l)$ and $\widetilde{D}^l=\widetilde{D}(q^l)$. To linearize the term $k(t_{l+1},q^{l+1},v^{l+1})$ in (4.1), we first introduce the position update formula. Given a parameter $\gamma\in[0,1]$ (fixed at the beginning of the simulation), we obtain the position at time t_{l+1} by the formula

$$q^{l+1}=q^l+h\left((1-\gamma)v^l+\gamma v^{l+1}\right).$$

For the term $k(t_{l+1},q^{l+1},v^{l+1})$ we have

$$\begin{aligned}
k\left(t_{l+1},q^{l+1},v^{l+1}\right)&=f_C\left(v^{l+1}\right)+k_1\left(t_{l+1},q^{l+1},v^{l+1}\right)\\
&=F\left(v^{l+1}\right)v^{l+1}+k_1\left(t_{l+1},q^{l+1},v^{l+1}\right),
\end{aligned}$$

where $f_C(v^{l+1})=F(v^{l+1})v^{l+1}$ are the Coriolis forces and $k_1(t_{l+1},q^{l+1},v^{l+1})$ are the external forces. A discussion related to this representation of the Coriolis forces is given at the end of this section. We replace the Coriolis term by

$$F\left(v^{l+1}\right)v^{l+1}\approx F\left(v^l\right)\left((1-\alpha)v^l+\alpha v^{l+1}\right)=F\left(v^l\right)v^l+\alpha F\left(v^l\right)\left(v^{l+1}-v^l\right). \tag{4.2}$$

The term $k_1(t_{l+1},q^{l+1},v^{l+1})$ is approximated as follows:

$$\begin{aligned}
k_1\left(t_{l+1},q^{l+1},v^{l+1}\right)&\approx(1-\alpha)k_1\left(t_l,q^l,v^l\right)+\alpha k_1\left(t_{l+1},q^{l+1},v^{l+1}\right),\\
&\approx(1-\alpha)k_1\left(t_l,q^l,v^l\right)+\alpha k_1\left(t_{l+1},q^l,v^l\right)\\
&\quad+\alpha\left(\widetilde{k}_{1q}\left(q^{l+1}-q^l\right)+\widetilde{k}_{1v}^l\left(v^{l+1}-v^l\right)\right),\\
&\approx(1-\alpha)k_1\left(t_l,q^l,v^l\right)+\alpha k_1\left(t_{l+1},q^l,v^l\right)\\
&\quad+\alpha h\widetilde{k}_{1q}^l v^l+\alpha\left(\widetilde{k}_{1v}^l+\gamma h\widetilde{k}_{1q}^l\right)\left(v^{l+1}-v^l\right),
\end{aligned} \tag{4.3}$$

where

$$\widetilde{k}_{1q}^{l} \approx k_{1q}\left(t_{l+1}, q^{l}, v^{l}\right), \quad \widetilde{k}_{1v}^{l} \approx k_{1v}\left(t_{l+1}, q^{l}, v^{l}\right)$$

are approximations of the Jacobians k_{1q} and k_{1v}, respectively. Combining the equations of motion with the joint constraints described at the velocity level and the frictional contact constraints, we obtain the following time-stepping scheme:

$$q^{l+1} = q^{l} + h\left((1-\gamma)v^{l} + \gamma v^{l+1}\right), \tag{4.4a}$$

(4.4b)

$$\widetilde{M}^{l} v^{l+1} - \sum_{i=1}^{m} \nu^{(i),l} c_{\nu}^{(i),l+1} - \sum_{j\in\mathcal{A}} (n^{(j),l} c_{n}^{(j),l+1} + D^{(j),l}\beta^{(j),l+1}) = \widetilde{M}^{l} v^{l} + \widetilde{k}^{l},$$

$$\left(\nu^{(i),l}\right)^{T}\left(\alpha v^{l+1} + (1-\alpha)v^{l}\right) = 0\,,\ i = 1, 2, \ldots m, \tag{4.4c}$$

$$0 \le \rho^{(j),l+1} := \left(n^{(j),l}\right)^{T}\left(\alpha v^{l+1} + (1-\alpha)v^{l}\right) \perp c_{n}^{(j),l+1} \ge 0, \quad j \in \mathcal{A}, \tag{4.4d}$$

(4.4e)

$$0 \le \sigma^{(j),l+1} := \lambda^{(j),l+1} e^{(j)} + \left(D^{(j),l}\right)^{T}\left(\alpha v^{l+1} + (1-\alpha)v^{l}\right) \perp \beta^{(j),l+1} \ge 0, \quad j \in \mathcal{A},$$

$$0 \le \zeta^{(j),l+1} := \mu^{(j)} c_{n}^{(j),l+1} - {e^{(j)}}^{T}\beta^{(j),l+1} \perp \lambda^{(j),l+1} \ge 0, \quad j \in \mathcal{A}, \tag{4.4f}$$

where $\nu^{(i),l} = \nu^{(i)}(q^{l})$, $n^{(j),l} = n^{(j)}(q^{l})$, $D^{(j),l} = D^{(j)}(q^{l})$, and

$$\widetilde{M}^{l} = \left(M - \alpha h\left(F(v^{l}) + \widetilde{k}_{1v}^{l}\right) - \alpha\gamma h^{2}\,\widetilde{k}_{1q}^{l}\right),$$

$$\widetilde{k}^{l} = h\left((1-\alpha)k_{1}\left(t_{l}, q^{l}, v^{l}\right) + \alpha k_{1}\left(t_{l+1}, q^{l}, v^{l}\right)\right) + (1-\alpha)hF(v^{l})v^{l} + \alpha h^{2}\widetilde{k}_{1q}^{l} v^{l}. \tag{4.5}$$

We note that (4.4a)–(4.4f) represent an MLCP. If at time-step l the index set of active contact constraints is given by $\mathcal{A} = \{j_1, j_2, \ldots, j_a\}$ and if we use (2.5), (2.10) together with

(4.6)

$$\widetilde{\lambda} = \left[\lambda^{(j_1)}, \lambda^{(j_2)}, \ldots, \lambda^{(j_a)}\right],\ \widetilde{\zeta} = \left[\zeta^{(j_1)}, \zeta^{(j_2)}, \ldots, \zeta^{(j_a)}\right],\ \widetilde{E} = \operatorname{diag}\left(e^{(j_1)}, e^{(j_2)}, \ldots, e^{(j_a)}\right)$$

$$\widetilde{\sigma} = \left[\sigma^{(j_1)}, \sigma^{(j_2)}, \ldots, \sigma^{(j_a)}\right],\ \widetilde{\rho} = \left[\rho^{(j_1)}, \rho^{(j_2)}, \ldots, \rho^{(j_a)}\right],\ \widetilde{\mu} = \operatorname{diag}\left(\mu^{(j_1)}, \mu^{(j_2)}, \ldots, \mu^{(j_a)}\right),$$

then the matrix form of the integration step is given by

$$\begin{bmatrix} \widetilde{M}^{l} & -\widetilde{\nu}^{l} & -\widetilde{n}^{l} & -\widetilde{D}^{l} & 0 \\ \left(\widetilde{\nu}^{l}\right)^{T} & 0 & 0 & 0 & 0 \\ \left(\widetilde{n}^{l}\right)^{T} & 0 & 0 & 0 & 0 \\ \left(\widetilde{D}^{l}\right)^{T} & 0 & 0 & 0 & \widetilde{E} \\ 0 & 0 & \widetilde{\mu} & -\widetilde{E}^{T} & 0 \end{bmatrix} \begin{bmatrix} v^{l+1} \\ \widetilde{c}_{\nu}^{l+1} \\ \widetilde{c}_{n}^{l+1} \\ \widetilde{\beta}^{l+1} \\ \widetilde{\lambda}^{l+1} \end{bmatrix} - \begin{bmatrix} \widetilde{M}^{l} v^{l} + \widetilde{k}^{l} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \widetilde{\rho}^{l+1} \\ \widetilde{\sigma}^{l+1} \\ \widetilde{\zeta}^{l+1} \end{bmatrix}, \tag{4.7}$$

$$0 \le \left[\widetilde{c}_{n}^{l+1}, \widetilde{\beta}^{l+1}, \widetilde{\lambda}^{l+1}\right] \perp \left[\widetilde{\rho}^{l+1}, \widetilde{\sigma}^{l+1}, \widetilde{\zeta}^{l+1}\right] \ge 0\,. \tag{4.8}$$

We denote by $\mathcal{L}(q^{l}, v^{l}, \widetilde{k}, h, \alpha, \gamma)$ the solution set of the MLCP (4.7).

We note that the choice $\alpha = \gamma = 1$ results in the scheme from [4, 5] (the former is obtained once we choose $\widetilde{k}_{1q} = 0$ and $\widetilde{k}_{1v} = 0$), and the choice $\alpha = \gamma = \frac{1}{2}$ results in a variant of the scheme from [39]. The choice $\gamma = \frac{1}{2}$, $\alpha = 1$ is similar to the midpoint

rule proposed by Moreau in [30, 31]. The time-stepping scheme in [39] detects (behind collisions) other type of events such as stick-slip transitions, take-off transitions, and changes in the active friction components. If the number of such changes is uniformly bounded as $h \to 0$, these transitions could be resolved in the same fashion in which collisions are here dealt with. For simplicity we restrict ourselves to collision detection only.

A collision occurs in the interval $(lh, (l+1)h]$ if $\Phi^{(j)}(q^l) > 0$ and $\Phi^{(j)}(\overline{q}^{l+1}) \leq 0$, where $\overline{q}^{l+1}$ is the position computed by the time-stepping scheme without including j in the active set at time t_{l+1}. The active set at t_{l+1} is taken as

$$\mathcal{A}(t_{l+1}) = \mathcal{A}^{l+1} = \left\{ j \ : \ \Phi^{(j)}(q^l) \leq 0 \right\}.$$

Whenever a collision is encountered, cubic interpolation is used to determine the precollision velocity and the position at which the collision occurs [39]. The detected position q^- and the precollision velocity v^- are used in the compression phase to obtain the new velocity. An MLCP of the same type as (4.7) is solved in the compression phase. More precisely, the solution set of the MLCP modeling the compression phase is $\mathcal{L}(q^-, v^-, 0, 0, 1, 1)$.

Collision detection may result in a nonuniform partition of the simulation interval $[0, T]$. More precisely, a collision may be detected at time t^* such that, for a given time-step h, $t^* \neq lh$ for any integer l. To make the upcoming proofs easier to follow, we enforce a uniform partition of $[0, T]$. When collision is detected at time $t^{*,l+1} \in (lh, (l+1)h]$, the collision is solved, resulting in the collision position $q^{-,l+1}$ and postcollision velocity $v^{+,l+1} \in \mathcal{L}(q^{-,l+1}, v^{-,l+1}, 0, 0, 1, 1)$. Instead of introducing the collision time t^* in the time partition of $[0, T]$ or solving another MLCP in the interval $(t^*, (l+1)h]$, we take

$$t_{l+1} = (l+1)h, \quad q^{l+1} = q^{-,l+1}, \text{ and } v^{l+1} = v^{+,l+1}.$$

Assuming that we do this for every collision and that the first integration step is not a collisional one, we have $t_l = lh$, for all l, and the scheme will keep a fixed time step throughout the integration process.

Note that, while this simplification possibly affects the accuracy of the scheme, our choice essentially represents only a notation convention. Indeed, if the external force k_1 does not depend on time, then the sequence of velocities and positions is identical to the one with the normal convention (where the collision time is considered one of the time points). If the force does depend on time, then the change in its value is only order $O(h)$, since from Assumption **(H7)**, which will be defined shortly, the number of collisions is bounded above and does not affect the convergence proofs.

We extend the numerical solution to time instants different from the ones given by the discrete solution, as follows. The velocity sequence $v^{h,\alpha}(t)$ is defined by

(4.9)
$$v^{h,\alpha}(t) = \begin{cases} v^{l+1,\alpha} & \text{if } t \in (lh, (l+1)h] \quad \text{and no collision in } (lh, (l+1)h], \\ v^{l+1} := v^{+,l+1} & \text{if } t \in (lh, (l+1)h] \quad \text{and collision detected in } (lh, (l+1)h], \\ v^0 & \text{if } t = 0, \end{cases}$$

where $v^{+,l+1}$ denotes the velocity at the end of the compression phase and where

$$v^{l+1,\alpha} = (1-\alpha)v^l + \alpha v^{l+1}. \tag{4.10}$$

The velocity function that uses no weighting is denoted by $v^h(\cdot)$ and is defined in a similar fashion:

(4.11)
$$v^h(t) = \begin{cases} v^{l+1} & \text{if } t \in (lh,(l+1)h] \quad \text{and no collision in } (lh,(l+1)h], \\ v^{l+1} := v^{+,l+1} & \text{if } t \in (lh,(l+1)h] \quad \text{and collision detected in } (lh,(l+1)h]. \end{cases}$$

For the position sequence, we take $q^{h,\alpha}(t)$ to be

(4.12a)
$$q^{h,\alpha}(t) = \begin{cases} \frac{1}{h}\left((t-t_l)q^{(l+1)} + (t_{l+1}-t)q^l\right) & \text{if } t \in (t_l = lh, t_{l+1} = (l+1)h], \\ q^0 & \text{if } t = 0, \end{cases}$$

(4.12b)
$$\text{where } q^{l+1} = \begin{cases} q^{(l)} + hv^{l+1,\alpha} & \text{if } t \in (lh,(l+1)h] \text{ and no collision in } (lh,(l+1)h], \\ q^{-,l+1} & \text{if } t \in (lh,(l+1)h] \text{ and collision detected in } (lh,(l+1)h]. \end{cases}$$

Here q^{l+1} is computed by the position update formula (4.4a), except for collisional instants (that is, a collision occurred in the $(lh,(l+1)h]$ interval), in which case $q^{l+1} := q^{-,l+1}$, where $q^{-,l+1}$ is the estimated collision position. Since the collision time $t^{*,l+1}$ is detected by solving

$$\Phi^{(j)}(\widetilde{q}(t)) = 0,$$

where $\widetilde{q} : [lh,(l+1)h] \to \mathbb{R}^s$ is the cubic interpolant of the data $\widetilde{q}(lh) = q^l$, $\frac{d\widetilde{q}}{dt}(lh) = v^l$, $\widetilde{q}((l+1)h) = \overline{q}^{l+1}$, $\frac{d\widetilde{q}}{dt}((l+1)h) = \overline{v}^{l+1}$ ($\overline{q}^{l+1}$ and $\overline{v}^{l+1}$ are obtained by applying a regular step with $j \notin \mathcal{A}$), and $q^{l+1} = q^{-,l+1} = \widetilde{q}(t^{*,l+1})$, we can guarantee that

$$\Phi^{(j)}\left(q^{l+1}\right) = \Phi^{(j)}\left(q^{-,l+1}\right) \geq -C_c h^2 \tag{4.13}$$

for a fixed constant C_c.

To obtain the convergence results, we use the following assumptions.

(H1) The nonpenetration constraints are twice continuously differentiable, and there exists B_H such that

$$\left\|\nabla_{qq}\Phi^{(j)}(q)\right\| \leq B_H \text{ for all } q \text{ and } j = 1,\ldots,p. \tag{4.14}$$

(H2) The functions $\Theta^{(i)}(q)$, $i = 1,\ldots,m$ are sufficiently smooth functions.

(H3) The generalized mass matrix M is constant, symmetric, and positive definite.

(H4) The total friction cone $\mathcal{FC}(q)$ is uniformly pointed with respect to all configurations q.

(H5) The norm of the external force increases at most linearly with the position and the velocity. That is,

$$\|k_1(t,q,v)\| \leq c_1 + c_2\|q\| + c_3\|v\|. \tag{4.15}$$

Here $k_1(q,v)$ denotes the external and inertial forces.
The Coriolis force is given by a bilinear operator

$$[f_C(v)]_i = \sum_{jk} f_{ijk} v_j v_k \ .$$

This is certainly true if the system is described by Newton–Euler equations in body coordinates [32, section 2.4], where the matrix $F(v)$ of entries

$$[F(v)]_{ij} = \sum_k f_{ijk} v_k$$

is antisymmetric in the sense that

$$u^T F(v)u = 0 \quad \forall u \ .$$

We also assume that the approximations $\tilde{k}_{1q}$ and $\tilde{k}_{1v}$ are bounded. More precisely,

$$\|\widetilde{k}_{1q}\| \le c_4, \ \ \|\widetilde{k}_{1v}\| \le c_5. \tag{4.16}$$

(H6) The contact data given by $\tilde{n}(q)$, $\tilde{D}(q)$ are globally Lipschitz continuous functions.

(H7) The number of collisions solved by the algorithm is uniformly upper bounded as $h \to 0$.

(H8) The external forces $k_1(t, q, v)$ are linear in v, and the approximation $\widetilde{k}_{1v}$ is constant.

Remark 4.1.

- Assumption **(H3)** is satisfied if we use the Newton–Euler formulation in body coordinates; see [32].
- Assumption **(H4)** implies that $\widetilde{\nu}(q)$ has uniform full rank. That is, there exists a constant $\kappa > 0$ such that

$$\sigma_{\min}(\widetilde{\nu}(q)) \ge \kappa \quad \forall\, q,$$

where $\sigma_{\min}(A)$ denotes the smallest singular value of the matrix A.

- Assumption **(H7)** is related to and implied by the assumption of non-Zeno behavior of the system: that the number of switching points in the dynamics (collisions and stick-slip transitions) is finite in any bounded time interval. [12, 42, 13]. While this is not unreasonable to expect when the restitution coefficient is 0 (as we assume here), we do not have a method of guaranteeing it a priori. We note that this is an issue only with trapezoidal-type schemes, and does not need to be assumed in the case of first-order schemes [6].
- The first part of assumption **(H8)** is fairly standard in a stability analysis of the time-stepping scheme. The second part is not needed to prove all of the results. More precisely, uniform boundedness of the numerical velocities as well as a uniform bound on the variation of the numerical velocities can be obtained without this assumption. We note that the **(H8)** assumption is satisfied when external forces include linear damping terms, by far the prevailing type of external velocity-dependent passive force.

The MLCP (4.7) has the same structure as the ones in [4, 39], and therefore the same solvability results can be used to show that the solution set $\mathcal{L}(q^l, v^l, \widetilde{k}, h, \alpha, \gamma)$ is not empty whenever the matrix $\widetilde{M}$ is positive definite. Since the mass matrix M is positive definite, the matrix $F(\cdot)$ is antisymmetric and the approximations used are bounded, it follows from (4.5) that $\widetilde{M}$ will be positive definite for sufficiently small values of h and for any value of the velocity.

Note also that in the presence of stiff forces the use of exact Jacobians k_{1q} and k_{1v} may force the simulation to choose a very small time-step h in order to ensure the positive definiteness of the matrix $\widetilde{M}$. In order to allow the simulation to proceed by using moderate values of the time-step h, appropriate negative semidefinite Jacobian approximations $\widetilde{k}_{1q}$ and $\widetilde{k}_{1v}$ may be used [39].

It is convenient for the proofs of the upcoming sections to separate the terms involving Coriolis forces in (4.4c). To this end, we introduce the following notation:

$$\begin{aligned} \overline{M}^l &= \left(M - \alpha h \widetilde{k}_{1v}^l - \alpha\gamma h^2 \widetilde{k}_{1q}^l\right), \\ \overline{k}^l &= h\left((1-\alpha)k_1\left(t_l, q^l, v^l\right) + \alpha k_1\left(t_{l+1}, q^l, v^l\right)\right) + \alpha h^2 \widetilde{k}_{1q}^l v^l. \end{aligned} \tag{4.17}$$

In terms of the new notation, (4.4c) is rewritten as

$$\overline{M}^l v^{l+1} - \sum_{i=1}^{m} \nu^{(i),l} c_\nu^{(i),l+1} - \sum_{j\in\mathcal{A}} \left(n^{(j),l} c_n^{(j),l+1} + D^{(j),l}\beta^{(j),l+1}\right) = \overline{M} v^l + \overline{k}^l + hF\left(v^l\right) v^{l+1,\alpha}. \tag{4.18}$$

5. Kinetic energy estimates. The following result establishes a uniform bound for the numerical velocities, as $h \to 0$. Since we are dealing only with inelastic collisions and the friction cone is uniformly pointed, the compression phase guarantees that the postcollision kinetic energy will be less than the precollision kinetic energy. Therefore we restrict the proof of the next result to the noncollisional case.

THEOREM 5.1. *If* **(H1)**–**(H8)** *are satisfied and* $\frac{1}{2} \le \alpha \le 1$, *then there is a constant* c *such that*

$$\left(v^l\right)^T M v^l \le \max\left\{\left(v^0\right)^T M v^0,\ \|q^0\| + 1\right\} e^{ct_l},\quad l = 0, 1, \ldots, \lfloor T/h \rfloor$$

for all sufficiently small h.

Proof. Suppose that no collisions are detected in the interval $[t_l, t_{l+1}]$. The new velocity v^{l+1} will be determined by solving the LCP (4.4c)–(4.4f).

Left multiplying (4.18) by $(v^{l+1,\alpha})^T$ and using the fact that $F(v^l)$ is a skew-symmetric matrix, we get that

$$\begin{aligned} \left(v^{l+1,\alpha}\right)^T \overline{M}^l v^{l+1} = \sum_{i=1}^{m} c_\nu^{(i),l+1}\left(\nu^{(i),l}\right)^T v^{l+1,\alpha} + \sum_{j\in\mathcal{A}}\Big\{ c_n^{(j),l+1}\left(n^{(j),l}\right)^T v^{l+1,\alpha} \\ + \left(\beta^{(j),l+1}\right)^T \left(D^{(j),l}\right)^T v^{l+1,\alpha}\Big\} + \left(v^{l+1,\alpha}\right)^T \overline{k}^l + \left(v^{l+1,\alpha}\right)^T \overline{M}^l v^l. \end{aligned} \tag{5.1}$$

Using (4.4c), we deduce that $(\nu^{(i),l})^T v^{l+1,\alpha} = 0$, $i = 1, 2, \ldots, m$. Also, using the contact constraints (4.4d), we obtain $c_n^{(j),l+1}(n^{(j),l})^T v^{l+1,\alpha} = 0$, $j \in \mathcal{A}$. Finally, from the frictional constraints (4.4e) and (4.4f), we get that

$$\begin{aligned} \left(\beta^{(j),l+1}\right)^T \left(D^{(j),l}\right)^T v^{l+1,\alpha} &= -\lambda^{(j),l+1}\left(\beta^{(j),l+1}\right)^T e^{(j)} \\ &= -\mu^{(j)} c_n^{(j),l+1} \lambda^{(j),l+1} \le 0\ \forall j \in \mathcal{A}. \end{aligned}$$

Then (5.1) implies that

$$\left(v^{l+1,\alpha}\right)^T \overline{M}^l v^{l+1} \le \left(v^{l+1,\alpha}\right)^T \overline{M}^l v^l + \left(v^{l+1,\alpha}\right)^T \overline{k}^l. \tag{5.2}$$

By expanding the left- and right-hand sides of the above inequality, we obtain

$$\begin{aligned}\left(v^{l+1,\alpha}\right)^T\overline{M}^l v^{l+1} &= \alpha {v^{l+1}}^T M v^{l+1} + (1-\alpha){v^l}^T M v^{l+1}\\ &\quad - h\alpha^2 {v^{l+1}}^T\left(\widetilde{k}^l_{1v} + \gamma h \widetilde{k}^l_{1q}\right) v^{l+1}\\ &\quad - h\alpha(1-\alpha){v^l}^T\left(\widetilde{k}^l_{1v} + \gamma h \overline{k}^l_{1q}\right) v^{l+1}, \qquad (5.3)\\ \left(v^{l+1,\alpha}\right)^T\left(\overline{M}^l v^l + \overline{k}^l\right) &= (1-\alpha){v^l}^T M v^l + \alpha {v^{l+1}}^T M v^l\\ &\quad - h\alpha^2 {v^{l+1}}^T\left(\widetilde{k}^l_{1v} + (\gamma-1) h \widetilde{k}^l_{1q}\right) v^l\\ &\quad - h\alpha(1-\alpha){v^l}^T\left(\widetilde{k}^l_{1v} + (\gamma-1) h \widetilde{k}^l_{1q}\right) v^l\\ &\quad + h\left(v^{l+1,\alpha}\right)^T\left((1-\alpha)k_1\left(t_{l+1}, q^l, v^l\right) + \alpha k_1\left(t_l, q^l, v^l\right)\right). \qquad (5.4)\end{aligned}$$

Using Assumption (4.15) **(H5)**, we are led to

$$\begin{aligned}\left(v^{l+1,\alpha}\right)^T\overline{M}^l v^{l+1} &\geq \alpha(1 - C_6 h)\|M^{1/2}v^{l+1}\|^2 - C_7 h\|M^{1/2}v^{l+1}\|\|M^{1/2}v^l\|\\ &\quad + (1-\alpha){v^l}^T M v^{l+1}, \qquad (5.5)\\ \left(v^{l+1,\alpha}\right)^T\left(\overline{M}^l v^l + \overline{k}^l\right) &\leq \alpha\left(-1 + \frac{1}{\alpha} + C_8 h\right)\|M^{1/2}v^l\|^2 + C_9 h\|M^{1/2}v^l\|\|M^{1/2}v^{l+1}\|\\ &\quad + C_{10}h(\alpha\|M^{1/2}v^{l+1}\| + (1-\alpha)\|M^{1/2}v^l\|)\\ &\quad \times(\|M^{1/2}v^l\| + \|q^l\| + 1) + \alpha {v^{l+1}}^T M v^l. \qquad (5.6)\end{aligned}$$

Let us denote

$$\rho_l = \left\|M^{1/2}v^l\right\|, \sigma_l = \left\|q^l\right\| + 1.$$

Note that $\alpha \geq \frac{1}{2}$ gives

$$2\alpha - 1 \geq 0 \Rightarrow (2\alpha-1)\left(v^{l+1}\right)^T M v^l \leq (2\alpha-1)\rho_{l+1}\rho_l. \qquad (5.7)$$

Dividing by α both sides of the inequality (5.2) and using the symmetry of the matrix M, the estimates (5.5)–(5.6), the implication (5.7), as well as the notation above, implies that

$$(1 - C_{11}h)\rho_{l+1}^2 \leq \left(-1 + \frac{1}{\alpha} + C_{11}h\right)\rho_l^2 + C_{11}h\sigma_l(\rho_l + \rho_{l+1}) + \left(2 - \frac{1}{\alpha}\right)\rho_l\rho_{l+1} \qquad (5.8)$$

for an appropriately defined constant C_{11}.

Consider now the case for which $\rho_l < \rho_{l+1}$. Dividing by ρ_{l+1} in (5.8) and using that $\rho_l/\rho_{l+1} < 1$ gives

$$(1 - C_{12}h)\rho_{l+1} \leq (1 + C_{12}h)\rho_l + C_{12}h\sigma_l \qquad (5.9)$$

for some constant $C_{12} \geq 0$ and all sufficiently small h. We can rewrite (5.9) in the form

$$\rho_{l+1} \leq (1 + C_{13}h)\rho_l + C_{13}h\sigma_l, \qquad (5.10)$$

with C_{13} appropriately chosen. It is straigthforward to see that, for the remaining case $\rho_{l+1} \leq \rho_l$, inequality (5.10) immediately follows. On the other hand, from (4.4a), we have

$$\|q^{l+1}\| \leq \|q^l\| + \|M^{-1/2}\| \left((1-\gamma)\|M^{1/2}v^l\| + \gamma\|M^{1/2}v^{l+1}\| \right). \tag{5.11}$$

Substituting the overestimate for ρ_{l+1}, (5.10), into (5.11) gives

$$\sigma_{l+1} \leq h\|M^{-1/2}\| \left(1 + \gamma C_{13} h\right) \rho_l + \left(1 + \gamma C_{13} h^2 \|M^{-1/2}\|\right) \sigma_l. \tag{5.12}$$

It follows that there is a constant C_{14} such that

$$\begin{aligned} \rho_{l+1} &\leq (1 + C_{14}h)\rho_l + C_{14}h\sigma_l, \\ \sigma_{l+1} &\leq C_{14}h\rho_l + (1 + C_{14}h)\sigma_l. \end{aligned}$$

By taking $c = 2C_{14}$, we have that, for all sufficiently small h, the following holds:

$$\left\| \begin{bmatrix} \rho_l \\ \sigma_l \end{bmatrix} \right\|_\infty \leq \left\| \begin{bmatrix} 1 + C_{14}h & C_{14}h \\ C_{14}h & 1 + C_{14}h \end{bmatrix} \right\|_\infty^l \left\| \begin{bmatrix} \rho_0 \\ \sigma_0 \end{bmatrix} \right\|_\infty = e^{ct_l} \left\| \begin{bmatrix} \rho_0 \\ \sigma_0 \end{bmatrix} \right\|_\infty,$$

which concludes the proof of our theorem. □

Remark 5.2. *The conclusion of Theorem* 5.1 *implies that both* $v^h(\cdot)$ *and* $v^{h,\alpha}(\cdot)$ *are uniformly bounded on* $[0, T]$, *as* $h \to 0$.

6. MDIs. In the following we use the setup and some of the results of [47]. Formally, we are looking at complementarity systems of the following form.

$$\frac{dq}{dt} = v, \tag{6.1}$$

$$M\frac{dv}{dt} = k(q, v) + \rho, \tag{6.2}$$

$$\Theta^{(i)}(q) = 0, \quad i = 1, 2, \ldots, m_J, \tag{6.3}$$

$$\nu^{(i)}(q)^T v = 0, \quad i = m_J + 1, m_J + 2, \ldots, m_J, \tag{6.4}$$

$$\Phi^{(j)}(q) \geq 0, \quad j = 1, \ldots, p, \tag{6.5}$$

$$\rho(t) = \overline{\rho}(t) + \sum_{j=1}^{p} \rho^{(j)}(t) \in \mathcal{FC}(q), \tag{6.6}$$

$$\overline{\rho}(t) \in \operatorname{span}\left\{ \nu^{(i)}(q(t)) \ : \ i = 1, \ldots, m \right\}, \tag{6.7}$$

$$\|\rho^{(j)}\|\Phi^{(j)}(q) = 0, \quad j = 1, 2, \ldots, p. \tag{6.8}$$

The differences between the above formulation and the one corresponding to the contact-only case consists in a different friction cone being used and the additional bilateral constraints enforced by (6.3). Here $\mathcal{FC}(q)$ is the total friction cone (it includes all constraint forces, bilateral and unilateral) as defined in the previous section. In what follows, we specify what we mean by a solution of (6.1)–(6.8). This is motivated by the fact that a strong solution may not exist in general [46].

In contact mechanics, measures appear as a result of the presence of impulsive forces, while inclusions appear as a result of the presence of Coulomb friction. Because of possible impulsive forces, the velocity of the system is no longer required to be an absolutely continuous function but rather a function of bounded variation.

We are going to replace the forces, as they are understood in general, by vector measures. A vector measure is defined in terms of its action on a continuous function. Assume now that $v : [0,T] \to \mathbb{R}^s$ is a function of bounded variation. That is, the total variation of v, $\bigvee_0^T v(\cdot)$, is finite. Here $\bigvee_0^T v(\cdot)$ is the supremum of the sums $\sum_{i=0}^{N-1} \|v(t_{i+1}) - v(t_i)\|$ over all finite partitions $a = t_0 < t_1 < \cdots < t_{N-1} < t_N = b$. We denote this by $v \in BV([0,T])$. It follows that the measure induced by v can be understood as a linear and continuous operator defined from $C([0,T])$, with values in $\mathbb{R}^s$. More precisely,

$$< dv, \phi >= \int_0^T \phi(t) dv(t), \tag{6.9}$$

where $\phi : [0,T] \to \mathbb{R}$ is continuous. The Riemann–Stieljes integral in (6.9), which exists because of $v(\cdot)$ being of bounded variation, can be approximated by finite Riemann sums:

$$\sum_{i=0}^{N-1} \phi(\tau_i)[v(t_{i+1}) - v(t_i)],$$

where $a = t_0 < \tau_1 < t_1 < \cdots < \tau_{N-1} < t_N = b$. Discontinuities in the velocity may lead to atoms of the measure dv. Therefore dv is not, in general, absolutely continuous with respect to the Lebesgue measure dt, and thus $\frac{dv}{dt}(\cdot)$ cannot be defined, in the usual sense, as a Radon–Nykodim derivative. To give a meaning to inclusions of the form

$$\frac{dv}{dt}(t) \in K(t), \quad \text{for } t \in [0,T], \tag{6.10}$$

we adopt the following definition [47].

DEFINITION 6.1 (MDI). *If $v \in BV([0,T])$ and $K(\cdot)$ is a convex-set valued mapping, we say that* (6.10) *holds if, for all continuous $\phi : [0,T] \to \mathbb{R}$, $\phi \geq 0$ and ϕ not identically zero, we have that*

$$\frac{\int_0^T \phi(t) dv(t)}{\int_0^T \phi(t) dt} \in \bigcup_{\tau : \phi(\tau) \neq 0} K(\tau).$$

DEFINITION 6.2 (weak solution of (6.1)–(6.8)). *We say that $q(t)$, $v(t)$ is a weak solution of* (6.1)–(6.8) *on $[0,T]$ if*

1. *$v(\cdot)$ is a function of bounded variation on $[0,T]$;*
2. *$q(\cdot)$ is an absolutely continuous function that satisfies*

$$q(t) = q(0) + \int_0^t v(\tau) d\tau \quad \text{for } t \in [0,T]; \tag{6.11}$$

3. *the measure $dv(t)$ must satisfy*

$$M \frac{dv}{dt} - k(q,v) \in \mathcal{FC}(q); \tag{6.12}$$

4. *$\Theta^{(i)}(q) = 0$, $i = 1, \ldots, m_J$ and $\nu^{(i)}(q)^T v = 0$ almost everywhere, $i = m_J + 1, m_J + 2, \ldots, m$;*
5. *$\Phi^{(j)}(q) \geq 0$, $j = 1, \ldots, p$.*

7. Uniform bound in variations. For the rest of the paper we consider (γ, α) satisfying

$$\gamma = \alpha \in \left[\frac{1}{2}, 1\right]. \tag{7.1}$$

Since $\gamma = \alpha$ and the number of collisions solved is uniformly upper bounded as $h \to 0$, we have, from (4.4a), that

$$q^{h,\alpha}(t) = q^{h,\alpha}(0) + \int_0^t v^{h,\alpha}(\tau) d\tau.$$

The uniform boundedness of the velocities implies that the sequence $\{q^{h,\alpha}(\cdot)\}$ is equicontinuous and equibounded. Therefore by the Arzela–Ascoli theorem, there exists a uniformly convergent subsequence, which we also denote by $q^{h,\alpha}(\cdot)$, that converges $q^{h,\alpha}(\cdot) \to q(\cdot)$ uniformly in $[0, T]$.

THEOREM 7.1. *$\bigvee_0^T v^{h,\alpha}(\cdot)$ is uniformly bounded as $h \to 0$, and there exists $v^*(\cdot)$ of bounded variation such that $v^{h,\alpha} \to v^*$ pointwise and $dv^{h,\alpha} \to dv^*$ weakly.*

We break the proof in five subsections, along the lines given in [47], with some modifications due to the presence of joint constraints. The main difference consists in the use of the reduced friction cone. In this context, using the regularity of the reduced friction cone, we first obtain a uniform bound on the sums $\sum_l \|\widetilde{c}_n^{l,h}\|$. This is used to obtain a similar result for the other constraint impulses. The proof then follows precisely the lines of [47], first by obtaining a local result for the velocity variation and then by using a compactness argument to extend this result to the entire time interval.

7.1. Use the regularity assumption on the reduced friction cone to obtain a bound on the sums $\sum_l \|\widetilde{c}_n^{l,h}\|$. Let $q(\cdot)$ be the limit of a uniformly convergent subsequence $q^{h,\alpha}(\cdot)$. Let t be a time instant in the interval $(0, T]$. From (3.9) it follows that there exist a unit vector $u_0(t)$ and a scalar $\zeta(t) > 0$ such that, for any $z = \widetilde{\nu}(q(t))\widetilde{c}_\nu + \widetilde{n}(q(t))\widetilde{c}_n + \widetilde{D}(q(t))\widetilde{\beta} \in \mathcal{FC}(q(t))$, we have

$$u_0{}^T(t)\widetilde{\nu}_\perp^T(q(t))z \geq \zeta(t)\|\widetilde{c}_n\|. \tag{7.2}$$

By the closed-graph property of the $\mathcal{FC}(q(t))$, it follows that there is $\eta(t) > 0$ and $h_0 > 0$ such that, for any t'' satisfying $|t'' - t| \leq \eta(t)$ and any $h \leq h_0$, we have

$$u_0{}^T(t)\widetilde{\nu}_\perp^T\left(q^h(t'')\right) z \geq \frac{1}{2}\zeta(t)\|\widetilde{c}_n\| \tag{7.3}$$

for any $z \in \mathcal{FC}(q^h(t''))$. Provided that both lh and $(l+1)h$ lie in the interval $[t - \eta(t), t + \eta(t)]$, the numerical scheme gives

$$\left(M - \alpha h \widetilde{k}_v^{l,h} - \alpha\gamma h^2 \widetilde{k}_q^{l,h}\right)\left(v^{l+1,h} - v^{l,h}\right) = \widetilde{k}^{l,h} + z^{l+1,h}, \tag{7.4}$$

with $z^{l+1,h} \in \widehat{\mathcal{FC}}(q^{(l),h})$. Let us denote

$$\tilde{\nu}_\perp^{l,h} := \widetilde{\nu}_\perp\left(q^{l,h}\right) \text{ and } \tilde{\nu}^{l,h} := \widetilde{\nu}\left(q^{l,h}\right).$$

From the joint constraint enforced at the velocity level, we have $(\widetilde{\nu}^{l,h})^T(\alpha v^{l+1,h} + (1-\alpha)v^{l,h}) = 0$ for all l. By using the orthogonal decomposition, we are led to

$$\begin{aligned} v^{l+1,h} &= \widetilde{\nu}_\perp^{l,h} w^{l+1,h} + \widetilde{\nu}^{l,h} u^{l+1,h}, \\ v^{l,h} &= \widetilde{\nu}_\perp^{l-1,h} w^{l,h} + \widetilde{\nu}^{l-1,h} u^{l,h}. \end{aligned} \tag{7.5}$$

Multiplying both equations in (7.5) on the left by $(\widetilde{\nu}_\perp^{l,h})^T M$ gives

$$\left(\widetilde{\nu}_\perp^{l,h}\right)^T M v^{l+1,h} = \left(\left(\widetilde{\nu}_\perp^{l,h}\right)^T M \widetilde{\nu}_\perp^{l,h}\right) w^{l+1,h} + \left(\widetilde{\nu}_\perp^{l,h}\right)^T M \widetilde{\nu}^{l,h} u^{l+1,h},$$

$$\left(\widetilde{\nu}_\perp^{l,h}\right)^T M v^{l,h} = \left(\left(\widetilde{\nu}_\perp^{l-1,h}\right)^T M \widetilde{\nu}_\perp^{l-1,h}\right) w^{l,h} + \left(\widetilde{\nu}_\perp^{l-1,h}\right)^T M \tilde{\nu}^{l-1,h} u^{l,h} + \mathcal{O}(h). \tag{7.6}$$

For the last equation in (7.6) we have used that $\widetilde{\nu}_\perp^{l,h} = \widetilde{\nu}_\perp^{l-1,h} + \mathcal{O}(h)$, which holds because of the sufficient smoothness of the joint gradients and the uniform boundness of the velocities. Thus, by using $\omega^{i+1,h} := ((\widetilde{\nu}_\perp^{i,h})^T M \widetilde{\nu}_\perp^{i,h}) w^{i+1,h}$ and $\omega^{i+1,h,\perp} := (\widetilde{\nu}_\perp^{i,h})^T M \widetilde{\nu}^{i,h} u^{i+1,h}$, we have, with respect to the new notation,

$$\left(\widetilde{\nu}_\perp^{l,h}\right)^T M v^{l+1,h} = \omega^{l+1,h} + \omega^{l+1,h,\perp} \text{ and } \left(\widetilde{\nu}_\perp^{l,h}\right)^T M v^{l,h} = \omega^{l,h} + \omega^{l,h,\perp} + \mathcal{O}(h). \tag{7.7}$$

We multiply (7.4) on the left by $\widetilde{\nu}_\perp^{l,h}$ to obtain

$$\omega^{l+1,h} - \omega^{l,h} + \omega^{l+1,h,\perp} - \omega^{l,h,\perp} = \widetilde{\nu}_\perp^{l,h} z^{l+1,h} + \mathcal{O}(h), \tag{7.8}$$

where we have used the fact that $\widetilde{k}_q^{l,h}$, $\widetilde{k}_v^{l,h}$, $\frac{1}{h}\widetilde{k}^{l,h}$ are uniformly bounded. It follows from (7.3) that

$$u_0{}^T(t)\left(\omega^{l+1,h} - \omega^{l,h} + \omega^{l+1,h,\perp} - \omega^{l,h,\perp}\right) + \mathcal{O}(h) \geq \frac{1}{2}\zeta(t)\|\widetilde{c}_n^{l+1,h}\|. \tag{7.9}$$

Set $l_{\min} = \lceil (t - \eta(t))/h \rceil$ and $l_{\max} = \lfloor (t + \eta(t))/h \rfloor$. Then

$$\sum_{l_{\min}}^{l_{\max}-1} u_0{}^T(t)\left(\omega^{l+1,h} - \omega^{l,h} + \omega^{l+1,h,\perp} - \omega^{l,h,\perp}\right) + \mathcal{O}(h(l_{\max} - l_{\min})) \geq \frac{1}{2}\zeta(t) \sum_{l_{\min}}^{l_{\max}-1} \left\|\widetilde{c}_n^{l+1,h}\right\|.$$

The sum on the left-hand side in the above inequality telescopes to

$$\begin{aligned}
&\sum_{l_{\min}}^{l_{\max}-1} u_0{}^T(t)\left(\omega^{l+1,h} - \omega^{l,h} + \omega^{l+1,h,\perp} - \omega^{l,h,\perp}\right) \\
&= u_0{}^T(t)\left(\omega^{l_{\max},h} - \omega^{l_{\min},h} + \omega^{l_{\max},h,\perp} - \omega^{l_{\min},h,\perp}\right) \\
&\leq \left\|\omega^{l_{\max},h}\right\| + \left\|\omega^{l_{\min},h}\right\| + \left\|\omega^{l_{\max},h,\perp}\right\| + \left\|\omega^{l_{\min},h,\perp}\right\|.
\end{aligned}$$

Using that $h(l_{\max} - l_{\min}) \leq \eta(t)$ and that $\omega^{l,h}$, $\omega^{l,h,\perp}$ are uniformly bounded (the uniform boundedness of the ω components results from the uniform boundedness of the velocities and the uniform linear independence of the columns of $\tilde{\nu}$) by a constant B_ω, we obtain

$$\sum_{l_{\min}}^{l_{\max}-1} \left\|\widetilde{c}_n^{l+1,h}\right\| \leq \frac{2}{\zeta(t)}\left(2B_\omega + C_1\eta(t)\right) \text{ uniformly as } h \to 0, \tag{7.10}$$

where the constant C_1 above corresponds to the term $\mathcal{O}(h(l_{\max} - l_{\min}))$.

7.2. Show that all of the other constraint impulses are bounded by the normal contact impulses. A bound on the tangential impulses $\tilde{\beta}^{l+1,h}$ is immediately obtained from the conic constraint:

$$\left\|\beta^{(j);l+1,h}\right\|_1 \leq \mu^{(j)} {c_n}^{(j);l+1,h}.$$

Thus, for the combined frictional impulses ${f_T}^{l+1,h} := \widetilde{D}^{l,h}\widetilde{\beta}^{l+1,h}$, we obtain

$$\sum_{l_{\min}}^{l_{\max}-1} \left\|{f_T}^{l+1,h}\right\| \leq C_2 \sum_{l_{\min}}^{l_{\max}-1} \left\|\tilde{c}_n^{l+1,h}\right\| \leq \frac{2C_2}{\zeta(t)}\left(2B_\omega + C_1\eta(t)\right), \tag{7.11}$$

with the last estimate holding uniformly as $h \to 0$. The constant C_2 above depends on the bounds on the frictional directions $d_i^{(j)}(q(\cdot))$, the friction coefficients, and the number of generators used in the polyhedral approximation of the friction cone.

To obtain a bound on $\sum_{l_{\min}}^{l_{\max}-1} \|{f_J}^{l+1,h,\alpha}\| := \sum_{l_{\min}}^{l_{\max}-1} \|\widetilde{\nu}^{l,h}(\alpha\tilde{c}^{l+1,h}+(1-\alpha)\tilde{c}^{l,h})\|$, we go back to

$$\left(M - \frac{h}{2}\widetilde{k}_v^{l,h} - \frac{h^2}{4}\widetilde{k}_q^{l,h}\right)\left(v^{l+1,h} - v^{l,h}\right) = \widetilde{k}^{l,h} + \widetilde{\nu}^{l,h}\widetilde{c}_\nu^{l+1,h} + \widetilde{n}^{l,h}\widetilde{c}_n^{l+1,h} + \widetilde{D}^{l,h}\widetilde{\beta}^{l+1,h},$$

which, together with the uniform bounds we have so far, implies that

$$v^{l+1,h} - v^{l,h} = M^{-1}\widetilde{\nu}^{l,h}\widetilde{c}_\nu^{l+1,h} + M^{-1}\widetilde{n}^{l,h}\widetilde{c}_n^{l+1,h} + M^{-1}\widetilde{D}^{l,h}\widetilde{\beta}^{l+1,h} + \mathcal{O}(h). \tag{7.12}$$

Equation (7.12), together with the uniform boundedness of the velocity sequence and the uniformly pointed friction cone assumption, implies that the impulse multipliers $\widetilde{c}_\nu^{l+1,h}, \widetilde{c}_n^{l+1,h}$, and $\widetilde{\beta}^{l+1,h}$ are bounded uniformly with respect to l. We define, for all indices l for which it makes sense, the following quantities:

$$\begin{aligned}
v^{l+1,h,\alpha} &:= \alpha v^{l+1,h} + (1-\alpha)v^{l,h},\\
\widetilde{c}_\nu^{l+1,h,\alpha} &:= \alpha\widetilde{c}_\nu^{l+1,h} + (1-\alpha)\widetilde{c}_\nu^{l,h},\\
\widetilde{c}_n^{l+1,h,\alpha} &:= \alpha\widetilde{c}_n^{l+1,h} + (1-\alpha)\widetilde{c}_n^{l,h},\\
\widetilde{\beta}^{l+1,h,\alpha} &:= \alpha\widetilde{\beta}^{l+1,h} + (1-\alpha)\widetilde{\beta}^{l,h}.
\end{aligned}$$

We note that the definition of our time-stepping scheme (4.4c) implies that

$$\left(\widetilde{\nu}^{l,h}\right)^T v^{l+1,h,\alpha} = 0, \tag{7.13}$$

and that the triangle inequality implies that

$$\left\|\widetilde{c}_n^{l,h,\alpha}\right\| \leq \alpha\left\|\widetilde{c}_n^{l+1,h}\right\| + (1-\alpha)\left\|\widetilde{c}_n^{l,h}\right\|, \quad \left\|\widetilde{\beta}^{l+1,h,\alpha}\right\| \leq \alpha\left\|\widetilde{\beta}^{l+1,h}\right\| + (1-\alpha)\left\|\widetilde{\beta}^{l,h}\right\|. \tag{7.14}$$

We multiply (7.12) by α and the same equation (7.12), with l replaced by $l-1$, by $(1-\alpha)$, and we add them. We obtain, by the uniform boundedness of the force multipliers and the uniform Lipschitz continuity of $\widetilde{\nu}(q)$, $\widetilde{n}(q)$, and $\widetilde{D}(q)$, that

$$v^{l+1,h,\alpha} - v^{l,h,\alpha} = M^{-1}\widetilde{\nu}^{l,h}\widetilde{c}_\nu^{l+1,h,\alpha} + M^{-1}\widetilde{n}^{l,h}\widetilde{c}_n^{l+1,h,\alpha} + M^{-1}\widetilde{D}^{l,h}\widetilde{\beta}^{l+1,h,\alpha} + \mathcal{O}(h). \tag{7.15}$$

We multiply (7.15) on the left by $(\widetilde{\nu}^{l,h})^T$ and use (7.13) at steps $l+1$ and l together with the fact that $\widetilde{\nu}^{l-1,h} = \widetilde{\nu}^{l,h} + \mathcal{O}(h)$. The result is

$$\mathcal{O}(h) = \left(\widetilde{\nu}^{l,h}\right)^T M^{-1}\widetilde{\nu}^{l,h}\widetilde{c}_\nu^{l+1,h,\alpha} + M^{-1}\widetilde{n}^{l,h}\widetilde{c}_n^{l+1,h,\alpha} + M^{-1}\widetilde{D}^{l,h}\widetilde{\beta}^{l+1,h,\alpha}.$$

Using the fact that the matrix $(\widetilde{\nu}^{l,h})^T M^{-1}\widetilde{\nu}^{l,h}$ is uniformly positive definite in the sense that its eigenvalues are bounded away from 0 uniformly with respect to $q^{l,h}$ as well as (7.14), we obtain a bound for the joint multipliers in terms of the normal contact impulses. More precisely, we have

$$\left\|\widetilde{c}_\nu^{l+1,h,\alpha}\right\| \le C_3\left\|\widetilde{c}_n^{l+1,h,\alpha}\right\| + \mathcal{O}(h) \le C_3\left(\alpha\left\|\widetilde{c}_n^{l+1,h,\alpha}\right\| + \left\|\widetilde{c}_n^{l,h,\alpha}\right\|\right) + \mathcal{O}(h),$$

where the constant C_3 can be chosen independent of l and h. Adding the above inequalities, we obtain

$$\sum_{l_{\min}}^{l_{\max}-1}\left\|{f_J}^{l+1,h,\alpha}\right\| := \sum_{l_{\min}}^{l_{\max}-1}\left\|\widetilde{\nu}^{l,h}\widetilde{c}_\nu^{l+1,h,\alpha}\right\| \le C_4\sum_{l_{\min}}^{l_{\max}-1}\left\|\widetilde{c}_n^{l+1,h,\alpha}\right\| \le \frac{2C_4}{\zeta(t)}\left(2B_\omega + C_1\eta(t)\right) + C_5\eta(t). \tag{7.16}$$

7.3. Obtain the bound for the variation of velocities on $[t-\eta(t), t+\eta(t)]$. As we have done in (7.4), denote by $z^{l+1,h,\alpha}$ the total constraint weighted impulse (combine total joint, normal, and tangential impulses) corresponding to step $l+1$, that is, $z^{l+1,h,\alpha} = \widetilde{\nu}^{l,h}\tilde{c}_\nu^{l+1,h,\alpha} + \widetilde{n}^{l,h}\widetilde{c}_n^{l+1,h,\alpha} + \widetilde{D}^{l,h}\widetilde{\beta}^{l+1,h,\alpha}$. From the derivations above we have that

$$\sum_{l_{\min}}^{l_{\max}-1}\left\|z^{l+1,h,\alpha}\right\| \le \frac{2C_6}{\zeta(t)}\left(2B_\omega + C_1\eta(t)\right) + C_7\eta(t), \tag{7.17}$$

where the constants above can be chosen independent of h. Now from (7.12) we have

$$\left\|v^{l+1,h,\alpha} - v^{l,h,\alpha}\right\| \le \left\|M^{-1}z^{l+1,h,\alpha}\right\| + \mathcal{O}(h),$$

and by adding, we obtain

$$\sum_{l_{\min}}^{l_{\max}-1}\left\|v^{l+1,h,\alpha} - v^{l,h,\alpha}\right\| \le \frac{2C_6}{\zeta(t)}\left(2B_\omega + C_1\eta(t)\right) + C_8\eta(t),$$

which shows that

$$\bigvee_{t-\eta(t)/2}^{t+\eta(t)/2} v^{h,\alpha}(\cdot) \text{ is uniformly bounded as } h \to 0.$$

7.4. Obtain the bound for the variation velocities on the entire time interval. Since $(t-\eta(t)/2, t+\eta(t)/2)$ is a covering of $[0,T]$, there is a finite subcovering

$$\{(t_i - \eta(t_i)/2, t_i + \eta(t_i)/2 \mid i = 1, \ldots, m_T\}.$$

Therefore, by summing the contributions corresponding to this finite set of subintervals, we obtain a uniform bound on $\bigvee_0^T v^{h,\alpha}(\cdot)$ as $h \to 0$. If we use the fact that $v^{h,\alpha}(\cdot)$ has bounded variation, then, by Helly's selection theorem, there exists a subsequence of $v^{h_k,\alpha}(\cdot)$ of $v^{h,\alpha}(\cdot)$ that converges pointwise to $v(\cdot)$ and has bounded variation. Since the limiting velocity $v(t)$ may not be well defined for every $t \in [0,T]$, *we assume without loss of generality,* [47], *that $v(\cdot)$ is right–continuous, i.e., $v(t) = v^+(t)$ for all* $t \in [0,T]$. The corresponding functions $q^{h_k,\alpha}(\cdot)$ converge to the indefinite integral of $v(\cdot)$ by the pointwise convergence theorem for Lebesgue integrals. We assume for simplicity that this is the entire sequence, and therefore $q^{h,\alpha}(\cdot) \to q(\cdot)$ and $v^{h,\alpha}(\cdot) \to v(\cdot)$.

7.5. Weak $*$ convergence. Since $\bigvee_0^T v^{h,\alpha}(\cdot)$ is uniformly bounded as $h \to 0$ and $v^{h,\alpha}(0) = v(0)$ and since $v^{h,\alpha}(\cdot) \to v(\cdot)$ pointwise, it follows that $dv^{h,\alpha} \to dv$ weakly *, that is,

$$\int_0^T \phi(t)^T dv^{h,\alpha}(t) \to \int_0^T \phi(t)^T dv(t)$$

for all continuous functions $\phi(t)$. Therefore, $dv^{h,\alpha}(\cdot) \to dv(\cdot)$ weak $*$ as Borel measures. The proof of Theorem 7.1 is complete.

8. Limits are solutions to the MDI. In this section we will use Assumptions **(H1)**–**(H8)** to prove that the limits are solutions to the rigid body MDI. We note that we cannot expect that the rigid body MDI will have unique solutions even in relatively simple cases [28, 48], a fact which can partly be justified by experimentally observed macroscopic behavior [28]. We therefore study the convergence of the subsequences of the time-stepping scheme to possibly multiple solutions of the rigid body MDI.

Assume (q, v) is a solution of the MDI of Definition (6.2). We write

$$v = \widetilde{\nu}(q)u + \widetilde{\nu}_\perp(q)w. \tag{8.1}$$

Since from the joint constraints the velocity v satisfies $(\widetilde{\nu}(q))^T v = 0$, we must have $u = 0$, which implies that the Borel measure dv (which is well defined since v is a function of bounded variation on $[0, T]$) satisfies $dv = d(\widetilde{\nu}_\perp(q)w)$. We can expand further to obtain, as detailed in Appendix A, that

$$\begin{aligned} dv = d\left(\widetilde{\nu}_\perp(q)w\right) &= \widetilde{\nu}_\perp(q)dw + \frac{\partial}{\partial q}\left(\widetilde{\nu}_\perp(q)w\right) dq \\ &= \widetilde{\nu}_\perp(q)dw + \frac{\partial}{\partial q}\left(\widetilde{\nu}_\perp(q)w\right) v dt = \widetilde{\nu}_\perp(q)dw + \left(\frac{\partial}{\partial q}\left(\widetilde{\nu}_\perp(q)w\right)\right)\widetilde{\nu}_\perp(q)w dt, \end{aligned} \tag{8.2}$$

where for the second last equality we have used (6.11) and for the last one we have used the fact that u from (8.1) is zero. Note that the second term in the last equality of (8.2) is a measure which is absolutely continuous with respect to the Lebesque measure dt. Motivated by the analysis above, we introduce the following definition which gives the MDI on the reduced cone.

Definition 8.1 (reduced weak solution of (6.1)–(6.8)). *We say that $q(t)$, $w(t)$ is a reduced weak solution of* (6.1)–(6.8) *on* $[0, T]$ *if*

1. *$w(\cdot)$ is a function of bounded variation on $[0, T]$;*
2. *$q(\cdot)$ is an absolutely continuous function that satisfies*

$$q(t) = q(0) + \int_0^t \widetilde{\nu}_\perp(q(\tau))w(\tau)d\tau \quad \text{for } t \in [0, T]; \tag{8.3}$$

3. *the measure $dw(t)$ must satisfy*

$$\left(\left(\widetilde{\nu}_\perp(q)\right)^T M\widetilde{\nu}_\perp(q)\right)\frac{dw}{dt} - k_{w,\perp}(t, q, w) \in \mathcal{FC}_r(q), \tag{8.4}$$

where

$$k_{w,\perp}(t, q, w) = \left(\widetilde{\nu}_\perp(q)\right)^T k_w(t, q, w) \tag{8.5}$$

and

$$k_w(t, q, w) = k(t, q, \widetilde{\nu}_\perp(q)w) - M\left(\left(\frac{\partial}{\partial q}\left(\widetilde{\nu}_\perp(q)w\right)\right)\widetilde{\nu}_\perp(q)w\right); \tag{8.6}$$

4. *$\Phi^{(j)}(q) \geq 0$, $j = 1, \ldots, p$.*

LEMMA 8.2. *If (q,w) is a reduced weak solution of* (6.1)–(6.8) *on* $[0,T]$ *in the sense of Definition* 8.1 *and* $\Theta^{(i)}(q(0))=0$, $i=1,2,\dots,m_J$, *then* $(q,v)=(q,\widetilde{\nu}_\perp(q)w)$ *is a weak solution of* (6.1)–(6.8) *on* $[0,T]$ *in the sense of Definition* 6.2.

Proof. By construction $(q,v)=(q,\widetilde{\nu}_\perp(q)w)$ and from conditions 1, 2, and 4 of Definition 8.1, it immediately follows that conditions 1, 2, and 5 of Definition 6.2 are satisfied. To prove that condition 4 of Definition 6.2 is satisfied, we use (8.3) to obtain, for $i=1,2,\dots,m_J$,

$$\begin{aligned}\Theta^{(i)}(q(t)) &= \Theta^{(i)}(q(0)) + \textstyle\int_0^t \left(\nu^{(i)}(q)\right)^T v(\tau)d\tau \\ &= \Theta^{(i)}(q(0)) + \textstyle\int_0^t \left(\left(\nu^{(i)}(q)\right)^T \widetilde{\nu}_\perp(q(\tau))\right) w(\tau)d\tau \\ &= \Theta^{(i)}(q(0)),\end{aligned}$$

where we have used the fact that $(\nu^{(i)})(q)^T\widetilde{\nu}_\perp(q)=0$, for $i=1,2,\dots,m_J$.

Since $\Theta^{(i)}(q(0))=0$, we have $\Theta^{(i)}(q(t))=0$ for all $t\in[0,T]$ and $i=1,2,\dots,m_J$. By a similar rationale we obtain that

$$\nu^{(i)}(q)^T v = \nu^{(i)}(q)^T\widetilde{\nu}_\perp(q)w$$

is zero almost everywhere for $i=m_J+1,m_J+2,\dots,m$, and therefore condition 4 of Definition 6.2 is satisfied.

To prove that (6.12) holds we mainly reverse the derivations in (8.2). That is, if (q,w) satisfies (8.4), then there exist $\overline{z}\in\mathcal{FC}(q)$ and a vector measure $\widetilde{d}_\nu\in\mathbb{R}^m$ such that

$$M\widetilde{\nu}_\perp(q)\frac{dw}{dt} - k_w(t,q,w) = \overline{z} + \widetilde{\nu}\widetilde{d}_\nu.$$

Since $\overline{z}\in\mathcal{FC}(q)$ implies that $\overline{z}+\widetilde{\nu}\widetilde{d}_\nu\in\mathcal{FC}(q)$ for any $\widetilde{d}_\nu\in\mathbb{R}^m$, we can write

$$M\widetilde{\nu}_\perp(q)\frac{dw}{dt} - k_w(t,q,w)\in\mathcal{FC}(q).$$

Using (8.2) together with (8.6) in the above inclusion gives

$$M\frac{dv}{dt} - k(t,q,w)\in\mathcal{FC}(q),$$

and therefore also condition 4 of Definition 6.2 is satisfied. This completes the proof of Lemma 8.2. ∎

8.1. The MDI for the limit. We start by writing (4.18) at step $(l+1)$ and step (l) as follows:

$$\overline{M}^l\left(v^{l+1}-v^l\right) - \left(\overline{k}^l + hF\left(v^l\right)v^{l+1,\alpha}\right) = z^{l+1}, \tag{8.7}$$

$$\overline{M}^{l-1}\left(v^{l}-v^{l-1}\right) - \left(\overline{k}^{l-1} + hF\left(v^{l-1}\right)v^{l,\alpha}\right) = z^{l}, \tag{8.8}$$

where $z^{k+1}\in\widehat{\mathcal{FC}}(q^k)$ and $\overline{M}^l$, $\overline{k}^l$ are given by (4.17).

Since the approximation $\widetilde{k}_{1q}$ is uniformly bounded, we have that

$$\overline{M}^k = M - \alpha h\widetilde{k}_{1v} + \mathcal{O}(h^2).$$

We now multiply (8.7) by α and (8.8) by $(1-\alpha)$ and add them up. We obtain

$$M\left(v^{l+1,\alpha}-v^{l,\alpha}\right)-\left(\alpha\overline{k}^{l}+(1-\alpha)\overline{k}^{l-1}\right)$$

$$-h\alpha\widetilde{k}_{1v}\left(v^{l+1,\alpha}-v^{l,\alpha}\right)-h\left(\alpha F\left(v^{l}\right)v^{l+1,\alpha}+(1-\alpha)F\left(v^{l-1}\right)v^{l,\alpha}\right)+\mathcal{O}(h^{2})=z^{l+1,\alpha}, \tag{8.9}$$

where we have used the fact that by Assumption **(H8)** $\widetilde{k}_{1v}^{l}$ is constant, i.e., $\widetilde{k}_{1v}^{l}=\widetilde{k}_{1v}$ for all l.

Using the fact that $F(\cdot)$ is a linear map, we have that

$$\begin{aligned}(1-\alpha)F\left(v^{l-1}\right)v^{l,\alpha}&=F\left((1-\alpha)v^{l-1}\right)v^{l,\alpha}\\&=F\left(v^{l,\alpha}\right)v^{l,\alpha}-\alpha F\left(v^{l}\right)v^{l,\alpha}.\end{aligned}$$

Then the Coriolis terms in (8.9) become

$$\alpha F\left(v^{l}\right)v^{l+1,\alpha}+(1-\alpha)F\left(v^{l-1}\right)v^{l,\alpha}=F\left(v^{l,\alpha}\right)v^{l,\alpha}+\alpha F\left(v^{l}\right)\left(v^{l+1,\alpha}-v^{l,\alpha}\right). \tag{8.10}$$

Since the sequence $v^{h}(\cdot)$ is uniformly bounded, $v^{h,\alpha}(t+h)\rightarrow v^{+}(t)$ and $v^{h,\alpha}(t)\rightarrow v(t)=v^{+}(t)$ a.e. on $[0,T]$, it follows that

$$F\left(v^{h}(t)\right)\left(v^{h,\alpha}(t+h)-v^{h,\alpha}(t)\right)\rightarrow 0 \text{ as } h\rightarrow 0,$$

for $t\in[0,T]-N$, where N is a set of Lebesque measure zero. The same reasoning applies for the term $\widetilde{k}_{1v}(v^{l+1,\alpha}-v^{l,\alpha})$, giving

$$\widetilde{k}_{1v}\left(v^{h,\alpha}(t+h)-v^{h,\alpha}(t)\right)\rightarrow 0 \text{ pointwise a.e. in } [0,T]. \tag{8.11}$$

Now by using the fact that $k_1(t,q,v)$ is linear in v as well as the fact that $\widetilde{k}_{1q}$ is bounded and $q^{l}=q^{l-1}+\mathcal{O}(h)$, we get that

$$\overline{k}^{h}(t):=\overline{k}^{h}\left(t,q^{h,\alpha}(t),v^{h,\alpha}(t)\right)\rightarrow k_1(t,q(t),v(t)) \text{ pointwise a.e. in } [0,T]. \tag{8.12}$$

Here

$$\begin{aligned}\overline{k}^{h}(t):=\overline{k}^{h}\left(t,q^{h,\alpha}(t),v^{h,\alpha}(t)\right)&=(1-\alpha)k_1\left(t,q^{h,\alpha}(t),v^{h,\alpha}(t)\right)\\&+\alpha k_1\left(t+h,q^{h,\alpha}(t),v^{h,\alpha}(t)\right)+\alpha\widetilde{k}_{1q}\left(t,q^{h,\alpha}(t),v^{h,\alpha}\right)\end{aligned}$$

is the function equivalent to the quantity $\frac{1}{h}\overline{k}^{l}$ from (4.17). Equation (8.12) implies that

$$\alpha\overline{k}^{h}(t)+(1-\alpha)\overline{k}^{h}(t-h)\rightarrow k_1\left(t,q^{h,\alpha}(t),v^{h,\alpha}(t)\right) \text{ pointwise a.e. in } [0,T]. \tag{8.13}$$

Using Assumptions **(H2)** and **(H6)**, we can write $z^{l+1,\alpha}$ in (8.9) as

$$z^{l+1,\alpha}=\overline{z}^{l+1}+\mathcal{O}\left(h\left\|v^{l+1,\alpha}-v^{l,\alpha}\right\|\right),$$

with $\overline{z}^{l+1}\in\widehat{\mathcal{FC}}(q^{l})$. This implies that, for all h sufficiently small,

$$M\frac{dv}{dt}^{h,\alpha}-\widehat{k}^{h}(t)\in\widehat{\mathcal{FC}}\left(q^{h,\alpha}(t)\right)\subset\mathcal{FC}\left(q^{h,\alpha}(t)\right), \tag{8.14}$$

where

$$\widehat{k}^h(t) = \alpha\overline{k}^h(t) + (1-\alpha)\overline{k}^h(t-h) + \left(\widetilde{k}_{1v} + F\left(v^h(t)\right)\right)\left(v^{h,\alpha}(t+h) - v^{h,\alpha}(t)\right) + F\left(v^{h,\alpha}(t)\right)v^{h,\alpha}(t) + \mathcal{O}(h) + \mathcal{O}\left(h\left\|v^{l+1,\alpha} - v^{l,\alpha}\right\|\right).$$

From (8.11)–(8.13) we can easily see that

$$\widehat{k}^h(t) \to k(t, q(t), v(t) = F(v(t))v(t) + k_1(t, q(t), v(t)) \text{ pointwise a.e. in } [0, T].$$

We now write

$$v^{h,\alpha}(t) = \widetilde{\nu}_\perp\left(q^{h,\alpha}(t)\right)w^{h,\alpha}(t) + \widetilde{\nu}\left(q^{h,\alpha}(t)\right)u^{h,\alpha}(t),$$

which gives

$$u^{h,\alpha}(t) = \left(\left(\widetilde{\nu}\left(q^{h,\alpha}(t)\right)\right)^T\widetilde{\nu}\left(q^{h,\alpha}(t)\right)\right)^{-1}\left(\widetilde{\nu}(q^{h,\alpha}(t))\right)^T v^{h,\alpha}(t). \tag{8.15}$$

Using a Taylor expansion together with Assumption **(H2)**, we obtain

$$\begin{aligned}\left(\widetilde{\nu}(q^{h,\alpha}(t))\right)^T v^{h,\alpha}(t) &= \left(\widetilde{\nu}(q^{h,\alpha}(t_l))\right)^T v^{h,\alpha}(t) \\ &\quad + \left(\frac{\partial}{\partial q}\left(\widetilde{\nu}^T(q)v^{h,\alpha}(t)\right)\Big|_{q=q^{h,\alpha}(t_l)}\right)\left(q^{h,\alpha}(t) - q^{h,\alpha}(t_l)\right) \\ &\quad + \mathcal{O}\left(\left\|q^{h,\alpha}(t) - q^{h,\alpha}(t_l)\right\|^2\right).\end{aligned} \tag{8.16}$$

Since the definition of the time-stepping scheme enforces $(\widetilde{\nu}(q^{h,\alpha}(t_l))^T v^{h,\alpha}(t) = 0$ for all $t \in (t_l, t_{l+1}]$ and since $q^{h,\alpha}(t) - q^{h,\alpha}(t_l) = (t - t_l)v^{h,\alpha}(t)$ for all $t \in [t_l, t_{l+1}]$, we have

$$\left(\widetilde{\nu}(q^{h,\alpha}(t))\right)^T v^{h,\alpha}(t) = (t-t_l)\left(\frac{\partial}{\partial q}\left(\widetilde{\nu}^T(q)v^{h,\alpha}(t)\right)\Big|_{q=q^{h,\alpha}(t_l)}\right)v^{h,\alpha}(t) + \mathcal{O}\left(h^2\right). \tag{8.17}$$

Combining (8.15) and (8.17) gives

$$\begin{aligned}&\left(\widetilde{\nu}_\perp(q^{h,\alpha}(t))\right)^T M\left(\widetilde{\nu}(q^{h,\alpha}(t))u^{h,\alpha}(t) - \widetilde{\nu}(q^{h,\alpha}(t-h))u^{h,\alpha}(t-h)\right) \\ &\qquad = \mathcal{O}\left(h\left\|v^{h,\alpha}(t) - v^{h,\alpha}(t-h)\right\|\right) + \mathcal{O}(h^2).\end{aligned} \tag{8.18}$$

Using a similar methodology one also gets

$$\begin{aligned}&\widetilde{\nu}_\perp\left(q^{h,\alpha}(t)\right)w^{h,\alpha}(t) - \widetilde{\nu}_\perp\left(q^{h,\alpha}(t-h)\right)w^{h,\alpha}(t-h) \\ &\quad = \widetilde{\nu}_\perp\left(q^{h,\alpha}(t)\right)\left(w^{h,\alpha}(t) - w^{h,\alpha}(t-h)\right) \\ &\qquad - \left(\frac{\partial}{\partial q}\left(\widetilde{\nu}^T(q)w^{h,\alpha}(t-h)\right)\Big|_{q=q^{h,\alpha}(t)}\right)\left(q^{h,\alpha}(t) - q^{h,\alpha}(t-h)\right) + \mathcal{O}(h^2) \\ &\quad = \widetilde{\nu}_\perp\left(q^{h,\alpha}(t)\right)\left(w^{h,\alpha}(t) - w^{h,\alpha}(t-h)\right) \\ &\qquad + h\left(\frac{\partial}{\partial q}\left(\widetilde{\nu}^T(q)w^{h,\alpha}(t-h)\right)\Big|_{q=q^{h,\alpha}(t)}\right)v^{h,\alpha}(t) + \mathcal{O}(h^2) \\ &\quad = \widetilde{\nu}_\perp\left(q^{h,\alpha}(t)\right)\left(w^{h,\alpha}(t) - w^{h,\alpha}(t-h)\right) \\ &\qquad + h\left(\frac{\partial}{\partial q}\left(\widetilde{\nu}_\perp^T(q)w^{h,\alpha}(t-h)\right)\Big|_{q=q^{h,\alpha}(t)}\right)\widetilde{\nu}_\perp(q^{h,\alpha}(t))w^{h,\alpha}(t) + \mathcal{O}(h^2),\end{aligned} \tag{8.19}$$

where for the last equality, we have used (8.17), which give $h\widetilde{\nu}(q^{h,\alpha}(t))u^{h,\alpha}(t) = \mathcal{O}(h^2)$. We use (8.19) to write

(8.20)
$$\begin{aligned}
&\left(\widetilde{\nu}_\perp\left(q^{h,\alpha}(t)\right)\right)^T M\Big(\widetilde{\nu}_\perp\left(q^{h,\alpha}(t)\right)w^{h,\alpha}(t) - \widetilde{\nu}_\perp\left(q^{h,\alpha}(t-h)\right)w^{h,\alpha}(t-h)\Big)\\
&\quad = \left(\left(\widetilde{\nu}_\perp^T M\widetilde{\nu}_\perp\right)\left(q^{h,\alpha}(t)\right)\right)\left(w^{h,\alpha}(t) - w^{h,\alpha}(t-h)\right)\\
&\qquad + h\left(\widetilde{\nu}_\perp(q^{h,\alpha}(t))\right)^T M\left(\frac{\partial}{\partial q}\left(\widetilde{\nu}_\perp^T(q)w^{h,\alpha}(t-h)\right)\Big|_{q=q^{h,\alpha}(t)}\right)\widetilde{\nu}_\perp(q^{h,\alpha}(t))w^{h,\alpha}(t)\\
&\qquad + \mathcal{O}(h^2).
\end{aligned}$$

Multiplying (8.14) on the left by $(\tilde{\nu}_\perp(q^{h,\alpha}(t))^T$ and using (8.18) and (8.20), we obtain

(8.21)
$$\begin{aligned}
&\left(\tilde{\nu}_\perp\left(q^{h,\alpha}(t)\right)\right)^T M\tilde{\nu}_\perp\left(q^{h,\alpha}(t)\right)\frac{dw}{dt}^{h,\alpha}\\
&\quad - \left(\widehat{k}^h_{w,\perp}(t) + \mathcal{O}\left(\left\|v^{h,\alpha}(t) - v^{h,\alpha}(t-h)\right\|\right) + \mathcal{O}(h)\right) \in \widehat{\mathcal{FC}}_r\left(q^{h,\alpha}(t)\right) \subset \mathcal{FC}_r\left(q^{h,\alpha}(t)\right),
\end{aligned}$$

where

$$\begin{aligned}
\widehat{k}^h_{w,\perp}(t) &= \left(\tilde{\nu}_\perp\left(q^{h,\alpha}(t)\right)\right)^T\\
&\quad\times\left(\widehat{k}^h(t) - M\left(\frac{\partial}{\partial q}\left(\widetilde{\nu}_\perp^T(q)w^{h,\alpha}(t-h)\right)\Big|_{q=q^{h,\alpha}(t)}\right)\widetilde{\nu}_\perp(q^{h,\alpha}(t))w^{h,\alpha}(t)\right).
\end{aligned} \tag{8.22}$$

Given that $q^{h,\alpha}(\cdot) \to q(\cdot)$ uniformly on $[0,T]$, $v^{h,\alpha}(\cdot) \to v(\cdot)$ a.e. on $[0,T]$ and $u^{h,\alpha}(\cdot) \to 0$ on $[0,T]$, we have

$$\left(\tilde{\nu}_\perp\left(q^{h,\alpha}(t)\right)\right)^T M\tilde{\nu}_\perp\left(q^{h,\alpha}\right) \to (\tilde{\nu}_\perp(q(t))^T M\tilde{\nu}_\perp(q(t) \text{ uniformly in } [0,T], \tag{8.23}$$

$$\widehat{k}^h_{w,\perp}(t) \to k_{w,\perp}(t,q(t),w(t)) \text{ pointwise a.e. on } [0,T]. \tag{8.24}$$

To obtain the MDI for the limits (q,w) we invoke [49, Theorem 4], stated in Appendix B, taking into account that (8.21), (8.23), and (8.24) are satisfied. In our case, the requirement of [49, Theorem 4] that $\min\{\|z\| | z \in K(w)\}$ is uniformly bounded is immediately satisfied because $K(w)$ are cones and always contain the zero element. Given also (8.23)–(8.24) as well as the fact that, from Lemma 3.5, $\mathcal{FC}_r(q)$ is uniformly pointed, we can apply the above result directly to obtain that the limits (q,w) satisfy the inclusion (8.4).

To complete this subsection, we note that, for any $t \in [0,T]$, we have that

$$\begin{aligned}
q^{h,\alpha}(t) - q^{h,\alpha}(0) &= \textstyle\int_{t_1}^{t_2} v^{h,\alpha}(\tau)d\tau\\
&= \textstyle\int_{t_1}^{t_2} \widetilde{\nu}_\perp\left(q^{h,\alpha}(\tau)\right)w^{h,\alpha}(\tau) + \widetilde{\nu}(q^{h,\alpha}(\tau))u^{h,\alpha}(\tau)d\tau.
\end{aligned}$$

Since $u^{h,\alpha}(\cdot) \to 0 = u(\cdot)$ (this results from (8.15) and (8.16) together with $(\widetilde{\nu}(q^{h,\alpha}(t_l))^T v^{h,\alpha}(t) = 0)$ as $h \to 0$ pointwise on $[0,T]$ and $q^{h,\alpha}(0) = q(0)$, we obtain that

$$q(t) = q(0) + \int_0^t \widetilde{\nu}_\perp(q(\tau))w(\tau)$$

as required by (8.3).

8.2. Feasibility of the limiting trajectories.

LEMMA 8.3. *Assume that*

$$\Theta^{(i)}\left(q^0\right) = 0, \;\; \Phi^{(j)}\left(q^0\right) \geq 0, \; i = 1, \ldots, m_J, \; j = 1, \ldots, p.$$

Then the limit $q(\cdot)$ is feasible in the sense that

$$\Theta^{(i)}(q(t)) = 0, \; i = 1, \ldots, m_J, \, \Phi^{(j)}(q(t)) \geq 0, \; j = 1, \ldots, p \text{ for all } t \in [0, T],$$

and

$$\nu^{(i)}(q(t)) = 0 \text{ a.e. } t \in [0, T], \; i = m_J + 1, m_J + 2, \ldots, m.$$

Proof. To prove the first part we note that by using the definition of the time-stepping scheme, the fact that the numerical velocities v^l are uniformly bounded as well as the fact that the algorithm solves a finite number of collisions in $[0, T]$, we obtain, for $i = 1, 2, \ldots, m$, that

$$\left\| \left(\nu^{(i)}\left(q^{h,\alpha}(t)\right)\right)^T v^{h,\alpha}(t) \right\| \leq C_1 h, \;\; \text{a.e. in } \; [0, T].$$

Taking the limit as $h \to 0$ gives

$$\left(\nu^{(i)}(q(t))\right)^T v(t) = 0 \;\; \text{a.e. in } \; [0, T], \; i = 1, 2, \ldots, m.$$

The last statement implies that for all $t \in [0, T]$ and all $i = 1, \ldots, m_J$, we have that

$$\begin{aligned} \Theta^{(i)}(q(t)) &= \Theta^{(i)}\left(q^0\right) + \textstyle\int_0^t \left(\nu^{(i)}(q(\tau))\right)^T v(\tau) d\tau \\ &= \Theta^{(i)}\left(q^0\right) \\ &= 0. \end{aligned}$$

To prove the second part, assume first that $\Phi^{(j)}(q^0) = 0$ for some $j \in \{1, \ldots, p\}$. This implies that $j \in \mathcal{A}$, and therefore

$$\left(n^{(j)}\left(q^0\right)\right)^T \left(\alpha v^1 + (1 - \alpha) v^0\right) = 0.$$

Using this we obtain that $\Phi^{(j)}(q^1) = \Phi^{(j)}(q^0) + \mathcal{O}(h^2)$ which implies, by assumption **(H1)**, that

$$\Phi^{(j)}\left(q^1\right) \geq -C_2 h^2,$$

where the constant C_2 depends on the uniform bound for the velocities and the constant B_H in (4.14). Assuming $\Phi^{(j)}(q^1) \leq 0$, i.e., $j \in \mathcal{A}$ at step 2, we can bound (in the same fashion as we did above) the negative part of $\Phi^{(j)}(\cdot)$ at the next step by $\Phi^{(j)}(q^2) \geq -2C_2 h^2$. We can continue this process until the first k for which $\Phi^{(j)}(q^k) \leq 0$ and $\Phi^{(j)}(q^{k+1}) > 0$. We obtain the estimate:

$$\Phi^{(j)}\left(q^l\right) \geq -l C_2 h^2 \geq -(C_2 \cdot T) h, \; l = 0, \ldots, k, \tag{8.25}$$

where to obtain the last inequality we have used that $k \leq l_{\max} = \left\lfloor \frac{T}{h} \right\rfloor$.

If $\Phi^{(j)}(q^0) > 0$, the only way to obtain $\Phi^{(j)}(q^k) < 0$ for some k is to have at least one collision occurring. Assume that this kth time step is the first collisional time step. We can guarantee by the collision-detection algorithm that $\Phi^{(j)}(q^k) \geq -C_3 h^2$ (for a fixed constant C_3), where q^k is the detected position for the collision. When computing the solution at step $(k + 1)$, the index j is a component of the active set. We have two possibilities for step $(k + 1)$:

- The nonpenetration constraint (j) leaves the active set , i.e., $\Phi^j(q^{k+1}) > 0$, in which case we can restart recursively; or,
- The nonpenetration constraint (j) remains in the active set , i.e., $\Phi^{(j)}(q^{k+1}) \leq 0$. In this case, we have $\Phi^{(j)}(q^{(k+1)}) \geq -(C_3 + C_2)h^2$. Continuing like this until step $(k + r + 1)$ where either take-off occurs or $(k + r + 1) \geq \frac{T}{h}$, we obtain the estimate

$$\Phi^{(j)}\left(q^{k+l}\right) \geq -(C_3 + lC_2)h^2 \geq -C_4 h,\ l = 0, \ldots, r. \tag{8.26}$$

Since the number of changes in the active set is uniformly upper bounded as $h \to 0$, we can separate the two cases above and combine (8.25)–(8.26) to obtain

$$\Phi^{(j)}\left(q^{l}\right) \geq -Ch,\ 0 \leq l \leq \left\lfloor \frac{T}{h} \right\rfloor.$$

It follows that $\Phi^{(j)}(q^{h,\alpha}(t)) \geq -Ch$ for h sufficiently small and all $t \in [0, T]$. Taking the limit as $h \to 0$ we obtain $\Phi^{(j)}(q(t)) \geq 0$, $t \in [0, T]$. ∎

We summarize the analysis above in the following result.

THEOREM 8.4. *Assume that $\gamma = \alpha \in [\frac{1}{2}, 1]$, and conditions* **(H1)**–**(H8)** *hold. Then there exists a subsequence $h_k \to 0$ such that*

1. *$q^{h_k,\alpha}(\cdot) \to q(\cdot)$ uniformly;*
2. *$v^{h_k,\alpha}(\cdot) \to v(\cdot)$ pointwise a.e;*
3. *$dv^{h_k,\alpha}(\cdot) \to dv(\cdot)$ weak * as Borel measures in $[0, T]$, and every such subsequence converges to a solution $(q(\cdot), v(\cdot))$ of the MDI (6.11)–(6.12).*

Therefore, $q(t), v(t)$ is a weak solution of our model.

9. Examples. In this section we present two numerical examples that illustrate some of the theoretical points made in this work. We mention that an example, involving a double pendulum colliding with a wall, of our scheme converging in the case that involves joints and collisions (and thus, discontinuities in the velocity solution) was already presented in [39].

9.1. A simple joint example. As an introductory example, consider the dynamics of the system $\ddot{q} = 0$, subject to the joint constraint $q = 0$, to which we apply the scheme (4.4) with parameters $\alpha = \frac{1}{2}$ and $\gamma = \frac{1}{2}$. If the initial conditions are $q = 0$ and $\dot{q} = 0$, then the exact solution satisfies $q(t) = 0$.

To model the effect of errors on initial conditions, we start with $q = 0$, $\dot{q} = \epsilon$. Our scheme produces $q^{l,\alpha} = 0$ and $v^l = (-1)^l \epsilon$. The total variation of the velocity for the time interval T is $\frac{2\epsilon T}{h}$, where h is the time step. Therefore, no matter how small the initial error, the total variation is unbounded, and the resulting velocity function does not converge pointwise as $h \to 0$. On the other hand, we can immediately see that $v^{l,\alpha} = 0$, and that the velocity function defined in our main result has bounded variation and is convergent pointwise. This also validates the fact that our bounded variation for $v^{l,\alpha}$ result holds irrespective of the initial error in constraint satisfaction, and that the same result cannot be proved for v^l (though for the case with exact satisfaction of the initial constraints we could neither prove nor disprove bounded variation of the velocity sequence).

Of course, this difficulty will disappear if we make $\epsilon = 0$. But on one hand, in practical examples exact satisfaction of the constraints is difficult to guarantee. And on the other hand, this example is indicative of the fact that $v^{l,\alpha}$ has a more stable behavior than v^l.

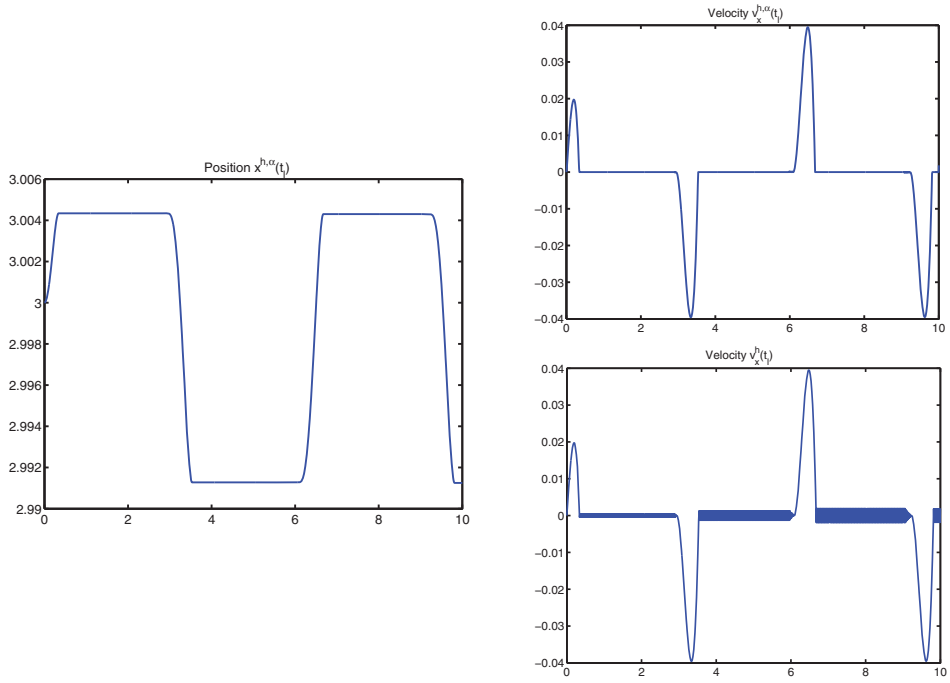


FIG. 9.1. *Numerical position and velocities for $T = 10$ (s), $\alpha = \gamma = 1/2$, and $h = 0.01$. The plot on the left shows positions $q_1^{h,\alpha}(t_l)$, while the two plots on the right show the velocity sequences $v_1^{h,\alpha}(t_l)$ at the top and $v_1^{h}(t_l)$ at the bottom.*

9.2. An example with stick-slip behavior. We want to further motivate our choice for the velocity sequence by looking at a very simple example, [39], with stick-slip behavior. In that example, a block of mass $m = 1$ is subjected to an exterior force $k(t) = 8\cos(t)$ and is sliding on a flat table with friction coefficient $\mu = 0.8$. The initial position of the block is $q_0 = (3, 0)^T$, and the initial velocity is $v_0 = (0, 0)^T$. The gravity $G = (0, -mg)^T$ is calculated, with $g = 9.81$. We compare the weighted numerical velocity sequence $v^{h,\alpha}(t)$ to the sequence $v^h(t)$ for $\alpha = \gamma = \frac{1}{2}$. The positions $q^{h,\alpha}(t_l)$ and velocities $v^h(t_l), v^{h,\alpha}(t_l)$, with $\alpha = \gamma = \frac{1}{2}$ are shown in Figure 9.1, and they indicate a typical stick-slip behavior. We note that the numerical velocities exhibit a quite different behavior, in line with our observations from the preceding sections. We see that, starting with the onset of sticking, the velocity sequence v^l exhibits oscillations that are not present in the sequence $v^{l,\alpha}$, which has the value 0 during the sticking phase. As opposed to the previous example, we do not obtain unbounded variation, though the total variation of the two velocity solutions is different. Nonetheless, the example illustrates the difficulty in obtaining a good behavior of the total variation of the velocity solution v^l, as opposed to $v^{l,\alpha}$, and justifies our choice of the latter for our convergence result.

10. Conclusions. In this work, we have defined a convergence framework for a class of time-stepping schemes for multirigid-body dynamics with joints, contact, and friction. In our framework the numerical solution is shown to converge to the solution of an MDI. *The novelty of our approach resides in the fact that convergence in an MDI sense of an LCP time-stepping scheme is proved, for the first time, for the case*

that involves joint constraints as well. We note that such a proof does not directly follow from representing a joint constraint (an equality constraint) as two opposite inequality constraints (contact constraints) and applying previous convergence results [46, 47], because the resulting system cannot possibly have a pointed friction cone, since any action can be realized with infinite multipliers by cancellation. The situation is analogous to the loss of the Mangasarian–Fromovitz constraint qualification in nonlinear programming when one equality constraint is represented as two inequality constraints [33]. In this work, results for cases involving joints are proved by defining the MDI with respect to an appropriately defined reduced friction cone.

The convergence framework presented here accommodates time-stepping methods based on semiexplicit Euler methods [4, 46] as well as various instances of the trapezoidal method that have been shown to have second-order convergence under certain assumptions [39]. An important step in the convergence proof, following the technique developed in [47], is the proof of the bounded variation of the discrete velocity sequence. We show that, although this may not hold for most trapezoidal-like methods for the natural discrete velocity sequence ($v(t) = v^{(l+1)}$, for $t \in (t_l, t_{l+1}]$, which is the one used in the seminal work [47]), it does hold for the modified velocity sequence $v(t) = \alpha v^{(l+1)} + (1-\alpha)v^{(l)}$ for $t \in (t_l, t_{l+1}]$, where α is the parameter used in the enforcement of the linearization of the geometrical constraints (contact and joint constraints). This point is reinforced by numerical examples.

Appendix A. The details in the derivation of (8.2). In this section we present the details of obtaining (8.2). The main result that we use can be found in [29, p. 9], and it is listed below.

LEMMA A.1 (see [29, p. 9]). *If u_1, $u_2 \in \mathrm{BV}([0,T],\mathbb{R}^k)$, then $d(u_1^T u_2)$ is a real Borel measure on $[0,T]$, which we write $d(u_1^T u_2) \in \mathcal{B}([0,T],\mathbb{R})$ and*

$$\begin{aligned} d\left(u_1^T u_2\right) &= \left(u_2^-\right)^T du_1 + \left(u_1^+\right)^T du_2 \\ &= \left(u_2^+\right)^T du_1 + \left(u_1^-\right)^T du_2, \end{aligned} \tag{A.1}$$

where for a function $f \in \mathrm{BV}([0,T],\mathbb{R}^k)$, f^+ (f^-) denotes the right-limit (left-limit) of f. More precisely $f^+(t) = \lim_{s\to t,\ s>t} f(s)$ ($f^-(t) = \lim_{s\to t,\ s<t} f(s)$), with the convention that if t is the right (left) endpoint of $[0,T]$, we take $f^+(t) = f(t)$ ($f^-(t) = f(t)$). Note that since f is of bounded variation, these limits exist for all t in $[0,T]$.

Proving (8.2). We recall that $q : [0,T] \to \mathbb{R}^s$ is a Lipschitz continuous function, $v = \widetilde{\nu}_\perp(q)w \in \mathrm{BV}([0,T],\mathbb{R}^s)$, and $\widetilde{\nu}_\perp : \mathbb{R}^s \to \mathbb{R}^{s\times(s-m)}$ is sufficiently smooth. We further assume that $v(\cdot) = v^+(\cdot)$ (Note that since $q(\cdot)$ is continuous and $\widetilde{\nu}_\perp(\cdot)$ is uniformly full column rank, this also implies that $w(\cdot)$ is equal to its right limit). To prove (8.2) the steps itemized below are followed.

- **Chain rule:** $d(\widetilde{\nu}_\perp(q)w) = \widetilde{\nu}_\perp(q)dw + \frac{\partial}{\partial q}\left(\widetilde{\nu}_\perp(q)w\right) dq$.
  For every $i \in \{1,\dots,s\}$ we apply (A.1), with

  $$u_1 = \left(\widetilde{\nu}_\perp(q)\right)_i \text{ and } u_2 = w.$$

  Here if A is a given matrix, A_i denotes its ith row written in column format. Since $q(\cdot)$ is Lipschitz continuous and $\widetilde{\nu}_\perp(\cdot)$ is sufficiently smooth, it follows that $u_1 \in \mathrm{BV}([0,T],\mathbb{R}^{s-m})$ and $u_1^+(t) = u_1^-(t) = u_1(t)$ for all $t \in [0,T]$. We also have $u_2 = w \in \mathrm{BV}([0,T],\mathbb{R}^{s-m})$. Using (A.1) we obtain

  $$(dv)_i = \left(d\left(\widetilde{\nu}_\perp(q)\right)_i\right)^T w + \left(\left(\widetilde{\nu}_\perp(q)\right)_i\right)^T dw, \tag{A.2}$$

where we have used the continuity of u_1 and right continuity of u_2. Since $\widetilde{\nu}_\perp(\cdot)$ is sufficiently smooth, we can write

$$d\left(\widetilde{\nu}_\perp(q)\right)_i = \left(\frac{\partial}{\partial q}\left(\left(\widetilde{\nu}_\perp\right)_i\right)(q)\right) dq, \tag{A.3}$$

where the $(s-m)\times s$ matrix in the right-hand side is the Jacobian of $(\widetilde{\nu}_\perp(q))_i$. Using (A.3) in (A.2) for all i gives the desired result, i.e.,

$$d\left(\widetilde{\nu}_\perp(q)w\right) = \widetilde{\nu}_\perp(q)dw + \frac{\partial}{\partial q}\left(\widetilde{\nu}_\perp(q)w\right) dq. \tag{A.4}$$

- **The differential vector measure induced by q:** $dq = vdt$.
  Since

$$q(t) = q(0) + \int_0^t v(\tau)d\tau,$$

for all $t \in [0,T]$ and v is bounded on $[0,T]$, it follows that dq is absolutely continuous w.r.t. the Lebesque measure dt, and the Radon–Nicodym derivative is

$$v = \frac{dq}{dt}.$$

Therefore we may write $dq = vdt$. Note that the Radon–Nicodym derivative (w.r.t. the Lebesque measure) is uniquely determined up to a set of (Lebesque) measure 0.

Appendix B. Theorem 4, [49]. *Suppose that $q_{\hat{n}}(\cdot)$ are continuous, $v_{\hat{n}}(\cdot)$ have uniformly bounded variation, and $k_{\hat{n}}(\cdot)$ are uniformly bounded, all on $[0,T]$, and $q_{\hat{n}}(\cdot) \to q(\cdot)$ uniformly, $v_{\hat{n}}(\cdot) \to v(\cdot)$ pointwise a.e., and $k_{\hat{n}}(\cdot) \to k(\cdot)$ pointwise a.e. Suppose also that $K : \mathbb{R}^n \Rightarrow \mathcal{C}(\mathbb{R}^n)$ has closed graph, $\min\{\|z\| | z \in K(w)\}$ is uniformly bounded, and $K(w)$ is pointed for all $w \in \mathbb{R}^n$. Then if*

$$\frac{dv_{\hat{n}}}{dt}(t) \in K(q_{\hat{n}}(t)) - k_{\hat{n}}(t)$$

for all $\hat{n}$, the limit satisfies

$$\frac{dv}{dt}(t) \in K(q(t)) - k(t).$$

Here $\mathcal{C}(\mathbb{R}^n)$ denotes all of the closed and convex subsets of $\mathbb{R}^n$.

Acknowledgments. We are grateful to two anonymous referees for suggestions that have substantially improved the paper and, in particular, for the suggestion of including nonholonomic constraints. We are grateful to Jeff Trinkle and Vijay Kumar for support and creative discussions.

REFERENCES

[1] M. Anitescu, J. F. Cremer, and F. A. Potra, *Formulating 3d contact dynamics problems*, Mech. Structures Mach., 24 (1996), pp. 405–437.

[2] M. Anitescu and G. D. Hart, *A constraint-stabilized time-stepping approach for rigid multibody dynamics with joints, contact and friction*, Internat. J. Numer. Methods Engrg., 60 (2004), pp. 2335–2371.

[3] M. Anitescu, F. A. Potra, and D. Stewart, *Time-stepping for three-dimensional rigid-body dynamics*, Comput. Methods Appl. Mech. Engrg., 177 (1999), pp. 183–197.

[4] M. Anitescu and F. A. Potra, *Formulating dynamic multi-rigid-body contact problems with friction as solvable linear complementarity problems*, Nonlinear Dynam., 14 (1997), pp. 231–247.

[5] M. Anitescu and F. A. Potra, *A time-stepping method for stiff multibody dynamics with contact and friction*, Internat. J. Numer. Methods Engrg., 55 (2002), pp. 753–784.

[6] M. Anitescu, *Optimization-based simulation of nonsmooth dynamics*, Math. Program., 105 (2006), pp. 113–143.

[7] U. M. Ascher and L. R. Petzold, *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*, SIAM, Philadelphia, 1998.

[8] D. Baraff, *Issues in computing contact forces for non-penetrating rigid bodies*, Algorithmica, 10 (1993), pp. 292–352.

[9] G. Bartels, T. Unger, D. Kadau, D. E. Wolf, and J. Kertsz, *The effect of contact torques on porosity of cohesive powders*, Granular Matter, 7 (2005), pp. 139–143.

[10] B. Brogliato, *Nonsmooth impact mechanics: Models, Dynamics and Control*, Springer-Verlag, London, 1996.

[11] B. Brogliatto, A. ten Dam, L. Paoli, F. Genot, and M. Abadie, *Numerical simulation of finite dimensional multibody not smooth mechanical systems*, Appl. Mech. Rev., 55 (2002), pp. 107–150.

[12] K. Camlibel and J. Schumacher, *On the Zeno behavior of linear complementarity systems*, in Proceedings of the 40th IEEE Conference on Decision and Control, 2001, pp. 346–351.

[13] M. K. Camlibel, J.-S. Pang, and J. Shen, *Conewise linear systems: Non-Zenoness and observability*, SIAM J. Control Optim., 45 (2006), pp. 1769–1800.

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

[15] B. R. Donald and D. K. Pai, *On the motion of compliantly connected rigid bodies in contact: A system for analyzing designs for assembly*, in Proceedings of the Conference on Robotics and Automation, IEEE, 1990, pp. 1756–1762.

[16] F. Camborde, C. Mariotti, and F. V. Donze, *Numerical study of rock and concrete behavior by discrete element modeling*, Comput. Geotechnics, 27 (2000), pp. 225–247.

[17] R. Featherstone, *Robot Dynamics Algorithms*, Kluwer Academic Publishers, Boston, 1987.

[18] C. Glocker and F. Pfeiffer, *An LCP-approach for multibody systems with planar friction*, in Proceedings of the CMIS 92 Contact Mechanics International Symposium, Lausanne, Switzerland, 1992, pp. 13–30.

[19] C. Glocker and F. Pfeiffer, *Multiple impacts with friction in rigid multi-body systems*, Nonlinear Dynam., 7 (1995), pp. 471–497.

[20] E. J. Gottlieb, M. J. McDonald, F. J. Oppel, J. B. Rigdon, and P. G. Xavier, *The UMBRA simulation framework as applied to building HLA federates*, in Proceedings of the 2002 Winter Simulation Conference, San Diego, California, 2002, pp. 981–989.

[21] H. D. Gougar, *Advanced Core Design and Fuel Management for Pebble-Bed Reactors*, Ph.D. thesis, Department of Nuclear Engineering, Pennsylvania State University, Dayton, PA, 2004.

[22] D. Goulding, J.-P. Hansen, and S. Melchionna, *Size selectivity of narrow pores*, Phys. Rev. Lett., 85 (2000), pp. 1132–1135.

[23] E. J. Haug, *Computer-Aided Kinematics and Dynamics of Mechanical Systems*, Prenctice-Hall, Englewood Cliffs, NJ, 1989.

[24] D. Helbing, I. Farkas, and T. Vicsek, *Simulating dynamical features of escape panic*, Nature, 407 (2000), pp. 487–490.

[25] M. Jean, *The nonsmooth contact dynamics method*, Comput. Methods Appl. Mech. Engrg., 177 (1999), pp. 235–257.

[26] W. Jorgensen, J. Chandrasekhar, J. Madura, R. Impey, and M. Klein, *Comparison of simple potential functions for simulating liquid water*, J. Chem. Phys., 79 (1983), pp. 926–935.

[27] D. Kadau, G. Bartels, L. Brendel, and D. Wolf, *Pore stabilization in cohesive granular systems*, Phase Transitions: A Multinational Journal, 76 (2003), pp. 315–331.

[28] R. I. Leine, B. Brogliato, and H. Nijmeijer, *Periodic motion and bifurcations induced by the painleve paradox*, Eur. J. Mech. A Solids, 21 (2002), pp. 869–896.
[29] M. D. P. Marques, *Differential inclusions in nonsmooth mechanical problems: Shocks and dry friction*, Progress in Nonlinear Differential Equations and Their Applications 9, Birkhäuser-Verlag, Basel, 1993.
[30] J. J. Moreau, *Unilateral constraints and dry friction in finite freedom dynamics*, in Nonsmooth Mechanics and Applications, CISM Courses and Lectures 302, J. Moreau and P. Panagiotopoulos, eds., Springer-Verlag, New York, 1988, pp. 1–82.
[31] J. J. Moreau, *Numerical aspects of the sweeping process*, Comput. Methods Appl. Mech. Engrg., 177 (1999), pp. 329–349.
[32] R. M. Murray, Z. Li, and S. S. Sastry, *A Mathematical Introduction to Robotic Manipulation*, CRC Press, Boca Raton, FL, 1993.
[33] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer-Verlag, Berlin, 1999.
[34] P. Painlevé, *Sur le lois du frottement de glissemment*, Comptes Rendus Acad. Sci. Paris, 121 (1895), pp. 112–115.
[35] J.-S. Pang, V. Kumar, and P. Song, *Convergence of time-stepping method for initial and boundary-value frictional compliant contact problems*, SIAM J. Numer. Anal., 43 (2005), pp. 2200–2226.
[36] J.-S. Pang and D. Stewart, *Differential variational inequalities*, Math. Program., 113 (2008), pp. 345–424.
[37] J.-S. Pang and D. Stewart, *Solution dependence on initial conditions in differential variational inequalities*, Set-Valued Anal., Math. Program., to appear.
[38] J.-S. Pang and J. C. Trinkle, *Complementarity formulations and existence of solutions of dynamic multi-rigid-body contact problems with Coulomb friction*, Math. Program., 73 (1996), pp. 199–226.
[39] F. A. Potra, M. Anitescu, B. Gavrea, and J. Trinkle, *A linearly implicit trapezoidal method for integrating stiff multibody dynamics with contact and friction*, Internat. J. Numer. Methods Engrg., 66 (2006), pp. 1079–1124.
[40] F. Radjai, M. Jean, J.-J. Moreau, and S. Roux, *Force distributions in dense two-dimensional granular systems*, Phys. Rev. Lett., 77 (1996), pp. 274–277.
[41] M. Saraniti, S. Aboud, and R. Eisenberg, *The simulation of ionic charge transport in biological ion channels: An introduction to numerical methods*, Rev. Comput. Chem., 22 (2006), pp. 229–293.
[42] J. Shen and J.-S. Pang, *Linear complementarity systems: Zeno states*, SIAM J. Control Optim., 44 (2005), pp. 1040–1066.
[43] P. Song, P. Kraus, V. Kumar, and P. Dupont, *Analysis of rigid-body dynamic models for simulation of systems with frictional contacts*, J. Appl. Mech., 68 (2001), pp. 118–128.
[44] P. Song, J.-S. Pang, and V. Kumar, *A semi-implicit time-stepping model for frictional compliant contact problems*, Internat. J. Numer. Methods Engrg., 60 (2004), pp. 267–279.
[45] D. E. Stewart and J. Trinkle, *An implicit time-stepping scheme for rigid body dynamics with inelastic collisions and Coulomb friction*, Internat. J. Numer. Methods Engrg., 39 (1996), pp. 281–287.
[46] D. E. Stewart, *Existence of solutions to rigid body dynamics and the Painlevé paradoxes*, C. R. Math. Acad. Sci. Paris, 325 (1997), pp. 689–693.
[47] D. E. Stewart, *Convergence of a time-stepping scheme for rigid body dynamics and resolution of Painlevé's problems*, Arch. Ration. Mech. Anal., 145 (1998), pp. 215–260.
[48] D. E. Stewart, *Rigid-body dynamics with friction and impact*, SIAM Rev., 42 (2000), pp. 3–39.
[49] D. E. Stewart, *Reformulations of measure differential inclusions and their closed graph property*, J. Differential Equations, 175 (2001), pp. 108–129.
[50] J. Trinkle, J.-S. Pang, S. Sudarsky, and G. Lo, *On dynamic multi-rigid-body contact problems with Coulomb friction*, Zeithschrift fur Angewandte Mathematik und Mechanik, 77 (1997), pp. 267–279.
[51] J. Tzitzouris, *Numerical Resolution of Frictional Multi-Rigid-Body Systems via Fully Implicit Time-Stepping and Nonlinear Complementarity*, Ph.D. thesis, Department of Mathematical Sciences, Johns Hopkins University, Baltimore, MD, 2001.
[52] A. Wallqvist and R. Mountain, *Molecular models of water: Derivation and description*, in Reviews in Computational Chemistry 13, John Wiley and Sons, New York, 1999, pp. 183–247.

SIAM J. OPTIM.
Vol. 19, No. 3, pp. 1003–1027

AN ACTIVE-SET NEWTON METHOD FOR MATHEMATICAL PROGRAMS WITH COMPLEMENTARITY CONSTRAINTS*

A. F. IZMAILOV† AND M. V. SOLODOV‡

Abstract. For a mathematical program with complementarity constraints (MPCC), we propose an active-set Newton method, which has the property of local quadratic convergence under the MPCC linear independence constraint qualification (MPCC-LICQ) and the standard second-order sufficient condition (SOSC) for optimality. Under MPCC-LICQ, this SOSC is equivalent to the piecewise SOSC on branches of MPCC, which is weaker than the special MPCC-SOSC often employed in the literature. The piecewise SOSC is also more natural than MPCC-SOSC because, unlike the latter, it has an appropriate second-order necessary condition as its counterpart. In particular, our assumptions for local quadratic convergence are weaker than those required by standard SQP when applied to MPCC and are equivalent to assumptions required by piecewise SQP for MPCC. Moreover, each iteration of our method consists of solving a linear system of equations instead of a quadratic program. Some globalization issues of the local scheme are also discussed, and illustrative examples and numerical experiments are presented.

Key words. mathematical program with complementarity constraints, active set method, Newton method, SQP, second-order sufficiency, quadratic convergence

AMS subject classifications. 90C30, 90C33, 49M37, 65K10

DOI. 10.1137/070690882

1. Introduction.

We consider a *mathematical program with complementarity constraints* (MPCC)

$$\min\ f(x) \quad \text{s.t.} \quad G(x) \geq 0,\ H(x) \geq 0,\ \langle G(x), H(x)\rangle \leq 0, \tag{1.1}$$

where $f : \mathbf{R}^n \to \mathbf{R}$ is a smooth function and G, $H : \mathbf{R}^n \to \mathbf{R}^m$ are smooth mappings (twice differentiable and possessing Lipschitzian second derivatives in a neighborhood of the solution of interest). We note that "usual" equality and inequality constraints can be added to our problem setting without any substantial difficulties. We shall consider the case when the problem has only complementarity constraints for the sake of simplicity. Note also that the last constraint in (1.1) could be written as an equality, which is more standard in the complementarity literature. However, it is known that in the context of MPCC, there are good numerical reasons to use the inequality formulation for this constraint. Also, this makes the associated set of Lagrange multipliers smaller, which has both numerical and theoretical advantages. MPCC is perhaps one of the most important instances of a *mathematical program*

*Received by the editors May 8, 2007; accepted for publication (in revised form) May 15, 2008; published electronically October 22, 2008.

http://www.siam.org/journals/siopt/19-3/69088.html.

†Moscow State University, Faculty of Computational Mathematics and Cybernetics, Department of Operations Research, Leninskiye Gori, GSP-2, 119899 Moscow, Russia (izmaf@ccas.ru). Research of this author is supported by the Russian Foundation for Basic Research grants 06-01-00530, 07-01-00270, 07-01-00416, and 07-01-90102-Mongm, and by Russian Federation President's Grant NS-693.2008.1 for the support of leading scientific schools. The author also thanks IMPA, where he was a visitor when this work was initiated.

‡IMPA, Estrada Dona Castorina 110, Jardim Botânico, Rio de Janeiro, RJ 22460-320, Brazil (solodov@impa.br). Research of this author is supported in part by CNPq grants 301508/2005-4, 490200/2005-2, and 550317/2005-8, by PRONEX–Optimization, and by FAPERJ grant E-26/151. 942/2004.

with equilibrium constraints, which has recently attracted considerable attention in the optimization literature; see [15, 16].

In order to explain the contribution of this work, some preliminaries from MPCC theory will be needed. To this end, let

$$L(x,\,\lambda) = f(x) - \langle \lambda_G,\, G(x)\rangle - \langle \lambda_H,\, H(x)\rangle + \lambda_0 \langle G(x),\, H(x)\rangle$$

be the standard Lagrangian of problem (1.1), where $x \in \mathbf{R}^n$ and $\lambda = (\lambda_G,\, \lambda_H,\, \lambda_0) \in \mathbf{R}^m \times \mathbf{R}^m \times \mathbf{R}$. As for any other mathematical program (MP), stationary points of (1.1) and the associated Lagrange multipliers are characterized by the Karush–Kuhn–Tucker (KKT) optimality system:

$$\begin{gathered} \frac{\partial L}{\partial x}(x,\,\lambda) = 0, \quad \lambda_0 \geq 0, \quad \langle G(x),\, H(x)\rangle \leq 0, \\ \lambda_G \geq 0,\ G(x) \geq 0,\ \langle \lambda_G,\, G(x)\rangle = 0, \quad \lambda_H \geq 0,\ H(x) \geq 0,\ \langle \lambda_H,\, H(x)\rangle = 0. \end{gathered} \tag{1.2}$$

In the above, we omit the condition $\lambda_0 \langle G(x),\, H(x)\rangle = 0$, because it is redundant (it follows from $\langle G(x),\, H(x)\rangle = 0$, which is implied by feasibility of x in (1.1)). For $\bar{x} \in \mathbf{R}^n$, let $\Lambda(\bar{x})$ stand for the set of Lagrange multipliers associated with $\bar{x}$, that is, the set of $\lambda = \bar{\lambda} = (\bar{\lambda}_G,\, \bar{\lambda}_H,\, \bar{\lambda}_0) \in \mathbf{R}^m \times \mathbf{R}^m \times \mathbf{R}$ satisfying (1.2) for $x = \bar{x}$. As is well known and can be easily checked, MPCC constraints violate the Mangasarian–Fromovitz constraint qualification and, even more so, the linear independence constraint qualification (LICQ), at every feasible point. Therefore, in general, $\bar{x}$ being a local solution of (1.1) does not guarantee that the set of Lagrange multipliers $\Lambda(\bar{x})$ is nonempty. Nevertheless, $\Lambda(\bar{x})$ happens to be nonempty in many cases of interest, and this became one of the common settings in MPCC literature.

Define further the so-called *MPCC-Lagrangian* of problem (1.1):

$$\mathcal{L}(x,\,\mu) = f(x) - \langle \mu_G,\, G(x)\rangle - \langle \mu_H,\, H(x)\rangle,$$

where $x \in \mathbf{R}^n$ and $\mu = (\mu_G,\, \mu_H) \in \mathbf{R}^m \times \mathbf{R}^m$. To a feasible point $\bar{x}$ we associate the index sets

$$\begin{gathered} I_G = I_G(\bar{x}) = \{i = 1, \ldots, m \mid G_i(\bar{x}) = 0\},\ I_H = I_H(\bar{x}) = \{i = 1, \ldots, m \mid H_i(\bar{x}) = 0\}, \\ I_0 = I_G \cap I_H. \end{gathered} \tag{1.3}$$

A feasible point $\bar{x}$ of (1.1) is said to be a *strongly stationary point* of this problem if there exists an *MPCC-multiplier* $\bar{\mu} = (\bar{\mu}_G,\, \bar{\mu}_H) \in \mathbf{R}^m \times \mathbf{R}^m$ satisfying

$$\frac{\partial \mathcal{L}}{\partial x}(\bar{x},\, \bar{\mu}) = 0,\ (\bar{\mu}_G)_{I_H \setminus I_G} = 0,\ (\bar{\mu}_H)_{I_G \setminus I_H} = 0,\ (\bar{\mu}_G)_{I_0} \geq 0,\ (\bar{\mu}_H)_{I_0} \geq 0, \tag{1.4}$$

where y_I stands for the subvector of the vector y, with components $y_i, i \in I$. Without nonnegativity conditions in (1.4), $\bar{x}$ is called a *weakly stationary point* of (1.1).

We say that *MPCC linear independence constraint qualification* (MPCC-LICQ) holds at $\bar{x}$ if the gradients

$$G_i'(\bar{x}),\ i \in I_G, \quad H_i'(\bar{x}),\ i \in I_H \quad \text{are linearly independent.} \tag{1.5}$$

It was shown in [18, Theorem 2] that if MPCC-LICQ holds at a local solution $\bar{x}$ of (1.1), then this point is strongly stationary, and the associated MPCC-multiplier $\bar{\mu}$ is unique.

The following proposition summarizes some results obtained in [7, Proposition 4.1] and [9, Proposition 1], which will be used in what follows. Its proof can be obtained by a direct computation. Let

$$\bar{\nu} = \max\left\{0,\ \max_{i\in I_G\setminus I_H}\left(-\frac{(\bar{\mu}_G)_i}{H_i(\bar{x})}\right),\ \max_{i\in I_H\setminus I_G}\left(-\frac{(\bar{\mu}_H)_i}{G_i(\bar{x})}\right)\right\}. \tag{1.6}$$

PROPOSITION 1.1. *A feasible point $\bar{x}$ of problem* (1.1) *is a stationary point of this problem if and only if it is a strongly stationary point of this problem. Moreover, if $\bar{\lambda} = (\bar{\lambda}_G, \bar{\lambda}_H, \bar{\lambda}_0)$ is a Lagrange multiplier associated with $\bar{x}$, then $\bar{\mu} = (\bar{\mu}_G, \bar{\mu}_H)$ defined by*

$$(\bar{\mu}_G)_i = \left(\bar{\lambda}_G\right)_i - \bar{\lambda}_0 H_i(\bar{x}),\ i \in I_G \setminus I_H, \quad (\bar{\mu}_G)_i = \left(\bar{\lambda}_G\right)_i,\ i \in I_H, \tag{1.7}$$

$$(\bar{\mu}_H)_i = \left(\bar{\lambda}_H\right)_i - \bar{\lambda}_0 G_i(\bar{x}),\ i \in I_H \setminus I_G, \quad (\bar{\mu}_H)_i = \left(\bar{\lambda}_H\right)_i,\ i \in I_G, \tag{1.8}$$

is an MPCC-multiplier associated with $\bar{x}$. Conversely, if $\bar{\mu} = (\bar{\mu}_G, \bar{\mu}_H)$ is an MPCC-multiplier associated with $\bar{x}$, then any $\bar{\lambda} = (\bar{\lambda}_G, \bar{\lambda}_H, \bar{\lambda}_0)$ satisfying (1.7)–(1.8) *and*

$$\bar{\lambda}_0 \geq \bar{\nu}, \tag{1.9}$$

with $\bar{\nu}$ defined in (1.6)*, is a Lagrange multiplier associated with $\bar{x}$.*

Furthermore, for any $\xi \in \mathbf{R}^n$ and any $\bar{\lambda} = (\bar{\lambda}_G, \bar{\lambda}_H, \bar{\lambda}_0) \in \mathbf{R}^m \times \mathbf{R}^m \times \mathbf{R}$ and $\bar{\mu} = (\bar{\mu}_G, \bar{\mu}_H) \in \mathbf{R}^m \times \mathbf{R}^m$ satisfying (1.7)–(1.8)*, it holds that*

$$\frac{\partial^2 L}{\partial x^2}\left(\bar{x}, \bar{\lambda}\right)[\xi, \xi] = \frac{\partial^2 \mathcal{L}}{\partial x^2}(\bar{x}, \bar{\mu})[\xi, \xi] + 2\bar{\lambda}_0 \sum_{i=1}^{m} \langle G_i'(\bar{x}), \xi\rangle \langle H_i'(\bar{x}), \xi\rangle. \tag{1.10}$$

In particular, if $\bar{\mu}$ is the unique MPCC-multiplier associated with $\bar{x}$ (e.g., under MPCC-LICQ (1.5)), then $\Lambda(\bar{x})$ is the ray defined by (1.7)–(1.9), with its origin corresponding to $\bar{\lambda}_0 = \bar{\nu}$.

It can be easily checked that the standard critical cone of problem (1.1) at $\bar{x}$ is given by

$$C(\bar{x}) = \left\{\xi \in \mathbf{R}^n \;\middle|\; \begin{array}{l} G'_{I_G\setminus I_H}(\bar{x})\xi = 0,\ H'_{I_H\setminus I_G}(\bar{x})\xi = 0,\ G'_{I_0}(\bar{x})\xi \geq 0,\ H'_{I_0}(\bar{x})\xi \geq 0, \\ \langle f'(\bar{x}), \xi\rangle \leq 0 \end{array}\right\}. \tag{1.11}$$

We say that *MPCC–second-order sufficient condition* (MPCC-SOSC) holds at a strongly stationary point $\bar{x}$ of problem (1.1), with the associated MPCC-multiplier $\bar{\mu}$, if

$$\frac{\partial^2 \mathcal{L}}{\partial x^2}(\bar{x}, \bar{\mu})[\xi, \xi] > 0 \quad \forall\, \xi \in C(\bar{x}) \setminus \{0\}. \tag{1.12}$$

Note that, for every $\xi \in C(\bar{x})$, we obtain from (1.11) that (1.10) takes the form

$$\frac{\partial^2 L}{\partial x^2}\left(\bar{x}, \bar{\lambda}\right)[\xi, \xi] = \frac{\partial^2 \mathcal{L}}{\partial x^2}(\bar{x}, \bar{\mu})[\xi, \xi] + 2\bar{\lambda}_0 \sum_{i\in I_0} \langle G_i'(\bar{x}), \xi\rangle \langle H_i'(\bar{x}), \xi\rangle, \tag{1.13}$$

where the last term in the right-hand side is nonnegative. Thus, according to Proposition 1.1, MPCC-SOSC implies the usual SOSC

$$\frac{\partial^2 L}{\partial x^2}\left(\bar{x}, \bar{\lambda}\right)[\xi, \xi] > 0 \quad \forall\, \xi \in C(\bar{x}) \setminus \{0\} \tag{1.14}$$

for any $\bar{\lambda}$ satisfying (1.7)–(1.9). In particular, under MPCC-LICQ (1.5), MPCC-SOSC (1.12) (with the unique MPCC-multiplier $\bar{\mu}$) implies SOSC (1.14), with any $\bar{\lambda}$ in the ray $\Lambda(\bar{x})$, including the origin of this ray.

It is important to point out that MPCC-SOSC is a rather strong condition. In particular, it cannot be linked to any second-order necessary condition for (1.1). By this we mean that a solution of (1.1) that satisfies MPCC-LICQ (1.5) (and thus is strongly stationary) does not have to satisfy the condition obtained from (1.12) by replacing the strict inequality by nonstrict. In our developments, we shall be making use of a SOSC weaker than (1.12), which also happens to be much more natural, because it is related to an appropriate second-order necessary condition for (1.1), as explained below.

For each partition (I_1, I_2) of I_0 (i.e., a pair of index sets such that $I_1 \cup I_2 = I_0$, $I_1 \cap I_2 = \emptyset$), define the *branch* (or *piece*) *MP* at $\bar{x}$ by

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & G_{(I_G \setminus I_H) \cup I_1}(x) = 0,\ H_{(I_H \setminus I_G) \cup I_2}(x) = 0,\ G_{I_2}(x) \geq 0,\ H_{I_1}(x) \geq 0. \end{array} \tag{1.15}$$

There is a finite number of such branch MPs, $\bar{x}$ is feasible for each of them, and in a neighborhood of $\bar{x}$ the feasible set of (1.1) is a union of feasible sets of all branch MPs. It is not difficult to see that the union of the critical cones of all branch MPs at $\bar{x}$ is given by

$$C_2(\bar{x}) = \left\{ \xi \in \mathbf{R}^n \;\middle|\; \begin{array}{l} \langle f'(\bar{x}),\, \xi \rangle \leq 0,\ G'_{I_G \setminus I_H}(\bar{x})\xi = 0,\ H'_{I_H \setminus I_G}(\bar{x})\xi = 0, \\ G'_{I_0}(\bar{x})\xi \geq 0,\ H'_{I_0}(\bar{x})\xi \geq 0,\ \langle G'_i(\bar{x}),\, \xi \rangle \langle H'_i(\bar{x}),\, \xi \rangle = 0,\ i \in I_0 \end{array} \right\}, \tag{1.16}$$

where the subscript "2" indicates that, unlike $C(\bar{x})$, this set takes into account the second-order information about the last constraint in (1.1). By direct comparison of (1.11) and (1.16), we have that

$$C_2(\bar{x}) \subset C(\bar{x}). \tag{1.17}$$

We say that *piecewise SOSC* holds at a strongly stationary point $\bar{x}$ of problem (1.1), with an associated MPCC-multiplier $\bar{\mu}$, if

$$\frac{\partial^2 \mathcal{L}}{\partial x^2}(\bar{x},\, \bar{\mu})[\xi,\, \xi] > 0 \quad \forall\, \xi \in C_2(\bar{x}) \setminus \{0\}. \tag{1.18}$$

From (1.4), it evidently follows that if $\bar{\mu} = (\bar{\mu}_G,\, \bar{\mu}_H)$ is an MPCC-multiplier associated with $\bar{x}$, then the pair $((\bar{\mu}_G)_{I_G},\, (\bar{\mu}_H)_{I_H})$ is a Lagrange multiplier associated with $\bar{x}$ for the branch MP (1.15). It follows that piecewise SOSC (1.18) implies SOSC for each branch at $\bar{x}$. This, in turn, guarantees that $\bar{x}$ is a strict local solution of (1.1). Thus, piecewise SOSC is indeed sufficient for optimality, even though it is evidently weaker than MPCC-SOSC (see (1.17)).

It is important to emphasize that under MPCC-LICQ (1.5), the condition obtained from (1.18) by replacing the strict inequality by nonstrict is necessary for optimality [18, Theorem 7]. In this sense, piecewise SOSC (1.18) is a more natural assumption than MPCC-SOSC (1.12), as the latter has no relation to any second-order necessary optimality condition.

Suppose that MPCC-LICQ (1.5) and piecewise SOSC (1.18) (with the unique MPCC-multiplier $\bar{\mu}$) hold at a strongly stationary point $\bar{x}$ of problem (1.1). From (1.13) and [9, Proposition 2], it follows that in this case either SOSC (1.14) holds

with all $\bar{\lambda}$ in the ray $\Lambda(\bar{x})$, or possibly there exists $\hat{\nu} \geq \bar{\nu}$ such that SOSC (1.14) does not hold for all $\bar{\lambda}$ corresponding to $\bar{\lambda}_0 \in [\bar{\nu}, \hat{\nu}]$, and holds for all $\bar{\lambda}$ corresponding to $\bar{\lambda}_0 > \hat{\nu}$. Conversely, if SOSC (1.14) holds for some $\bar{\lambda} \in \Lambda(\bar{x})$, from (1.13) and (1.16), taking also into account (1.17), it is easy to see that piecewise SOSC (1.18) holds as well. Thus, under MPCC-LICQ, SOSC (with some multiplier) is equivalent to piecewise SOSC.

Despite the inevitable violation of standard constraint qualifications, there exists some numerical evidence of good performance of sequential quadratic programming (SQP) algorithms for MPCCs (see [6]). Moreover, [7] gives some theoretical justification for local superlinear convergence of the SQP algorithm for MPCC under a set of assumptions that includes MPCC-LICQ and MPCC-SOSC, among other things. However, it is very easy to provide examples satisfying all *natural* in MPCC context requirements (say, MPCC-LICQ and piecewise SOSC), and such that SQP does not possess superlinear convergence; see, e.g., the example in [7, section 7.3], discussed also in detail in [11, section 6]. This means that the existing evidence supporting the use of standard optimization algorithms (say, SQP) for MPCC cannot be regarded as completely satisfactory, and it still makes sense to develop special algorithms which take into account special structure of MPCC, and which are guaranteed to achieve quadratic convergence under more natural assumptions.

Let us recall now the main idea of the piecewise SQP algorithm, suggested originally in [17] for MPs with linear complementarity constraints and then extended in [15] to the nonlinear case. An iteration of piecewise SQP is organized as follows: identify *any* branch MP valid at the solution $\bar{x}$ that is being approximated, and perform a step of standard SQP for this branch. In order to identify a valid branch MP, it suffices to (over)estimate the sets $I_G \setminus I_H$ and $I_H \setminus I_G$ (see (1.15)). Locally, this comes for free, with no significant computational cost and with no assumptions needed. However, in order to justify the overall superlinear convergence of piecewise SQP, one needs to guarantee superlinear convergence of SQP for each branch, and dual convergence to the same multiplier for all branches. This results in the following set of assumptions: MPCC-LICQ (1.5) and piecewise SOSC (1.18) at the solution $\bar{x}$.

In this paper, we suggest a local algorithm based on the following idea (to some extent motivated by the development in [13]). Instead of an arbitrary valid branch, we identify the index sets I_G and I_H and perform the Newton–Lagrange steps for the following purely equality-constrained *tightened MP*:

$$\min\ f(x) \quad \text{s.t.} \quad G_{I_G}(x) = 0,\ H_{I_H}(x) = 0. \tag{1.19}$$

Note that this problem is not a branch MP, in general, but its feasible set is contained in the feasible sets of all branch MPs.

For quadratic convergence of the Newton–Lagrange method for (1.19), we need to assume MPCC-LICQ (1.5) and SOSC for this problem, the latter being evidently guaranteed by piecewise SOSC (1.18). Local identification of I_G and I_H uses the procedure suggested in [4] and the error bound following from [8, Lemma 2] and [5, Theorem 2] (see (2.6)). The identification technique based on this combination of tools (first used for problems without any regularity assumptions on constraints in [10]) still costs nothing computationally. The error bound requires some $\bar{\lambda} \in \Lambda(\bar{x})$ satisfying SOSC (1.14). According to our discussion above, the existence of such $\bar{\lambda}$ can again be guaranteed under MPCC-LICQ (1.5) and piecewise SOSC (1.18). Hence, we obtain local quadratic convergence of our algorithm under the same set of assumptions as for piecewise SQP: MPCC-LICQ (1.5) and piecewise SOSC (1.18) at $\bar{x}$. At the same

time, our local algorithm enjoys the advantage of being quadratic program (QP)-free: it requires solving only one linear system per iteration. Of course, within a local framework, this may not always be a big advantage. Note, however, that globalized Algorithm 3.2 in section 3.1 is QP-free globally.

2. Local algorithms. As is well known, the KKT system (1.2) can be written in the form

$$\Phi(x,\,\lambda) = 0,$$

where $\Phi : \mathbf{R}^n \times (\mathbf{R}^m \times \mathbf{R}^m \times \mathbf{R}) \to \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{R}^m \times \mathbf{R}$,

$$\Phi(x,\,\lambda) = \left(\frac{\partial L}{\partial x}(x,\,\lambda),\, \rho(\lambda_G,\, G(x)),\, \rho(\lambda_H,\, H(x)),\, \rho(\lambda_0,\, -\langle G(x),\, H(x)\rangle)\right),$$

and $\rho : \mathbf{R} \times \mathbf{R} \to \mathbf{R}$ is a complementarity function (that is, a function such that $\rho(a,\,b) = 0 \iff a \geq 0,\ b \geq 0,\ ab = 0$) applied componentwise. In what follows, we shall make use of two complementarity functions, namely, the natural residual $\rho(a,\,b) = \min\{a,\,b\}$ and the Fischer–Burmeister function $\rho(a,\,b) = \sqrt{a^2+b^2} - a - b$. The corresponding version of Φ will be denoted by Φ_{NR} and Φ_{FB}, respectively. As is well known, both these mappings are semismooth (and in particular, locally Lipschitz). Moreover, according to [19], these two complementarity functions are equivalent in terms of their growth rates. This means that, throughout the paper, Φ_{NR} can actually be replaced by Φ_{FB} without any changes in the analysis or results.

ALGORITHM 2.1. `Preliminary step.` Fix $\theta \in (0,\,1)$. Choose $x^0 \in \mathbf{R}^n$, $\lambda^0 = (\lambda_G^0,\, \lambda_H^0,\, \lambda_0^0) \in \mathbf{R}^m \times \mathbf{R}^m \times \mathbf{R}$.

`Identification step.` Compute the index sets

$$I_G = I_G\left(x^0,\,\lambda^0\right) = \left\{i = 1,\ldots,m \mid G_i\left(x^0\right) \leq \|\Phi_{NR}\left(x^0,\,\lambda^0\right)\|^\theta\right\}, \tag{2.1}$$

$$I_H = I_H\left(x^0,\,\lambda^0\right) = \left\{i = 1,\ldots,m \mid H_i\left(x^0\right) \leq \|\Phi_{NR}\left(x^0,\,\lambda^0\right)\|^\theta\right\}. \tag{2.2}$$

`Main step.` Generate the sequence $\{(x^k,\,\mu^k)\}$, with $\mu^k = (\mu_G^k,\,\mu_H^k) \in \mathbf{R}^m \times \mathbf{R}^m$, as follows.

• Generate the sequence $\{(x^k,\,(\mu_G^k)_{I_G},\,(\mu_H^k)_{I_H})\}$ by the Newton–Lagrange method for tightened MP (1.19) (that is, the Newton method applied to the Lagrange optimality system of this problem) starting from $(x^0,\,(\mu_G^0)_{I_G},\,(\mu_H^0)_{I_H})$, with $(\mu_G^0)_{I_G}$ and $(\mu_H^0)_{I_H}$ defined by

$$\left(\mu_G^0\right)_i = \left(\lambda_G^0\right)_i - \lambda_0^0 H_i\left(x^0\right),\ i \in I_G \setminus I_H, \quad \left(\mu_G^0\right)_i = \left(\lambda_G^0\right)_i,\ i \in I_G \cap I_H, \tag{2.3}$$

$$\left(\mu_H^0\right)_i = \left(\lambda_H^0\right)_i - \lambda_0^0 G_i\left(x^0\right),\ i \in I_H \setminus I_G, \quad \left(\mu_H^0\right)_i = \left(\lambda_H^0\right)_i,\ i \in I_G \cap I_H. \tag{2.4}$$

• Set

$$\left(\mu_G^k\right)_{I_H \setminus I_G} = 0, \quad \left(\mu_H^k\right)_{I_G \setminus I_H} = 0 \quad \forall\, k = 0, 1, \ldots. \tag{2.5}$$

THEOREM 2.1. *Let $\bar{x}$ be a local solution of MPCC* (1.1)*, and assume that MPCC-LICQ* (1.5) *holds at $\bar{x}$. Furthermore, let $\bar{\mu}$ be the (unique) MPCC-multiplier associated with $\bar{x}$, and suppose that $(x^0,\,\lambda^0)$ is close enough to $(\bar{x},\,\bar{\lambda})$, with some $\bar{\lambda} \in \Lambda(\bar{x})$ satisfying SOSC* (1.14).

Then Algorithm 2.1 *correctly generates the sequence $\{(x^k,\,\mu^k)\}$, which converges quadratically to $(\bar{x},\,\bar{\mu})$.*

Proof. According to [8, Lemma 2] and [5, Theorem 2], SOSC (1.14) implies the existence of $c > 0$ such that the error bound

$$\operatorname{dist}((x, \lambda), \{\bar{x}\} \times \Lambda(\bar{x})) \leq c\|\Phi_{NR}(x, \lambda)\| \tag{2.6}$$

holds for all $(x, \lambda) \in \mathbf{R}^n \times (\mathbf{R}^m \times \mathbf{R}^m \times \mathbf{R})$ close enough to $(\bar{x}, \bar{\lambda})$. Since (x^0, λ^0) is close enough to $(\bar{x}, \bar{\lambda})$, from [4, Theorem 2.2] it follows that the index sets $I_G = I_G(x^0, \lambda^0)$ and $I_H = I_H(x^0, \lambda^0)$, computed according to (2.1) and (2.2), coincide with $I_G = I_G(\bar{x})$ and $I_H = I_H(\bar{x})$, respectively, defined in (1.3).

Furthermore, the point $\bar{x}$ is a local solution of tightened MP (1.19), and MPCC-LICQ (1.5) means that LICQ holds at $\bar{x}$ for the constraints of (1.19). In particular, $\bar{x}$ is a stationary point of (1.19), and from (1.4) it evidently follows that $((\bar{\mu}_G)_{I_G}, (\bar{\mu}_H)_{I_H})$ is the unique Lagrange multiplier associated with this stationary point.

Stationarity of $\bar{x}$ in (1.19) evidently implies that

$$\langle f'(\bar{x}), \xi\rangle = 0 \quad \forall\, \xi \in \ker G'_{I_G}(\bar{x}) \cap \ker H'_{I_H}(\bar{x}),$$

where $\ker A$ stands for the kernel (null space) of a linear operator A. Hence, by (1.11),

$$\ker G'_{I_G}(\bar{x}) \cap \ker H'_{I_H}(\bar{x}) \subset C(\bar{x}).$$

From Proposition 1.1 (see (1.10)) and SOSC (1.14), it now follows that

$$\frac{\partial^2 \mathcal{L}}{\partial x^2}(\bar{x}, \bar{\mu})[\xi, \xi] = \frac{\partial^2 L}{\partial x^2}\left(\bar{x}, \bar{\lambda}\right)[\xi, \xi] > 0 \quad \forall\, \xi \in \left(\ker G'_{I_G}(\bar{x}) \cap \ker H'_{I_H}(\bar{x})\right) \setminus \{0\},$$

and according to the equalities $(\bar{\mu}_G)_{I_H \setminus I_G} = 0$ and $(\bar{\mu}_H)_{I_G \setminus I_H} = 0$ in (1.4), the latter means that SOSC holds at $\bar{x}$ for tightened MP (1.19) (with the unique associated multiplier $((\bar{\mu}_G)_{I_G}, (\bar{\mu}_H)_{I_H})$.

Finally, since (x^0, λ^0) is close enough to $(\bar{x}, \bar{\lambda})$, the pair $((\mu^0_G)_{I_G}, (\mu^0_H)_{I_H})$ defined by (2.3)–(2.4) will be close enough to $((\bar{\mu}_G)_{I_G}, (\bar{\mu}_H)_{I_H})$ (recall that, according to Proposition 1.1, the latter satisfies (1.7)–(1.8)). From the standard convergence result for the Newton–Lagrange method, it now follows that the sequence $\{(x^k, (\mu^k_G)_{I_G}, (\mu^k_H)_{I_H})\}$ is correctly defined and converges quadratically to the point $(\bar{x}, (\bar{\mu}_G)_{I_G}, (\bar{\mu}_H)_{I_H})$. At the same time, according to (1.4) and (2.5), it holds that

$$\left(\mu^k_G\right)_{I_H \setminus I_G} = (\bar{\mu}_G)_{I_H \setminus I_G} = 0, \quad \left(\mu^k_H\right)_{I_G \setminus I_H} = (\bar{\mu}_H)_{I_G \setminus I_H} = 0 \quad \forall\, k = 0, 1, \ldots.$$

This completes the proof. ∎

Let us discuss briefly the assumptions of Theorem 2.1. These assumptions are, in a sense, "minimal." In particular, none of them can be removed, as illustrated next.

MPCC-LICQ (1.5) is needed for nondegeneracy of constraints of tightened MP (1.19), which is clearly necessary for the approach to be valid; otherwise, the linearized constraints of tightened MP can be inconsistent arbitrarily close to a solution. To this end, consider, e.g., $n = 2$, $m = 1$, $f(x) = x_2^2/2$, $G(x) = x_1 + x_1^2/2$, and $H(x) = x_1 + x_1^2$. Then $\bar{x} = 0$ is a strongly stationary point of (1.1) satisfying SOSC (1.14) but violating MPCC-LICQ (1.5). It is easily seen that linearization of tightened MP (1.19) at any $x \in \mathbf{R}^2$ with $x_1 \neq 0$ is inconsistent.

The role of SOSC (1.14) is twofold. First, it is used for identification of the index sets I_G and I_H. To see that without SOSC identification can be incorrect, consider $n = 2$, $m = 1$, $f(x) = x_1$, $G(x) = x_1$, and $H(x) = x_2$. Then $\bar{x} = 0$ is a solution of (1.1) satisfying MPCC-LICQ (1.5) but violating SOSC (1.14) (one can

even add, e.g., $|x_2|^3$ to the objective function in order to make this solution strict). Take $\lambda^0 = (1, 0, 0) \in \Lambda(\bar{x})$, and let $x_1^0 \geq 0$, $x_2^0 \geq 0$. Then $\|\Phi_{NR}(x^0, \lambda^0)\| = x_1^0$, and for any fixed $\theta \in (0, 1)$, by taking $x_2^0 = (x_1^0)^{\theta/2}$, we obtain a point $x^0 = (x_1^0, x_2^0)$ which can be arbitrarily close to $\bar{x}$, while (2.2) will always (incorrectly) identify I_H as empty at such point.

Finally, even if the identification is correct, SOSC (1.14) is still needed as it guarantees SOSC for tightened MP (1.19). Let, e.g., $n = 2$, $m = 1$, $f(x) = x_1 + |x_2|^3$, $G(x) = x_1$, and $H(x) \equiv 1$. Then $\bar{x} = 0$ is a solution satisfying MPCC-LICQ (1.5) but violating SOSC (1.14). Moreover, tightened MP is also violating SOSC. It can be checked directly that the convergence rate of the Newton–Lagrange method for tightened MP (1.19) is only linear.

Note that the presented algorithm appears more suitable for globalization than, say, piecewise SQP. This is because Algorithm 2.1 uses as a dual starting point an approximation of Lagrange multiplier rather than an approximation of MPCC-multiplier. The proximity to points satisfying KKT system (1.2) (and hence, to Lagrange multipliers) can be controlled via some globally defined merit functions (like the norm of Φ_{NR} or Φ_{FB}). By contrast, the definition (1.4) of MPCC-multipliers involves the index sets I_G and I_H depending on a specific $\bar{x}$, and it seems difficult to suggest a reasonable globally defined merit function characterizing MPCC-multipliers.

Furthermore, having in mind globalization of convergence, it can be useful to consider a modified algorithm, with `Identification step` being performed not only once (at the beginning of the process) but before each iteration of `Main step`. Identification is a very cheap procedure and, therefore, this modification will not increase computational costs significantly. However, in this case we will need to generate not only the sequence $\{(x^k, \mu^k)\}$ but also an appropriate sequence $\{\lambda^k\} \subset \mathbf{R}^m \times \mathbf{R}^m \times \mathbf{R}$, and redefine I_G and I_H accordingly:

$$I_G = I_G\left(x^k, \lambda^k\right) = \left\{i = 1, \ldots, m \mid G_i\left(x^k\right) \leq \|\Phi_{NR}\left(x^k, \lambda^k\right)\|^\theta\right\}, \tag{2.7}$$

$$I_H = I_H\left(x^k, \lambda^k\right) = \left\{i = 1, \ldots, m \mid H_i\left(x^k\right) \leq \|\Phi_{NR}\left(x^k, \lambda^k\right)\|^\theta\right\} \tag{2.8}$$

for each $k = 0, 1, \ldots$. Clearly, for all the conclusions of Theorem 2.1 to remain valid for this modified algorithm, it suffices to show that $\{\lambda^k\}$ stays close to $\bar{\lambda}$, which can be achieved by keeping it close to λ^0. In particular, one can just take $\lambda^k = \lambda^0$ $\forall\, k = 1, 2, \ldots$. Another option, which seems more suitable for globalization purposes (and which is more in the spirit of SQP methods), is realized in the following method.

ALGORITHM 2.2. `Preliminary step.` Fix $\theta \in (0, 1)$. Set $k = 0$ and choose $x^0 \in \mathbf{R}^n$ and $\lambda^0 = (\lambda_G^0, \lambda_H^0, \lambda_0^0) \in \mathbf{R}^m \times \mathbf{R}^m \times \mathbf{R}$.

`Identification step.` Define the index sets I_G and I_H according to (2.7) and (2.8). If $k = 0$, define $(\mu_G^0)_{I_G}$ and $(\mu_H^0)_{I_H}$ according to (2.3)–(2.4).

`Main step.` Compute (x^{k+1}, μ^{k+1}) as follows.

• The triple $(x^{k+1}, (\mu_G^{k+1})_{I_G}, (\mu_H^{k+1})_{I_H})$ is generated by the step of Newton–Lagrange method for tightened MP (1.19) from the point $(x^k, (\mu_G^k)_{I_G}, (\mu_H^k)_{I_H})$.

• Set $(\mu_G^{k+1})_{I_H \setminus I_G} = 0$, $(\mu_H^{k+1})_{I_G \setminus I_H} = 0$.

Set

$$\nu_{k+1} = \max\left\{0, \max_{i \in I_G \setminus I_H}\left(-\frac{\left(\mu_G^{k+1}\right)_i}{H_i(x^{k+1})}\right), \max_{i \in I_H \setminus I_G}\left(-\frac{\left(\mu_H^{k+1}\right)_i}{G_i(x^{k+1})}\right)\right\}, \tag{2.9}$$

and define $\lambda^{k+1} = (\lambda_G^{k+1}, \lambda_H^{k+1}, \lambda_0^{k+1})$ as follows:

$$\lambda_0^{k+1} = \max\left\{\nu_{k+1}, \lambda_0^k\right\}, \tag{2.10}$$

$$(2.11)\quad \left(\lambda_G^{k+1}\right)_i = \left(\mu_G^{k+1}\right)_i + \lambda_0^{k+1} H_i\left(x^{k+1}\right), i \in I_G \setminus I_H,\ \left(\lambda_G^{k+1}\right)_i = \left(\mu_G^{k+1}\right)_i, i \in I_H,$$

$$(2.12)\quad \left(\lambda_H^{k+1}\right)_i = \left(\mu_H^{k+1}\right)_i + \lambda_0^{k+1} G_i\left(x^{k+1}\right), i \in I_H \setminus I_G,\ \left(\lambda_H^{k+1}\right)_i = \left(\mu_H^{k+1}\right)_i, i \in I_G.$$

Adjust k by 1 and go to `Identification step`.

For purposes of convergence analysis, we need to introduce some auxiliary dual estimates. Suppose that, for some $k = 0, 1, \ldots$, Algorithm 2.2 correctly defined x^k, λ^k, and μ^k. Define $\hat{\lambda}^k = (\hat{\lambda}_G^k, \hat{\lambda}_H^k, \hat{\lambda}_0^k) \in \mathbf{R}^m \times \mathbf{R}^m \times \mathbf{R}$ as follows:

$$(2.13)\quad \hat{\lambda}_0^k = \lambda_0^k,$$

$$(2.14)\quad \left(\hat{\lambda}_G^k\right)_i = (\bar{\mu}_G)_i + \lambda_0^k H_i(\bar{x}),\ i \in I_G \setminus I_H, \quad \left(\hat{\lambda}_G^k\right)_i = (\bar{\mu}_G)_i,\ i \in I_H,$$

$$(2.15)\quad \left(\hat{\lambda}_H^k\right)_i = (\bar{\mu}_H)_i + \lambda_0^k G_i(\bar{x}),\ i \in I_H \setminus I_G, \quad \left(\hat{\lambda}_H^k\right)_i = (\bar{\mu}_H)_i,\ i \in I_G.$$

According to (1.7)–(1.8) and (2.13)–(2.15), it holds that

$$(2.16)\quad \|\hat{\lambda}^k - \bar{\lambda}\| \le \left(1 + \max\left\{\max_{i \in I_G \setminus I_H} H_i(\bar{x}), \max_{i \in I_H \setminus I_G} G_i(\bar{x})\right\}\right) |\lambda_0^k - \bar{\lambda}_0|,$$

and hence, $\hat{\lambda}^k$ is close to $\bar{\lambda}$ provided (x^k, λ^k) is close enough to $(\bar{x}, \bar{\lambda})$.

From Theorem 2.1, we obtain that if (x^k, λ^k) is close enough to $(\bar{x}, \bar{\lambda})$, then the points x^{k+1} and μ^{k+1} will be correctly defined by Algorithm 2.2, and

$$(2.17)\quad \left\|x^{k+1} - \bar{x}\right\| = O\left(\left\|\left(x^k - \bar{x}, \mu^k - \bar{\mu}\right)\right\|^2\right) = O\left(\left\|\left(x^k - \bar{x}, \lambda^k - \hat{\lambda}^k\right)\right\|^2\right),$$

$$(2.18)\quad \left\|\mu^{k+1} - \bar{\mu}\right\| = O\left(\left\|\left(x^k - \bar{x}, \mu^k - \bar{\mu}\right)\right\|^2\right) = O\left(\left\|\left(x^k - \bar{x}, \lambda^k - \hat{\lambda}^k\right)\right\|^2\right).$$

Furthermore, according to (1.6)–(1.8), (2.9), (2.11)–(2.15), (2.17), and (2.18), we obtain the estimates

$$(2.19)\quad |\nu_{k+1} - \bar{\nu}| = O\left(\left\|x^{k+1} - \bar{x}\right\|\right) + O\left(\left\|\mu^{k+1} - \bar{\mu}\right\|\right) = O\left(\left\|\left(x^k - \bar{x}, \lambda^k - \hat{\lambda}^k\right)\right\|^2\right),$$

$$\begin{aligned}(2.20)\quad \left\|\lambda^{k+1} - \hat{\lambda}^k\right\| &= O\left(\left\|x^{k+1} - \bar{x}\right\|\right) + O\left(\left\|\mu^{k+1} - \bar{\mu}\right\|\right) + O\left(|\lambda_0^{k+1} - \lambda_0^k|\right) \\ &= O\left(\left\|\left(x^k - \bar{x}, \lambda^k - \hat{\lambda}^k\right)\right\|^2\right) + O\left(|\lambda_0^{k+1} - \lambda_0^k|\right),\end{aligned}$$

$$\begin{aligned}(2.21)\quad \left\|\lambda^{k+1} - \bar{\lambda}\right\| &= O\left(\left\|x^{k+1} - \bar{x}\right\|\right) + O\left(\left\|\mu^{k+1} - \bar{\mu}\right\|\right) + O\left(|\lambda_0^{k+1} - \bar{\lambda}_0|\right) \\ &= O\left(\left\|\left(x^k - \bar{x}, \lambda^k - \hat{\lambda}^k\right)\right\|^2\right) + O\left(|\lambda_0^{k+1} - \bar{\lambda}_0|\right).\end{aligned}$$

Let us consider separately the two cases

$$(2.22)\quad \bar{\lambda}_0 > \bar{\nu} \quad \text{or} \quad \bar{\lambda}_0 = \bar{\nu}$$

(see (1.9)).

LEMMA 2.2. *Let $\bar{x}$ be a local solution of MPCC* (1.1)*, and assume that MPCC-LICQ* (1.5) *holds at $\bar{x}$. Furthermore, suppose that, for some $k = 0, 1, \ldots$, Algorithm* 2.2 *generated points x^k, λ^k, and μ^k such that (x^k, λ^k) is close enough to $(\bar{x}, \bar{\lambda})$, with some $\bar{\lambda} \in \Lambda(\bar{x})$ satisfying SOSC* (1.14).

Then the points x^{k+1}, λ^{k+1}, and μ^{k+1} will be correctly generated by Algorithm 2.2*, and if the first relation in* (2.22) *holds, with $\bar{\nu}$ defined according to* (1.6)*, then*

$$\lambda_0^{k+1} = \lambda_0^k, \tag{2.23}$$

$$\left\|\lambda^{k+1} - \hat{\lambda}^k\right\| = O\left(\left\|\left(x^k - \bar{x},\, \lambda^k - \hat{\lambda}^k\right)\right\|^2\right). \tag{2.24}$$

Proof. By the first relation in (2.22), we obtain that if (x^k, λ^k) is close enough to $(\bar{x}, \bar{\lambda})$, then

$$\lambda_0^k > \bar{\lambda}_0 - \left(\bar{\lambda}_0 - \bar{\nu}\right)/2 = \left(\bar{\nu} + \bar{\lambda}_0\right)/2, \quad \nu_{k+1} < \bar{\nu} + \left(\bar{\lambda}_0 - \bar{\nu}\right)/2 = \left(\bar{\nu} + \bar{\lambda}_0\right)/2,$$

with inequality in the last relation being implied by (2.16) and (2.19). Thus $\lambda_0^k > \nu_{k+1}$, and by (2.10), we obtain (2.23). Estimate (2.24) follows from (2.20) and (2.23). ∎

LEMMA 2.3. *Under the assumptions of Lemma* 2.2*, if the second relation in* (2.22) *holds, with $\bar{\nu}$ defined according to* (1.6)*, then the following estimates are valid.*

If $\lambda_0^{k+1} = \nu_{k+1}$, then

$$\left\|\lambda^{k+1} - \bar{\lambda}\right\| = O\left(\left\|\left(x^k - \bar{x},\, \lambda^k - \hat{\lambda}^k\right)\right\|^2\right), \tag{2.25}$$

$$\left\|\hat{\lambda}^{k+1} - \bar{\lambda}\right\| = O\left(\left\|\left(x^k - \bar{x},\, \lambda^k - \hat{\lambda}^k\right)\right\|^2\right). \tag{2.26}$$

If $\lambda_0^{k+1} = \lambda_0^k$, then (2.24) *holds.*

Proof. If $\lambda_0^{k+1} = \nu_{k+1}$, then by (2.16), (2.19), (2.21), and the second relation in (2.22), we obtain the estimates

$$\begin{aligned}
\left\|\lambda^{k+1} - \bar{\lambda}\right\| &= O\left(\left\|\left(x^k - \bar{x},\, \lambda^k - \hat{\lambda}^k\right)\right\|^2\right) + O\left(\left|\lambda_0^{k+1} - \bar{\lambda}_0\right|\right) \\
&= O\left(\left\|\left(x^k - \bar{x},\, \lambda^k - \hat{\lambda}^k\right)\right\|^2\right) + O\left(\left|\nu_{k+1} - \bar{\nu}\right|\right) = O\left(\left\|\left(x^k - \bar{x},\, \lambda^k - \hat{\lambda}^k\right)\right\|^2\right), \\
\left\|\hat{\lambda}^{k+1} - \bar{\lambda}\right\| &= O\left(\left|\lambda_0^{k+1} - \bar{\lambda}_0\right|\right) = O\left(\left|\nu_{k+1} - \bar{\nu}\right|\right) = O\left(\left\|\left(x^k - \bar{x},\, \lambda^k - \hat{\lambda}^k\right)\right\|^2\right).
\end{aligned}$$

This proves (2.25) and (2.26).

If $\lambda_0^{k+1} = \lambda_0^k$, then estimate (2.24) follows immediately from (2.20). ∎

We are now in position to prove convergence of Algorithm 2.2.

THEOREM 2.4. *Under the assumptions of Theorem* 2.1*, Algorithm* 2.2 *correctly generates the sequence $\{(x^k, \lambda^k, \mu^k)\}$ such that $\{(x^k, \mu^k)\}$ converges quadratically to $(\bar{x}, \bar{\mu})$. Moreover, if the first relation in* (2.22) *holds, then the sequence $\{(x^k, \lambda^k)\}$ converges quadratically to $(\bar{x}, \hat{\lambda}^0)$, with $\hat{\lambda}^0$ defined according to* (2.13)–(2.15)*, and $\hat{\lambda}^0 \in \Lambda(\bar{x})$.*

Proof. If the first relation in (2.22) holds, then employing (2.13)–(2.15), (2.16), (2.17), and Lemma 2.2 (see (2.23) and (2.24)), it can be shown (by standard argument) that if (x^0, λ^0) is close enough to $(\bar{x}, \bar{\lambda})$, then each further step of Algorithm 2.2 will produce a pair (x^{k+1}, λ^{k+1}), with $\lambda_0^{k+1} = \lambda_0^k = \lambda_0^0$, and this new pair will be close to $(\bar{x}, \hat{\lambda}^k) = (\bar{x}, \hat{\lambda}^0)$, which in turn is close to $(\bar{x}, \bar{\lambda})$. Then by the same argument as in the proof of Theorem 2.1, for any k, the index sets $I_G(x^k, \lambda^k)$ and $I_H(x^k, \lambda^k)$ computed according to (2.7) and (2.8) will coincide with $I_G = I_G(\bar{x})$ and $I_H = I_H(\bar{x})$ defined in (1.3), respectively. This means that Algorithm 2.2 generates

exactly the same trajectory $\{(x^k, \mu^k)\}$ as Algorithm 2.1, and quadratic convergence follows now from Theorem 2.1. Furthermore, quadratic convergence of $\{(x^k, \lambda^k)\}$ to $(\bar{x}, \hat{\lambda}^0)$ follows from (2.17), (2.24), and the above-established equality $\hat{\lambda}^k = \hat{\lambda}^0$ $\forall\, k$. Finally, $\hat{\lambda}^0 \in \Lambda(\bar{x})$ according to (2.13)–(2.15), the first relation in (2.22), and Proposition 1.1.

We proceed with the case when the second relation in (2.22) holds. Again we need to show that if (x^0, λ^0) is close enough to $(\bar{x}, \bar{\lambda})$, then $\{(x^k, \lambda^k)\}$ stays close to $(\bar{x}, \bar{\lambda})$. Then the needed assertion will follow the same way as for the previous case.

From (2.16)–(2.17) and (2.24)–(2.26), it follows that, for any $q \in (0, 1/2]$, there exists $\varepsilon > 0$ such that for all (x^k, λ^k) satisfying $\|x^k - \bar{x}\| < \varepsilon$ and $\|\lambda^k - \bar{\lambda}\| < \varepsilon$ the following estimates are valid.

If $\lambda_0^{k+1} = \nu_{k+1}$, then

$$\left\|(x^{k+1} - \bar{x},\, \lambda^{k+1} - \bar{\lambda})\right\| \le q \left\|\left(x^k - \bar{x},\, \lambda^k - \hat{\lambda}^k\right)\right\|, \tag{2.27}$$

$$\left\|\left(x^{k+1} - \bar{x},\, \hat{\lambda}^{k+1} - \bar{\lambda}\right)\right\| \le q \left\|\left(x^k - \bar{x},\, \lambda^k - \hat{\lambda}^k\right)\right\|. \tag{2.28}$$

If $\lambda_0^{k+1} = \lambda_0^k$, then

$$\left\|\left(x^{k+1} - \bar{x},\, \lambda^{k+1} - \hat{\lambda}^k\right)\right\| \le q \left\|\left(x^k - \bar{x},\, \lambda^k - \hat{\lambda}^k\right)\right\|. \tag{2.29}$$

Let (x^0, λ^0) be close enough to $(\bar{x}, \bar{\lambda})$, so that

$$\left\|\left(x^0 - \bar{x},\, \lambda^0 - \hat{\lambda}^0\right)\right\| < \delta, \quad \left\|\left(x^0 - \bar{x},\, \hat{\lambda}^0 - \bar{\lambda}\right)\right\| < \delta, \tag{2.30}$$

where $\delta > 0$ satisfies the inequality

$$(q+1)\delta \le \varepsilon \tag{2.31}$$

(see (2.16)). We now prove by induction that $\forall\, k = 1, 2, \ldots$, it holds that

$$\left\|\left(x^k - \bar{x},\, \lambda^k - \hat{\lambda}^k\right)\right\| < \delta, \tag{2.32}$$

$$\left\|\left(x^k - \bar{x},\, \hat{\lambda}^k - \bar{\lambda}\right)\right\| < \delta, \tag{2.33}$$

$$\left\|(x^k - \bar{x},\, \lambda^k - \bar{\lambda})\right\| < \varepsilon. \tag{2.34}$$

Let $k = 1$. If $\lambda_0^1 = \nu_1$, then by (2.27), (2.30), and (2.31), we obtain

$$\left\|(x^1 - \bar{x},\, \lambda^1 - \bar{\lambda})\right\| \le q \left\|\left(x^0 - \bar{x},\, \lambda^0 - \hat{\lambda}^0\right)\right\| < q\delta < \varepsilon, \tag{2.35}$$

i.e., (2.34) holds for $k = 1$. Furthermore, by (2.28), (2.30), and by the inequality $q < 1$,

$$\left\|\left(x^1 - \bar{x},\, \hat{\lambda}^1 - \bar{\lambda}\right)\right\| \le q \left\|\left(x^0 - \bar{x},\, \lambda^0 - \hat{\lambda}^0\right)\right\| < q\delta < \delta, \tag{2.36}$$

i.e., (2.33) holds for $k = 1$. Finally, by (2.35), (2.36), and by the inequality $q \le 1/2$,

$$\left\|\left(x^1 - \bar{x},\, \lambda^1 - \hat{\lambda}^1\right)\right\| \le \left\|(x^1 - \bar{x},\, \lambda^1 - \bar{\lambda})\right\| + \left\|\hat{\lambda}^1 - \bar{\lambda}\right\| < 2q\delta \le \delta,$$

i.e., (2.32) holds for $k = 1$.

On the other hand, if $\lambda_0^1 = \lambda_0^0$, then by (2.13)–(2.15) $\hat{\lambda}^1 = \hat{\lambda}^0$, and by (2.29), (2.30),

$$\left\|\left(x^1 - \bar{x},\, \lambda^1 - \hat{\lambda}^0\right)\right\| \le q\left\|\left(x^0 - \bar{x},\, \lambda^0 - \hat{\lambda}^0\right)\right\| < q\delta,$$

and hence, by the inequality $q < 1$, we have that

$$\left\|\left(x^1 - \bar{x},\, \lambda^1 - \hat{\lambda}^1\right)\right\| = \left\|\left(x^1 - \bar{x},\, \lambda^1 - \hat{\lambda}^0\right)\right\| < q\delta < \delta, \tag{2.37}$$

i.e., (2.32) holds for $k = 1$. Furthermore, by (2.30), we have that

$$\left\|\left(x^1 - \bar{x},\, \hat{\lambda}^1 - \bar{\lambda}\right)\right\| = \left\|\left(x^1 - \bar{x},\, \hat{\lambda}^0 - \bar{\lambda}\right)\right\| < \delta,$$

i.e., (2.33) holds for $k = 1$. Finally, by (2.31), (2.33) for $k = 1$, and (2.37), we obtain

$$\left\|(x^1 - \bar{x},\, \lambda^1 - \bar{\lambda})\right\| \le \left\|\left(x^1 - \bar{x},\, \lambda^1 - \hat{\lambda}^1\right)\right\| + \left\|\hat{\lambda}^1 - \bar{\lambda}\right\| < q\delta + \delta \le \varepsilon,$$

i.e., (2.34) holds for $k = 1$.

Now suppose that the hypothesis is valid for $k = s$. If $\lambda_0^{s+1} = \nu_{s+1}$, then by (2.27), (2.30), and (2.31), we obtain that

$$\left\|(x^{s+1} - \bar{x},\, \lambda^{s+1} - \bar{\lambda})\right\| \le q\left\|\left(x^s - \bar{x},\, \lambda^s - \hat{\lambda}^s\right)\right\| < q\delta < \varepsilon, \tag{2.38}$$

i.e., (2.34) holds for $k = s + 1$. Furthermore, by (2.28), (2.30), and by the inequality $q < 1$,

$$\left\|\left(x^{s+1} - \bar{x},\, \hat{\lambda}^{s+1} - \bar{\lambda}\right)\right\| \le q\left\|\left(x^s - \bar{x},\, \lambda^s - \hat{\lambda}^s\right)\right\| < q\delta < \delta, \tag{2.39}$$

i.e., (2.33) holds for $k = s+1$. Finally, by (2.38), (2.39), and by the inequality $q \le 1/2$,

$$\left\|\left(x^{s+1} - \bar{x},\, \lambda^{s+1} - \hat{\lambda}^{s+1}\right)\right\| \le \left\|(x^{s+1} - \bar{x},\, \lambda^{s+1} - \bar{\lambda})\right\| + \left\|\hat{\lambda}^{s+1} - \bar{\lambda}\right\| < 2q\delta \le \delta,$$

i.e., (2.32) holds for $k = s + 1$.

On the other hand, if $\lambda_0^{s+1} = \lambda_0^s$, then by (2.13)–(2.15) $\hat{\lambda}^{s+1} = \hat{\lambda}^s$, and by (2.29), (2.30), we have that

$$\left\|\left(x^{s+1} - \bar{x},\, \lambda^{s+1} - \hat{\lambda}^s\right)\right\| \le q\left\|\left(x^s - \bar{x},\, \lambda^s - \hat{\lambda}^s\right)\right\| < q\delta,$$

and hence, by the inequality $q < 1$,

$$\left\|\left(x^{s+1} - \bar{x},\, \lambda^{s+1} - \hat{\lambda}^{s+1}\right)\right\| = \left\|\left(x^{s+1} - \bar{x},\, \lambda^{s+1} - \hat{\lambda}^s\right)\right\| < q\delta < \delta, \tag{2.40}$$

i.e., (2.32) holds for $k = s + 1$. Furthermore, by (2.30), we obtain

$$\left\|\left(x^{s+1} - \bar{x},\, \hat{\lambda}^{s+1} - \bar{\lambda}\right)\right\| = \left\|\left(x^{s+1} - \bar{x},\, \hat{\lambda}^s - \bar{\lambda}\right)\right\| < \delta,$$

i.e., (2.33) holds for $k = s + 1$. Finally, by (2.31), (2.33) for $k = s + 1$, and (2.40), we derive that

$$\left\|(x^{s+1} - \bar{x},\, \lambda^{s+1} - \bar{\lambda})\right\| \le \left\|\left(x^{s+1} - \bar{x},\, \lambda^{s+1} - \hat{\lambda}^{s+1}\right)\right\| + \left\|\hat{\lambda}^{s+1} - \bar{\lambda}\right\| < q\delta + \delta \le \varepsilon,$$

i.e., (2.34) holds for $k = s + 1$. This completes the proof by induction. □

3. Globalization issues. In this section, we discuss some possible ways of globalizing the local scheme presented above. The first approach is based on a generic outer phase steering the iterates toward stationary points. This globalization uses a test of linear decrease for the KKT residual to decide when active-set steps are successful. We also give a specific implementation of this approach along the lines of hybrid semismooth Newton methods for mixed complementarity problems, for which both global convergence and superlinear rate of convergence can be formally proved under reasonable assumptions. The second approach below is based on SQP. It is therefore quite close in spirit to existing algorithms, and can be easily incorporated into them. However, this method may converge to weakly (i.e., not only strongly) stationary points. We do not provide a formal convergence analysis for this method. The reason is that such analysis would primarily concern the study of global convergence properties of standard linesearch SQP algorithms for MPCCs, which is a general issue not related specifically to local algorithms suggested above.

3.1. Hybrid globalization. We next show how our local algorithm can be embedded into any globally convergent scheme. By this we mean that having chosen and fixed some outer-phase global strategy which is guaranteed to produce primal-dual iterates converging to stationary (in some sense) points of MPCC (1.1), the role of our local method is to force quadratic convergence rate under natural assumptions stated above. The key to this construction is the proof that close to a solution with stated properties, the Newton–Lagrange step for (1.19) provides quadratic (hence, also arbitrarily fast linear) decrease for the Fischer–Burmeister residual Φ_{FB} of the KKT system (1.2) for MPCC (1.1).

ALGORITHM 3.1. `Preliminary step.` Fix θ, $q \in (0, 1)$. Set $k = 0$, and choose $x^0 \in \mathbf{R}^n$ and $\lambda^0 = (\lambda_G^0, \lambda_H^0, \lambda_0^0) \in \mathbf{R}^m \times \mathbf{R}^m \times \mathbf{R}$.

`Identification step.` Define the index sets I_G and I_H according to (2.7) and (2.8). If $k = 0$, or if I_G or I_H does not coincide with its counterpart computed at the previous iteration, or if $I_G \cup I_H \neq \{1, \ldots, m\}$, go to `Outer-phase step`.

`Active-set step.` If the current point (x^k, λ^k) was generated by `Outer-phase step`, set $\tilde{k} = k$, store $(x^{\tilde{k}}, \lambda^{\tilde{k}})$, and define $(\mu_G^k)_{I_G}$ and $(\mu_H^k)_{I_H}$ by

$$(\mu_G^k)_i = (\lambda_G^k)_i - \lambda_0^k H_i(x^k), \ i \in I_G \setminus I_H, \tag{3.1}$$

$$(\mu_H^k)_i = (\lambda_H^k)_i - \lambda_0^k G_i(x^k), \ i \in I_H \setminus I_G, \tag{3.2}$$

$$(\mu_G^k)_i = (\lambda_G^k)_i, \ (\mu_H^k)_i = (\lambda_H^k)_i, \ i \in I_G \cap I_H. \tag{3.3}$$

Compute (x^{k+1}, μ^{k+1}) as follows.
• The triple $(x^{k+1}, (\mu_G^{k+1})_{I_G}, (\mu_H^{k+1})_{I_H})$ is generated by the step of Newton–Lagrange method for tightened MP (1.19) from the point $(x^k, (\mu_G^k)_{I_G}, (\mu_H^k)_{I_H})$.
• $(\mu_G^{k+1})_{I_H \setminus I_G} = 0$, $(\mu_H^{k+1})_{I_G \setminus I_H} = 0$.

If there exists $i \in I_G \setminus I_H$ such that $H_i(x^{k+1}) = 0$, or there exists $i \in I_H \setminus I_G$ such that $G_i(x^{k+1}) = 0$, go to `Outer-phase step`. Otherwise, define $\lambda^{k+1} = (\lambda_G^{k+1}, \lambda_H^{k+1}, \lambda_0^{k+1})$ according to (2.9)–(2.12). If the point (x^{k+1}, λ^{k+1}) is well defined and satisfies the condition

$$\left\|\Phi_{FB}(x^{k+1}, \lambda^{k+1})\right\| \leq q \left\|\Phi_{FB}(x^k, \lambda^k)\right\|, \tag{3.4}$$

adjust k by 1, and go to `Identification step`.

`Outer-phase step.` If the current point (x^k, λ^k) was generated by `Active-set step`, set $k = \tilde{k}$ and $(x^k, \lambda^k) = (x^{\tilde{k}}, \lambda^{\tilde{k}})$.

Compute (x^{k+1}, λ^{k+1}) according to the outer-phase strategy. Adjust k by 1, and go to `Identification step`.

Global convergence properties of Algorithm 3.1 are quite transparent. By (3.4), we immediately obtain the following result.

THEOREM 3.1. *Let $\{(x^k, \lambda^k)\}$ be a trajectory generated by Algorithm* 3.1*, and suppose that all the iterates in this trajectory with k large enough are generated by* `Active-set step` *of the algorithm. Then*

$$\Phi_{FB}\left(x^k, \lambda^k\right) \to 0 \quad as \quad k \to \infty. \tag{3.5}$$

In particular, the primal part of any accumulation point of $\{(x^k, \lambda^k)\}$ is strongly stationary for (1.1)*, while the dual part is an associated Lagrange multiplier.*

Except for the case considered in Theorem 3.1, the only other possibility is that all the iterates are generated by the outer-phase strategy (because unsuccessful active-set iterates are eventually discarded). In this case, the method inherits global convergence of the outer strategy. Possible choices of outer strategies will be discussed below.

To prove quadratic convergence of Algorithm 3.1, some work is required. We start with the following dual estimate.

LEMMA 3.2. *Let $\bar{x}$ be a strongly stationary point of MPCC* (1.1)*, and assume that MPCC-LICQ* (1.5) *holds at $\bar{x}$. Let $\bar{\lambda} \in \Lambda(\bar{x})$.*

Then there exists $c > 0$ such that, for each (x^k, λ^k) close enough to $(\bar{x}, \bar{\lambda})$, it holds that

$$\left\|\lambda^k - \hat{\lambda}^k\right\| \le c \operatorname{dist}\left(\lambda^k, \Lambda(\bar{x})\right), \tag{3.6}$$

where $\hat{\lambda}^k$ is defined according to (2.13)–(2.15).

Proof. We argue by contradiction. If $\lambda^k \in \Lambda(\bar{x})$, then by Proposition 1.1 and by (2.13)–(2.15), we have that $\lambda^k = \hat{\lambda}^k$, and (3.6) holds with any $c \ge 0$. Suppose that there exists a sequence $\{(x^k, \lambda^k)\}$ convergent to $(\bar{x}, \bar{\lambda})$ such that $\lambda^k \notin \Lambda(\bar{x})$ $\forall\, k$, and

$$\left\|\lambda^k - \hat{\lambda}^k\right\| / \operatorname{dist}\left(\lambda^k, \Lambda(\bar{x})\right) \to \infty \text{ as } k \to \infty. \tag{3.7}$$

Let $\bar{\lambda}^k$ be the orthogonal projection of λ^k onto $\Lambda(\bar{x})$. Then (3.7) is equivalent to

$$\left\|\lambda^k - \bar{\lambda}^k\right\| / \left\|\lambda^k - \hat{\lambda}^k\right\| \to 0 \text{ as } k \to \infty. \tag{3.8}$$

For each k, we have that

$$\left(\lambda^k - \hat{\lambda}^k\right) / \left\|\lambda^k - \hat{\lambda}^k\right\| = \left(\lambda^k - \bar{\lambda}^k\right) / \left\|\lambda^k - \hat{\lambda}^k\right\| + \left(\bar{\lambda}^k - \hat{\lambda}^k\right) / \left\|\lambda^k - \hat{\lambda}^k\right\|.$$

Observe that in this equality the left-hand side has unit norm and belongs to the "vertical" hyperplane $\lambda_0 = 0$; the first term in the right-hand side tends to 0 as $k \to \infty$, by (3.8); while the second term in the right-hand side belongs to the straight line containing the ray $\Lambda(\bar{x})$ (see Proposition 1.1 and (2.13)–(2.15)), which does not belong to the "vertical" hyperplane. The contradiction is now evident. ∎

THEOREM 3.3. *Let $\{(x^k, \lambda^k)\}$ be a trajectory generated by Algorithm* 3.1*, and suppose that this trajectory has an accumulation point $(\bar{x}, \bar{\lambda})$, with $\bar{x}$ being a strongly stationary point of problem* (1.1) *and $\bar{\lambda}$ being an associated Lagrange multiplier, satisfying MPCC-LICQ* (1.5) *and SOSC* (1.14)*.*

Then the entire trajectory $\{(x^k, \lambda^k)\}$ *converges to* $(\bar{x}, \bar{\lambda})$*, and the rate of convergence is quadratic.*

Proof. Let (x^k, λ^k) be close to $(\bar{x}, \bar{\lambda})$. Furthermore, let (x^{k+1}, λ^{k+1}) be computed by the `Active-set step` of Algorithm 3.1 (this point is correctly defined, according to Theorem 2.4).

We next construct $\bar{\lambda}^{k+1} \in \Lambda(\bar{x})$ satisfying the estimate

$$\left\|\lambda^{k+1} - \bar{\lambda}^{k+1}\right\| = O\left(\left\|\left(x^k - \bar{x}, \lambda^k - \hat{\lambda}^k\right)\right\|^2\right). \tag{3.9}$$

This is done separately for the two possible cases in (2.22).

If the first relation in (2.22) holds, we define $\bar{\lambda}^{k+1} = \hat{\lambda}^k$. In this case, by Proposition 1.1, by (2.13)–(2.15), and by the proximity of λ_0^k to $\bar{\lambda}_0$, we have that $\hat{\lambda}^k \in \Lambda(\bar{x})$. The estimate (3.9) now follows from (2.24). Let the second relation in (2.22) hold. If $\lambda_0^{k+1} = \nu_{k+1}$, then set $\bar{\lambda}^{k+1} = \bar{\lambda}$. In this case, estimate (3.9) follows from (2.25). If $\lambda_0^{k+1} = \lambda_0^k$, then $\lambda_0^k > \nu_{k+1}$, and we define $\bar{\lambda}^{k+1}$ as follows.

If $\lambda_0^k \geq \bar{\nu}$, then set $\bar{\lambda}^{k+1} = \hat{\lambda}^k$. In this case, $\hat{\lambda}^k \in \Lambda(\bar{x})$ according to Proposition 1.1 and (2.13)–(2.15), and estimate (3.9) follows from (2.24).

If $\lambda_0^k < \bar{\nu}$, then $\nu_{k+1} < \lambda_0^k < \bar{\nu}$, and by (2.16), (2.19), and the second relation in (2.22), we have that

$$\left\|\hat{\lambda}^k - \bar{\lambda}\right\| = O\left(\left|\lambda_0^k - \bar{\lambda}_0\right|\right) = O(|\nu_{k+1} - \bar{\nu}|) = O\left(\left\|\left(x^k - \bar{x}, \lambda^k - \hat{\lambda}^k\right)\right\|^2\right). \tag{3.10}$$

Set $\bar{\lambda}^{k+1} = \bar{\lambda}$. Then estimate (3.9) follows from (2.24), (3.10), and from the inequality

$$\left\|\lambda^{k+1} - \bar{\lambda}^{k+1}\right\| \leq \left\|\lambda^{k+1} - \hat{\lambda}^k\right\| + \left\|\hat{\lambda}^k - \bar{\lambda}\right\|.$$

Set $\varphi_{FB}(x, \lambda) = \|\Phi_{FB}(x, \lambda)\|^2$, $x \in \mathbf{R}^n$, $\lambda \in \mathbf{R} \times \mathbf{R}^m \times \mathbf{R}^m$. As is well known, the function φ_{FB} is smooth, and since $(\bar{x}, \bar{\lambda}^{k+1})$ is a global unconstrained minimizer of this function, we obtain the equalities

$$\varphi_{FB}\left(\bar{x}, \bar{\lambda}^{k+1}\right) = 0, \quad \varphi'_{FB}\left(\bar{x}, \bar{\lambda}^{k+1}\right) = 0. \tag{3.11}$$

Recall that, under our assumptions, the error bound (2.6) holds for all (x, λ) close enough to $(\bar{x}, \bar{\lambda})$. Then, by (2.17), (3.9), (3.11), and by Lemma 3.2, we obtain that

$$\begin{aligned}
&\left\|\Phi_{FB}\left(x^{k+1}, \lambda^{k+1}\right)\right\|^2 = \varphi_{FB}\left(x^{k+1}, \lambda^{k+1}\right) \\
&= \varphi_{FB}\left(x^{k+1}, \lambda^{k+1}\right) - \varphi_{FB}\left(\bar{x}, \bar{\lambda}^{k+1}\right) \\
&= \left\langle \varphi'_{FB}\left(\bar{x}, \bar{\lambda}^{k+1}\right), \left(x^{k+1} - \bar{x}, \lambda^{k+1} - \bar{\lambda}^{k+1}\right)\right\rangle + O\left(\left\|\left(x^{k+1} - \bar{x}, \lambda^{k+1} - \bar{\lambda}^{k+1}\right)\right\|^2\right) \\
&= O\left(\left\|\left(x^{k+1} - \bar{x}, \lambda^{k+1} - \bar{\lambda}^{k+1}\right)\right\|^2\right) = O\left(\left\|\left(x^k - \bar{x}, \lambda^k - \hat{\lambda}^k\right)\right\|^4\right) \\
&= O\left(\left(\|x^k - \bar{x}\| + \operatorname{dist}\left(\lambda^k, \Lambda\left(\bar{x}\right)\right)\right)^4\right) = O\left(\left\|\Phi_{NR}(x^k, \lambda^k)\right\|^4\right) = O\left(\left\|\Phi_{FB}(x^k, \lambda^k)\right\|^4\right),
\end{aligned}$$

where the last relation follows from the equivalence of $\|\Phi_{NR}(\cdot)\|$ and $\|\Phi_{FB}(\cdot)\|$ in terms of their growth rates [19].

Evidently, the above relation implies (3.4) for any fixed $q \in (0, 1)$, if (x^k, λ^k) is close enough to $(\bar{x}, \bar{\lambda})$. This implies that the `Active-set step` will be accepted, and Algorithm 3.1 will be further working identically to the (local) Algorithm 2.2. The result now follows from Theorem 2.4. □

One possible choice of the outer-phase algorithm is the elastic mode SQP method discussed in section 3.2 below. Another possibility is to use the merit function φ_{FB} in order to organize the outer phase as well, by means of globalizing the semismooth Newton method applied to the equation $\Phi_{FB}(x, \lambda) = 0$. The resulting algorithm is in the spirit of the method for complementarity problems in [14], and its extension to globalization of an active-set method for mixed complementarity problems in [3, section 3]. One advantage of such a scheme is that one can guarantee the overall monotonicity of the sequence $\{\|\Phi_{FB}(x^k, \lambda^k)\|\}$, and thus no backup safeguards are needed when entering the active-set phase (i.e., global convergence can be proved without such safeguards). That is why we present this scheme as a separate algorithm.

ALGORITHM 3.2. `Preliminary step.` Fix $\theta, q, \varepsilon, \tau \in (0, 1)$, $\delta, \gamma > 0$. Set $k = 0$, and choose $x^0 \in \mathbf{R}^n$ and $\lambda^0 = (\lambda_G^0, \lambda_H^0, \lambda_0^0) \in \mathbf{R}^m \times \mathbf{R}^m \times \mathbf{R}$.

`Identification step.` Define the index sets I_G and I_H according to (2.7) and (2.8). If $k = 0$ or if I_G or I_H does not coincide with its counterpart computed at the previous iteration or if $I_G \cup I_H \neq \{1, \ldots, m\}$, go to `SNM–FB step`.

`Active-set step.` If the current point (x^k, λ^k) was generated by `SNM–FB step`, define $(\mu_G^k)_{I_G}$ and $(\mu_H^k)_{I_H}$ by (3.1)–(3.3). Compute (x^{k+1}, μ^{k+1}) as follows.

• The triple $(x^{k+1}, (\mu_G^{k+1})_{I_G}, (\mu_H^{k+1})_{I_H})$ is generated by the step of Newton–Lagrange method for tightened MP (1.19) from the point $(x^k, (\mu_G^k)_{I_G}, (\mu_H^k)_{I_H})$.

• $(\mu_G^{k+1})_{I_H \setminus I_G} = 0$, $(\mu_H^{k+1})_{I_G \setminus I_H} = 0$.

If there exists $i \in I_G \setminus I_H$ such that $H_i(x^{k+1}) = 0$ or there exists $i \in I_H \setminus I_G$ such that $G_i(x^{k+1}) = 0$, go to `SNM–FB step`. Otherwise, define $\lambda^{k+1} = (\lambda_G^{k+1}, \lambda_H^{k+1}, \lambda_0^{k+1})$ according to (2.9)–(2.12). If the point (x^{k+1}, λ^{k+1}) is well-defined and satisfies the condition (3.4), adjust k by 1, and go to `Identification step`.

`SNM–FB step.` Compute $\Lambda_k \in \partial_B \Phi_{FB}(x^k, \lambda^k)$ and

$$\left(x^{k+1}, \lambda^{k+1}\right) = \left(x^k, \lambda^k\right) - \Lambda_k^{-1} \Phi_{FB}\left(x^k, \lambda^k\right).$$

If this point is well-defined and (3.4) holds, and satisfies the condition, adjust k by 1, and go to `Identification step`.

If x^{k+1} is well-defined but (3.4) does not hold, set $d^k = x^{k+1} - x^k$. If

$$\left\langle \varphi'_{FB}\left(x^k, \lambda^k\right), d^k \right\rangle \leq -\gamma \|d^k\|^\delta,$$

go to `Linesearch step`.

`Gradient step.` Set $d^k = -\varphi'_{FB}(x^k, \lambda^k)$.

`Linesearch step.` Compute the stepsize parameter α_k according to the Armijo rule: $\alpha_k = \tau^s$, where s is the smallest nonnegative integer satisfying

$$\varphi_{FB}\left(\left(x^k, \lambda^k\right) + \tau^s d^k\right) \leq \varphi_{FB}\left(x^k, \lambda^k\right) + \varepsilon \tau^s \left\langle \varphi'_{FB}\left(x^k, \lambda^k\right), d^k \right\rangle.$$

Set $(x^{k+1}, \lambda^{k+1}) = (x^k, \lambda^k) + \alpha_k d^k$, adjust k by 1, and go to `Identification step`.

THEOREM 3.4. *Let $\{(x^k, \lambda^k)\}$ be a trajectory generated by Algorithm* 3.2.

Then any accumulation point $(\bar{x}, \bar{\lambda})$ of $\{(x^k, \lambda^k)\}$ satisfies $\varphi'_{FB}(\bar{x}, \bar{\lambda}) = 0$.

Furthermore, if there exists an infinite subsequence of $\{(x^k, \lambda^k)\}$ such that all the iterates in this subsequence are generated by `Active-set step`*, then* (3.5) *holds. In that case, the primal part of any accumulation point of $\{(x^k, \lambda^k)\}$ is strongly stationary in* (1.1)*, while the dual part is an associated Lagrange multiplier.*

Proof. If there exists an infinite subsequence of $\{(x^k, \lambda^k)\}$ such that all the iterates in this subsequence are generated by `Active-set step` of the algorithm, then (3.5) follows immediately from (3.4) and the fact that the values of φ_{FB} are nonincreasing

along the trajectories of the algorithm. The only other possibility is that the "tail" of the trajectory is generated by the outer-phase algorithm, in which case the result can be obtained extending [14, Theorem 3.1] to the setting of mixed complementarity problems. □

Finally, to obtain the rate of convergence result, one should just repeat the proof of Theorem 3.3, with Algorithm 3.1 replaced by Algorithm 3.2.

THEOREM 3.5. *Let* $\{(x^k, \lambda^k)\}$ *be a trajectory generated by Algorithm* 3.2*, and suppose that this trajectory has an accumulation point* $(\bar{x}, \bar{\lambda})$*, with* $\bar{x}$ *being a strongly stationary point of problem* (1.1) *and* $\bar{\lambda}$ *being an associated Lagrange multiplier, satisfying MPCC-LICQ* (1.5) *and SOSC* (1.14)*.*

Then the entire trajectory $\{(x^k, \lambda^k)\}$ *converges to* $(\bar{x}, \bar{\lambda})$*, and the rate of convergence is quadratic.*

We have thus developed a QP-free algorithm for MPCC, with justified global convergence and quadratic rate of convergence under MPCC-LICQ and the usual SOSC (1.14).

3.2. Globalization based on SQP with linesearch. Introducing slack variables, MPCC (1.1) can be equivalently written in the form

$$\min_{(x,\,y,\,z)} f(x) \quad \text{s.t.} \quad G(x) = y,\ H(x) = z,\ y \geq 0,\ z \geq 0,\ \langle y, z\rangle \leq 0. \tag{3.12}$$

As is well known, this reformulated MPCC has the same properties (MPCC constraint qualifications and SOSC) as (1.1), while being preferable for numerical solution by SQP [6, 7].

We first discuss the outer (elastic mode SQP) phase of the algorithm stated below. When SQP is applied to MPCC, under natural assumptions SQP subproblems can be infeasible, even arbitrarily close to a solution. Thus some kind of constraints relaxation (known as *elastic mode*; see, e.g., [1]) has to be used. Let $u^k = (x^k, y^k, z^k) \in \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{R}^m$ be the current primal iterate, and let $\lambda_0^k \geq 0$ be the current estimate of the Lagrange multiplier corresponding to the last constraint in (3.12). We suggest partial relaxation of SQP constraints, which gives the following subproblems:

$$\begin{array}{ll} \min\limits_{(d,\,t)} & \left\langle f'\left(x^k\right), \xi\right\rangle + \frac{1}{2}\langle \mathcal{H}_k\xi, \xi\rangle + \lambda_0^k\langle \eta, \zeta\rangle + ct \\ \text{s.t.} & -te \leq y^k - G\left(x^k\right) + \eta - G'\left(x^k\right)\xi \leq te, \\ & -te \leq z^k - H\left(x^k\right) + \zeta - H'\left(x^k\right)\xi \leq te, \\ & y^k + \eta \geq 0,\ z^k + \zeta \geq 0,\ \left\langle y^k, z^k\right\rangle + \left\langle z^k, \eta\right\rangle + \left\langle y^k, \zeta\right\rangle \leq 0, \end{array} \tag{3.13}$$

where $d = (\xi, \eta, \zeta) \in \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{R}^m$, $t \in \mathbf{R}$, $\mathcal{H}_k$ is an $n \times n$ positive definite symmetric matrix, $c > 0$ is the (penalty) parameter, and $e \in \mathbf{R}^m$ is the vector of ones.

If (d^k, t_k) is a solution of (3.13), then the next iterate is defined by $u^{k+1} = u^k + \alpha_k d^k$, where $\alpha_k \in (0, 1]$ is the stepsize parameter. Choosing $y^0 \geq 0$ and $z^0 \geq 0$, by the first two constraints in the last line of (3.13), it evidently holds that $y^k \geq 0$ and $z^k \geq 0$ for all k. The last three constraints in (3.13) are then always consistent (for example, $\eta = -y^k$ and $\zeta = 0$ satisfies this part of constraints), while the other constraints in (3.13) are consistent due to the elastic mode. It follows that subproblems (3.13) are always feasible. Furthermore, the objective function in (3.13) is bounded below on the nonempty feasible set. Hence, by the Frank–Wolfe theorem [2, Theorem 2.8.1], the subproblem (3.13) has a solution.

Taking into account that $y^k \geq 0$ and $z^k \geq 0$ for all k, the following penalty function can be used in the linesearch procedure for choosing the stepsize parameter:

for $u = (x,\, y,\, z) \in \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{R}^m$,

$$\varphi_c(u) = f(x) + c\psi(u) + c\langle y,\, z\rangle, \quad \psi(u) = \|(y,z) - (G(x), H(x))\|_\infty\,. \tag{3.14}$$

If $(d^k,\, t^k)$ is a solution of SQP subproblem (3.13), then, by direct computation of directional derivative and by standard argument, it can be seen that d^k is a direction of descent for φ_c, provided c is large enough. This justifies the linesearch procedure along the direction obtained from (3.13).

Now let $\lambda^k = (\lambda_G^k,\, \lambda_H^k,\, \lambda_0^k)$ be the current estimate of the Lagrange multipliers corresponding to inequality constraints in (3.12). It can be easily seen that such λ^k is a natural approximation of Lagrange multipliers of the original MPCC (1.1). Define the index sets I_G and I_H according to (2.7) and (2.8), respectively. Once we have reasons to believe that the index sets I_G and I_H give a correct identification, we shall set the corresponding slacks to zero ($y_{I_G}^k = 0,\, z_{I_H}^k = 0$) and switch to the inner (active-set) phase. We note that identification cannot be correct if $I_G \cup I_H \neq \{1, \ldots, m\}$. Another sign of incorrect identification is when the sets I_G and I_H are not yet stable (i.e., change from one iteration to the next). The inner phase consists in applying SQP to the tightened MP

$$\min_{(x,\,y,\,z)} f(x) \quad \text{s.t.} \quad G(x) = y,\ H(x) = z,\ y_{I_G} = 0,\ z_{I_H} = 0, \tag{3.15}$$

i.e., we find a solution d^k of

$$\begin{aligned} \min{}_d \quad & \left\langle f'\left(x^k\right),\, \xi\right\rangle + \tfrac{1}{2}\langle \mathcal{H}_k\xi,\, \xi\rangle \\ \text{s.t.} \quad & y^k - G\left(x^k\right) + \eta - G'\left(x^k\right)\xi = 0,\ y_{I_G}^k + \eta_{I_G} = 0, \\ & z^k - H\left(x^k\right) + \zeta - H'\left(x^k\right)\xi = 0,\ z_{I_H}^k + \zeta_{I_H} = 0, \end{aligned} \tag{3.16}$$

and set $u^{k+1} = u^k + \alpha_k d^k$, with some $\alpha_k \in (0,\, 1]$. Infeasibility of the active-set subproblem (3.16) is again one of the signs of incorrect identification, in which case we go back to the outer phase. We shall show below that if the subproblem (3.16) is feasible, its solution provides a direction of descent for the same penalty function (3.14) that is used in the outer phase. This justifies incorporating the active-set phase into the global SQP framework.

Having in mind fast local convergence, the matrices $\mathcal{H}_k$ in (3.13) and (3.16) should in some specific sense (i.e., not necessarily on the whole space) "approximate" the Hessians with respect to x of the Lagrangians of (3.12) and (3.15), respectively, at the limiting primal-dual solution. It can be easily checked that both these Hessians coincide with $\frac{\partial^2 \mathcal{L}}{\partial x^2}(\bar{x},\, \bar{\mu})$, where $\bar{x}$ is the primal limiting solution, while $\bar{\mu} = (\bar{\mu}_G,\, \bar{\mu}_H)$ is the part of dual limiting solution, corresponding to the first two constraints in (3.12) and (3.15). (For problem (3.12), $\bar{\mu}$ is an MPCC-multiplier associated with $\bar{x}$, by necessity. For problem (3.15), this is the case as well, if the index sets I_G and I_H are correctly identified and provided $\bar{x}$ is a strongly stationary point of MPCC (1.1) with unique associated MPCC-multiplier.) In order to approximate $\frac{\partial^2 \mathcal{L}}{\partial x^2}(\bar{x},\, \bar{\mu})$, one might need to compute an approximation $\mu^k = (\mu_G^k,\, \mu_H^k)$ of $\bar{\mu}$. Within the inner phase, these estimates can be computed directly as Lagrange multipliers corresponding to the first two constraints in (3.16). Within the outer phase, they can be derived from λ^k by the equalities

$$\mu_G^k = \lambda_G^k - \lambda_0^k H\left(x^k\right), \quad \mu_H^k = \lambda_H^k - \lambda_0^k G\left(x^k\right) \tag{3.17}$$

(note that these formulas do not use identification of active indices).

We proceed to formally state the proposed algorithm.

ALGORITHM 3.3. `Preliminary step.` Fix $\theta, \varepsilon, \tau \in (0, 1)$, and $c > 0$. Set $k = 0$, and choose $u^0 = (x^0, y^0, z^0) \in \mathbf{R}^n \times \mathbf{R}^m_+ \times \mathbf{R}^m_+$ and $\lambda^0 = (\lambda^0_G, \lambda^0_H, \lambda^0_0) \in \mathbf{R}_+ \times \mathbf{R}^m_+ \times \mathbf{R}^m_+$.

`Identification step.` Define the index sets I_G and I_H according to (2.7) and (2.8). If $k = 0$ or if I_G or I_H does not coincide with its counterpart computed at the previous iteration or if $I_G \cup I_H \neq \{1, \ldots, m\}$, go to `Elastic mode SQP step`.

`Active-set step.` If d^{k-1} was generated by `Elastic mode SQP step`, set $\tilde{k} = k$, store $u^{\tilde{k}}$ and $\lambda^{\tilde{k}}$, redefine $u^k = (x^k, y^k, z^k)$ by setting $y^k_{I_G} = 0$, $z^k_{I_H} = 0$, and define $\mu^k = (\mu^k_G, \mu^k_H)$ by (3.1)–(3.3). and

$$\left(\mu^k_G\right)_i = 0,\ i \in I_H \setminus I_G, \quad \left(\mu^k_H\right)_i = 0,\ i \in I_G \setminus I_H. \tag{3.18}$$

Using μ^k, choose an $n \times n$ positive definite symmetric matrix $\mathcal{H}_k$. If (3.16) is infeasible, go to `Elastic mode SQP step`.

Compute $d^k = (\xi^k, \eta^k, \zeta^k)$ as a solution of (3.16) and $\mu^{k+1} = (\mu^{k+1}_G, \mu^{k+1}_H)$ as an associated Lagrange multiplier corresponding to the first two constraints in (3.16). Set $\tilde{x}^{k+1} = x^k + \xi^k$. If there exists $i \in I_G \setminus I_H$ such that $H_i(\tilde{x}^{k+1}) = 0$ or there exists $i \in I_H \setminus I_G$ such that $G_i(\tilde{x}^{k+1}) = 0$, go to `Elastic mode SQP step`. Otherwise, define $\lambda^{k+1} = (\lambda^{k+1}_G, \lambda^{k+1}_H, \lambda^{k+1}_0)$ according to (2.9)–(2.12), with x^{k+1} replaced by $\tilde{x}^{k+1}$, and go to `Linesearch step`.

`Elastic mode SQP step.` If d^{k-1} was generated by `Active-set step`, redefine $u^k = (x^k, y^k, z^k)$ by setting

$$y^k_i = 0 \quad \forall\, i = 1, \ldots, m \text{ such that } y^k_i < 0, \tag{3.19}$$

$$z^k_i = 0 \quad \forall\, i = 1, \ldots, m \text{ such that } z^k_i < 0 \tag{3.20}$$

and if

$$\varphi_c\left(u^k\right) > \varphi_c\left(u^{\tilde{k}}\right), \tag{3.21}$$

then set $k = \tilde{k}$, $u^k = u^{\tilde{k}}$, and $\lambda^k = \lambda^{\tilde{k}}$.

Using $\mu^k = (\mu^k_G, \mu^k_H)$ computed according to (3.17), choose an $n \times n$ positive definite symmetric matrix $\mathcal{H}_k$.

Compute (d^k, t^k) as a solution of (3.13) and $\lambda^{k+1} = (\lambda^{k+1}_G, \lambda^{k+1}_H, \lambda^{k+1}_0)$ as an associated Lagrange multiplier corresponding to inequality constraints in (3.13).

`Linesearch step.` Compute the stepsize parameter α_k according to the Armijo rule: $\alpha_k = \tau^s$, where s is the smallest nonnegative integer satisfying

$$\varphi_c\left(u^k + \tau^s d^k\right) \leq \varphi_c\left(u^k\right) + \varepsilon\tau^s \varphi'_c\left(u^k; d^k\right). \tag{3.22}$$

Set $u^{k+1} = u^k + \alpha_k d^k$, adjust k by 1, and go to `Identification step`.

Observe that the active-set iterations always start with $u^k = (x^k, y^k, z^k)$ satisfying complementarity. Indeed, the SQP iterations in the elastic mode start with $y^k \geq 0$, $z^k \geq 0$ and maintain nonnegativity. Furthermore, active-set iterations start with $(y^k)_{I_G} = 0$, $(z^k)_{I_H} = 0$, where $I_G \cup I_H = \{1, \ldots, m\}$. The only way complementarity can be violated during a sequence of active-set steps is when some component of y or z becomes negative. Obviously, this can happen only for indices which are not in I_G in the case of y and not in I_H in the case of z. Once a component becomes negative, this index is immediately added to the corresponding set (see (2.7), (2.8)), which makes the sets change. In such a case, we get out of the active-set phase, restore nonnegativity (see (3.19), (3.20)), and if such a point breaks monotonicity

of the sequence of the penalty function values (that is, if (3.21) happens), we go back to the last iterate preceding the active-set phase (which was determined to be premature).

We next show that when within `Active-set step` of Algorithm 3.3 the subproblem (3.16) is feasible, the generated direction d^k is of descent for the penalty function (3.14) at u^k, and hence the linesearch procedure along this direction is well-defined.

LEMMA 3.6. *Let $d^k = (\xi^k, \eta^k, \zeta^k)$ and $\mu^{k+1} = (\mu_G^{k+1}, \mu_H^{k+1})$ be computed within* `Active-set step` *of Algorithm* 3.3 *from the primal-dual solution of* (3.16).

Then

$$\varphi_c'\left(u^k; d^k\right) \le -\left\langle \mathcal{H}_k \xi^k, \xi^k \right\rangle - \left(c - \left\|\mu^{k+1}\right\|_1\right) \psi\left(u^k\right). \tag{3.23}$$

In particular, d^k is a direction of descent for φ_c, provided either $\xi^k \neq 0$ or $c > \|\mu^{k+1}\|_1$ and $\psi(u^k) > 0$.

Proof. First note that, as observed above, $y^k \ge 0$, $z^k \ge 0$, because otherwise the index sets would have changed, and we would not be solving (3.16). Furthermore, recall that we set

$$y_{I_G}^k = 0, \quad z_{I_H}^k = 0 \tag{3.24}$$

when the algorithm enters the active-set phase. Moreover, these equalities are preserved within this phase, because the last line in (3.16) implies

$$\eta_{I_G}^k = 0, \quad \zeta_{I_H}^k = 0. \tag{3.25}$$

We thus have that whenever $z_i^k > 0$, it holds that $i \notin I_H$. Since the algorithm can enter the active-set phase only with $I_G \cup I_H = \{1, \ldots, m\}$, we have that $i \in I_G$. Therefore, $y_i^k = 0$ by the first equality in (3.24), and hence $\eta_i^k = 0$ by the first equality in (3.25). This shows that $\langle z^k, \eta^k \rangle = 0$. Analogously, $\langle y^k, \zeta^k \rangle = 0$. It follows that the directional derivative of the term in the definition (3.14) of φ_c that penalizes complementarity violation is equal to zero.

By direct computation of the directional derivative and by standard argument,

$$\psi'\left(u^k; d^k\right) = -\psi\left(u^k\right). \tag{3.26}$$

Furthermore, by the Lagrange optimality conditions for (3.16), it holds that

$$f'\left(x^k\right) + \mathcal{H}_k \xi^k - \left(G'\left(x^k\right)\right)^{\mathrm{T}} \mu_G^{k+1} - \left(H'\left(x^k\right)\right)^{\mathrm{T}} \mu_H^{k+1} = 0, \tag{3.27}$$

$$\left(\mu_G^{k+1}\right)_{I_H \setminus I_G} = 0, \quad \left(\mu_H^{k+1}\right)_{I_G \setminus I_H} = 0. \tag{3.28}$$

Taking again into account the structure of constraints in (3.16), from (3.24)–(3.28) we obtain

$$\begin{aligned}
&\left\langle f'\left(x^k\right), \xi^k \right\rangle = -\left\langle \mathcal{H}_k \xi^k, \xi^k \right\rangle + \left\langle \mu_G^{k+1}, G'\left(x^k\right)\xi^k \right\rangle + \left\langle \mu_H^{k+1}, H'\left(x^k\right)\xi^k \right\rangle \\
&= -\left\langle \mathcal{H}_k \xi^k, \xi^k \right\rangle + \sum_{i \in I_G} \left(\mu_G^{k+1}\right)_i \left(y_i^k - G_i\left(x^k\right) + \eta_i^k\right) \\
&\quad + \sum_{i \in I_H} \left(\mu_H^{k+1}\right)_i \left(z_i^k - H_i\left(x^k\right) + \zeta_i^k\right) \\
&= -\left\langle \mathcal{H}_k \xi^k, \xi^k \right\rangle + \sum_{i \in I_G} \left(\mu_G^{k+1}\right)_i \left(y_i^k - G_i\left(x^k\right)\right) + \sum_{i \in I_H} \left(\mu_H^{k+1}\right)_i \left(z_i^k - H_i\left(x^k\right)\right) \\
&\le -\left\langle \mathcal{H}_k \xi^k, \xi^k \right\rangle + \left\|\mu^{k+1}\right\|_1 \psi\left(u^k\right),
\end{aligned}$$

where $\mu^{k+1} = (\mu_G^{k+1}, \mu_H^{k+1})$. Combining the latter with (3.26), we obtain (3.23). ∎

If $\xi^k = 0$ and $\psi(u^k) = 0$, we obtain from (3.14) and from the constraints of (3.16) that $d^k = 0$ and the point u^k is feasible in (3.12). Furthermore, (3.27) and (3.28) show that x^k is a weakly stationary point of (1.1). Otherwise, the linesearch procedure is well-defined and results in the decrease of the penalty function value with respect to $\varphi_c(u^k)$. Overall, the method generates iterates such that the sequence $\{\varphi_c(u^k)\}$ is nonincreasing, as in standard SQP framework.

4. Numerical examples. In this section, we illustrate behavior of the algorithms discussed above by some numerical examples. In what follows, `Linearization` is the linesearch SQP method with $\mathcal{H}_k$ being the identity matrix, applied to the original problem formulation (without slacks), while `SQP−slacks` is the linesearch SQP method with $\mathcal{H}_k = \frac{\partial^2 \mathcal{L}}{\partial x^2}(x^k, \mu^k)$ applied to the problem formulation with slacks. The first simple choice of $\mathcal{H}_k$ is motivated by robustness (if $\mathcal{H}_k$ is not positive definite, then the subproblems sometimes do not have a solution, while more sophisticated choices of positive definite matrices require complex quasi-Newton implementations). The second choice of $\mathcal{H}_k$ is motivated by its efficiency (when subproblems are solvable). SQP-type methods were all implemented in their basic form, without elastic mode (which corresponds to setting $t = 0$), without any attempts to modify $\mathcal{H}_k$ with respect to the two alternative choices above, and without any tools for avoiding the Maratos effect. While without a doubt important for any professional implementation, all those details have no real bearing for illustrating our proposal for forcing fast local convergence by the active-set phase. Linesearch parameters were chosen as follows: $\varepsilon = 0.1$ and $\tau = 0.5$. We used the simplest update rule for penalty parameters: $c_0 = \|\lambda^1\|_\infty + 1$, and then for each $k = 1, 2, \ldots$, we set $c_k = c_{k-1}$ if $c_{k-1} \geq \|\lambda^{k+1}\|_\infty$, and $c_k = \|\lambda^{k+1}\|_\infty + 1$ otherwise. The other implemented methods are the following. `SNM−FB` is Algorithm 3.2 without `Active-set step` and with parameters $\delta = 2.1$, $\gamma = 10^{-9}$, $\varepsilon = 10^{-4}$, and $\tau = 0.5$. `Linearization+AS` and `SQP−slacks+AS` are the modifications of algorithms `Linearization` and `SQP−slacks`, respectively, supplied with the option of switching to `Active-set step`, implemented as specified in Algorithm 3.1. Finally, `SNM−FB+AS` is precisely Algorithm 3.2. The identification test parameter and the linear decrease parameter were chosen as follows: $\theta = 0.5$, $q = 0.9$. All computations were performed in Matlab environment, with the QP-subproblems solved by the built-in Matlab QP-solver. We used the stopping criterion of the form

$$\left\|\Phi_{FB}\left(x^k, \lambda^k\right)\right\| < 10^{-7}. \tag{4.1}$$

We start with reporting some local runs of the algorithms discussed above for the following example, which is a modified version of `ralph2` in MacMPEC [12]. A separate consideration of this example is due to the fact that it is known to violate MPCC-SOSC, and so we expect that our method may behave better than SQP. The problem `ralph2` is modified by introducing higher-order nonlinear terms, in order to prevent the tendency for finite termination, which is quite common for SQP in the cases of "simple" (affine) constraints.

Example 4.1. The problem

$$\min x_1^2 + x_2^2 - 4x_1x_2 + x_2^3 \text{ s.t. } x_1 + x_2^2/2 \geq 0, x_2 - x_1^2 \geq 0, \left(x_1 + x_2^2/2\right)\left(x_2 - x_1^2\right) \leq 0,$$

has two local solutions $\bar{x}^1 = (0, 0)$ and $\bar{x}^2 = (1, 1)$, the latter being global, both satisfying MPCC-LICQ (1.5). The first solution satisfies piecewise SOSC (1.18) but violates MPCC-SOSC (1.12), while the second satisfies MPCC-SOSC (1.12).

We use the primal starting points close to $\bar{x}^1$ to facilitate convergence to this solution. Selected results for $\lambda_G^0 = 0.01$, $\lambda_H^0 = 0.02$, $\lambda_0^0 = 5$ are presented in Table 4.1.

TABLE 4.1
Example 4.1*, local runs.*

Algorithm	$x^0 = 10^{-3}\times$				
	(10, 1)	(7, 3)	(5, 5)	(3, 7)	(1, 10)
SQP−slacks	11	10	10	10	4
SNM−FB	9	9	10	10	5
SQP−slacks+AS	**3**	**3**	**3**	**3**	**3**
SNM−FB+AS	**4**	**3**	**3**	**3**	**3**

For each run, we report the number of iterations before convergence was declared. Bold-faced numbers mean that convergence was achieved by active-set steps.

These results evidently demonstrate that, in this case, the active-set phase is useful. And this is precisely our message—we do not claim that it should always result in faster convergence than that for some nonmodified method, but it is easy to incorporate, is useful at least *sometimes*, and works as it is supposed to. To give some more validation of our claim, in the rest of this section we present numerical results for global convergence of our algorithms on some small test problems derived from MacMPEC [12]. The set of test problems was obtained as follows. We select all the problems in MacMPEC satisfying the following criteria: they have no more than 10 variables, and they do not have any inequality constraints apart from complementarity constraints (to be consistent with the problem setting of the paper). This makes 37 problems. Furthermore, we ignore the simple bounds (again in order to be consistent with the problem setting of the paper), which of course may sometimes affect the solutions/stationary points of these problems. Finally, `ralph1` suggests two different objective functions, and we use both, labeling the corresponding problems `ralph11` and `ralph12`. Thus, we end up with 38 problems.

We performed the runs of each algorithm from the same randomly generated starting points. Primal starting points were generated in a cubic neighborhood around the solution (a feasible point with the objective function value equal to the optimal value reported in MacMPEC; these points were found in the course of our experiments), with the edge of the cube equal to 20. Dual starting points for equality constraints were generated the same way, but around 0, while for dual starting points corresponding to the complementarity constraints multipliers there was the additional nonnegativity restriction. In the process of collecting information, we disregard the runs when at least one of the QP-employing algorithms fails because of a failure of the QP solver (such failures must be avoided in professional implementations by using elastic mode and modifications of the Hessian, or quasi-Newton updates with appropriate line-search, etc.; in any case these failures are concerned with the outer phase, rather than the use of the active-set step). Thus, we keep generating random starting points until we have 100 that do not cause QP solver failures.

When reporting the results, we count the cases of failure (when convergence was not achieved after 50 steps), the cases of convergence (to KKT points), and provide some details about convergence. We are not concerned whether the obtained KKT point is a local/global solution or not (this, once again, has to do mostly with behavior of the outer phase).

Columns of Tables 4.2 and 4.3 contain average/summarized information on the performance of each algorithm for 100 runs from random starting points. First row of each cell contains average characteristics over successful runs: iteration count, last active-set steps, overall count of active-set steps. Thus the average number of useless

TABLE 4.2
Results on MacMPEC problems.

Problem	Algorithm					
	Linearization	Linearization+AS	SQP−slacks	SQP−slacks+AS	SNM−FB	SNM−FB+AS
bard1	15.2 1	6.4/0.9/0.9 0/87	3.4 0	3.0/0.3/0.3 0/27	14.0 24	7.5/1.0/1.0 19/79
bard1m	14.2 17	11.8/0.2/0.3 5/18	1.3 1	1.6/0.03/0.2 0/3	15.4 49	8.9/0.8/1.2 43/45
dempe	5.5 41	22/5.7/6.6 1/72	8.1 0	8.4/0.6/3.2 0/36	22.9 66	21.9/6.5/7.2 56/44
desilva	13.2 0	10.6/3.0/6.0 0/99	7.0 0	7.8/3.4/5.0 5/67	16.2 41	11.4/5.0/6.6 24/77
ex9.2.1	2.0 4	2.0/0/0 4/0	2.5 8	2.5/0/0 8/0	9.6 26	8.4/0.9/1.0 26/72
ex9.1.4	5.9 0	5.8/0.02/0.02 0/2	2.0 0	2.0/0/0 0/0	30.5 58	30.2/0.4/0.4 59/18
ex9.2.1	4.4 20	4.4/0.07/0.07 19/6	1.7 4	1.7/0.01/0.06 4/1	13.3 36	10.6/0.5/0.9 34/35
ex9.2.4	20.3 0	10.8/0.4/0.4 0/43	2.6 0	2.6/0.02/0.02 0/2	13.1 17	10.7/0.3/0.3 17/22
ex9.2.5	13.9 0	5.8/0.3/0.6 0/32	3.1 0	3.1/0.1/0.3 0/12	19.4 26	16.6/1.0/1.0 17/79
ex9.2.7	4.2 14	4.3/0.06/0.06 13/5	1.8 3	1.7/0.06/0.07 3/6	13.2 36	10.7/0.6/0.9 36/38
ex9.2.8	14.4 0	14.4/0.4/0.4 0/37	2.5 0	2.5/0/0 0/0	10.8 66	7.3/1.0/1.0 55/45
ex9.2.9	8.9 4	8.7/0.6/0.6 4/59	2.5 0	2.5/0.03/0.03 0/3	11.9 28	7.5/1.0/1.0 23/77
flp2	17.6 11	10.2/0.7/1.8 0/69	2.3 0	2.4/0.01/0.5 0/1	11.3 26	12.1/1.0/1.6 20/78
gauvin	4.3 0	3.6/0.3/0.3 0/30	3.2 0	3.0/0.1/0.2 0/13	12.0 18	8.6/1.0/1.1 22/75
jr1	2.9 0	2.9/0.01/0.5 0/1	2.5 0	3.0/0.01/0.7 0/1	8.7 14	7.9/1.0/1.7 5/93
jr2	4.1 0	4.1/0.03/0.2 0/3	4.3 3	4.2/0.1/0.8 8/11	9.1 14	6.2/0.9/1.5 15/79
kth1	5.7 0	5.6/0.09/0.09 0/9	1.8 0	1.8/0.03/0.03 0/3	11.4 48	3.4/1.0/1.0 20/80
kth2	5.8 0	5.7/0.09/0.2 0/9	2.5 0	2.5/0.1/0.3 0/10	9.8 39	8.1/1.0/1.6 25/73
kth3	4.0 0	3.0/0.4/0.4 0/38	3.0 0	2.8/0.4/0.6 0/44	9 12	6.8/1.0/1.4 11/85

active-set steps (eventually disregarded by backup safeguards) for Linearization+AS and SQP−slack+AS equals the difference between the third and the second number. Second row of each cell contains the overall number of failures and those cases when convergence was achieved by active-set steps. Note that what should be compared is the behavior of a given outer-phase algorithm with and without using the AS step. For Linearization and SNM−FB, in many cases using active-set step helps in terms of either robustness, efficiency, or both. SQP−slacks is very efficient by itself, and

TABLE 4.3
Results on MacMPEC problems.

Problem	Algorithm					
	Linearization	Linearization+AS	SQP−slacks	SQP−slacks+AS	SNM−FB	SNM−FB+AS
nash1	7.1 27	8.6/0.3/0.7 12/29	2.4 0	2.3/0.06/0.4 1/5	9.8 32	11.3/1.0/1.7 32/67
outrata31	20.8 3	20.7/0.3/0.3 3/63	11.7 19	13.7/0.2/1.1 23/13	16.5 44	15.6/0.6/1.8 44/28
outrata32	35.3 26	30.3/0.9/1.8 14/43	10.7 30	14.0/0.3/0.9 34/18	17.3 37	16.7/0.4/1.9 33/29
outrata33	25 8	22.4/0.7/6.6 4/48	11.3 10	15.7/0.4/0.7 16/31	17.0 36	17.3/0.1/1.1 32/5
outrata34	39.7 12	23.1/1.6/1.6 9/91	11.0 38	18.2/0.6/0.7 42/35	14.2 29	14.2/0.6/0.6 29/42
ralph11	3.5 26	3.4/0.3/0.3 15/26	1 0	1/0/0 0/0	10 57	2.7/0.9/0.9 7/83
ralph12	6.6 39	6.3/0.5/0.5 21/37	1.4 0	14/0/0 0/0	 100	2.1/1.0/1.0 17/83
ralph2	2.0 0	1.9/0.1/0.1 0/10	3.8 0	2.0/0.3/0.3 0/25	9.2 35	2.7/1.0/1.0 1/99
scholtes1	3.7 57	2.8/0.9/1.0 57/37	7.1 0	7.7/0.1/4.6 1/1	11.9 27	10.7/0.9/3.1 8/87
scholtes2	 100	9.1/7.1/7.4 33/67	11.2 2	10.1/7.1/7.2 0/100	12.9 19	10.1/7.5/7.5 2/98
scholtes3	14.9 0	14.7/0.3/0.9 0/34	3.2 0	3.1/0.3/0.7 0/34	10.5 29	4.3/1.0/1.0 66/33
scholtes5	2.6 0	2.6/0/0.01 0/0	2.3 6	2.6/0/0 6/0	10.0 14	9.8/0/0.08 13/0
scale1	4 88	3.9/0.2/1.7 87/2	3.7 27	3.5/0.1/0.9 21/10	7.9 11	5.1/0.9/1.4 52/44
scale2	35.8 0	11.2/1.0/2.5 0/95	3.6 0	3.6/0.04/0.2 0/4	6.6 1	5.2/1.0/1.0 0/96
scale3	28.5 9	11.6/1.0/2.6 9/91	3.0 13	2.5/0.09/0.5 13/8	7.1 1	4.9/1.0/1.1 0/99
scale4	22.2 95	20.2/0.2/3.2 95/1	2.4 46	2.4/0/1.0 46/0	9.1 69	26.6/0.02/0.9 57/1
scale5	26.4 95	11.1/1.0/2.8 91/9	4.1 0	4.1/0/0.4 0/0	16.7 82	16.0/0.8/0.8 80/16
sl1	13.1 49	13.1/0/0.08 49/0	1.2 0	1.2/0/0 0/0	16.8 66	14.0/0.4/0.4 60/17
stackelberg1	17.2 4	2/1/1 0/100	2.2 0	2/1/1 0/100	6.0 0	2/1/1 0/100

in our implementation, active-set steps sometimes improve or harm it just slightly, being overall comparable. Recall, however, Example 4.1, which puts in evidence that active-set steps do outperform SQP in the case where MPCC-SOSC does not hold. Also, it should be kept in mind that SQP−slacks (just as standard SQP) does not possess fully justified superlinear convergence, unlike active-set steps. One can see from Tables 4.2 and 4.3 that, apart from being intended for the cases of weaker SOSC, active-set step usually does not harm at all, neither robustness nor efficiency. The

number of disregarded active-set steps remains very low. In many cases, active-set step is used just once, on the last iteration, which means that the corresponding outer-phase algorithm without active-set step could not possibly converge faster (usually it converges slower, at least in the cases of `Linearization` and `SNM−FB`). In some cases, the active-set strategy helps seriously.

REFERENCES

[1] M. Anitescu, *On using the elastic mode in nonlinear programming approaches to mathematical programs with complementarity constraints*, SIAM J. Optim., 15 (2005), pp. 1203–1236.

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

[3] A. N. Daryina, A. F. Izmailov, and M. V. Solodov, *Numerical results for a globalized active-set Newton method for mixed complementarity problems*, Comput. Appl. Math., 24 (2005), pp. 293–316.

[4] F. Facchinei, A. Fischer, and C. Kanzow, *On the accurate identification of active constraints*, SIAM J. Optim., 9 (1999), pp. 14–32.

[5] A. Fischer, *Local behavior of an iterative framework for generalized equations with nonisolated solutions*, Math. Program., 94 (2002), pp. 91–124.

[6] R. Fletcher and S. Leyffer, *Solving mathematical programs with equilibrium constraints as nonlinear programs*, Optim. Methods Softw., 19 (2004), pp. 15–40.

[7] R. Fletcher, S. Leyffer, D. Ralph, and S. Scholtes, *Local convergence of SQP methods for mathematical programs with equilibrium constraints*, SIAM J. Optim., 17 (2006), pp. 259–286.

[8] W. W. Hager and M. S. Gowda, *Stability in the presence of degeneracy and error estimation*, Math. Program., 85 (1999), pp. 181–192.

[9] A. F. Izmailov, *Mathematical programs with complementarity constraints: Regularity, optimality conditions, and sensitivity*, Comput. Math. Math. Phys., 44 (2004), pp. 1145–1164.

[10] A. F. Izmailov and M. V. Solodov, *Newton-type methods for optimization problems without constraint qualifications*, SIAM J. Optim., 15 (2004), pp. 210–228.

[11] A. F. Izmailov and M. V. Solodov, *On attraction of Newton-type iterates to multipliers violating second-order sufficiency conditions*, Math. Program., 117 (2009), pp. 271–304.

[12] S. Leyffer, *MacMPEC: AMPL collection of MPECs*, www.mcs.anl.gov/~leyffer/MacMPEC/.

[13] G. H. Lin and M. Fukushima, *Hybrid approach with active set identification for mathematical programs with complementarity constraints*, J. Optim. Theory Appl., 128 (2006), pp. 1–28.

[14] T. De Luca, F. Facchinei, and C. Kanzow, *A theoretical and numerical comparison of some semismooth algorithms for complementarity problems*, Comput. Optim. Appl., 16 (2000), pp. 173–205.

[15] Z.-Q. Luo, J.-S. Pang, and D. Ralph, *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press, Cambridge, UK, 1996.

[16] J. V. Outrata, M. Kocvara, and J. Zowe, *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints: Theory, Applications and Numerical Results*, Kluwer Academic Publishers, Dordrecht, 1998.

[17] D. Ralph, *Sequential quadratic programming for mathematical programs with linear complementarity constraints*, in Computational Techniques and Applications: CTAC95, R. L. May and A. K. Easton, eds., World Scientific, Singapore, 1996, pp. 663–668.

[18] H. Scheel and S. Scholtes, *Mathematical programs with complementarity constraints: Stationarity, optimality and sensitivity*, Math. Oper. Res., 25 (2000), pp. 1–22.

[19] P. Tseng, *Growth behavior of a class of merit functions for the nonlinear complementarity problem*, J. Optim. Theory Appl., 89 (1996), pp. 17–37.

SIAM J. OPTIM.
Vol. 19, No. 3, pp. 1028–1047

A REGULARIZED SMOOTHING NEWTON METHOD FOR SYMMETRIC CONE COMPLEMENTARITY PROBLEMS*

LINGCHEN KONG†, JIE SUN‡, AND NAIHUA XIU†

Abstract. This paper extends the regularized smoothing Newton method in vector complementarity problems to symmetric cone complementarity problems (SCCP), which includes the nonlinear complementarity problem, the second-order cone complementarity problem, and the semidefinite complementarity problem as special cases. In particular, we study strong semismoothness and Jacobian nonsingularity of the total natural residual function for SCCP. We also derive the uniform approximation property and the Jacobian consistency of the Chen–Mangasarian smoothing function of the natural residual. Based on these properties, global and quadratical convergence of the proposed algorithm is established.

Key words. symmetric cone complementarity problem, monotonicity, natural residual function, regularized smoothing method, quadratic convergence

AMS subject classifications. 65K05, 90C33

DOI. 10.1137/060676775

1. Introduction.

We are interested in the following symmetric cone complementarity problem (SCCP): Find vectors $x, y \in \mathcal{J}$ such that

$$x \in K, \quad y = F(x) \in K, \quad \langle x, y \rangle = 0, \tag{1.1}$$

where $\mathcal{J}$ is an n-dimensional real Euclidean space, $\mathcal{A} := (\mathcal{J}, \langle \cdot, \cdot \rangle, \circ)$ is a Euclidean Jordan algebra, K is a symmetric cone in $\mathcal{A}$ (see section 2), and $F : \mathcal{J} \to \mathcal{J}$ is a continuously differentiable mapping. Problem (1.1) includes the semidefinite complementarity problem (SDCP), the second-order cone complementarity problem (SOCCP), and the nonlinear complementarity problem (NCP) as special cases. The SCCP has wide applications; in particular, it may arise from the KKT system of a nonlinear optimization problem. The SCCP has been the focal point of some recent research; see, e.g., [12, 13, 14, 22, 23, 29, 33, 37].

We intend to design an algorithm for SCCPs, which is called the regularized smoothing Newton method. In the setting of NCP, various regularized smoothing methods have been tested, which, in addition to their simplicity of implementation, have the advantage of being able to solve some ill-posed problems. Recently, there are some papers studying the smoothing Newton methods with or without regularization for SOCCP and SDCP; see, e.g., [3, 4, 5, 6, 7, 11, 15, 16, 32, 35]. These papers either address the case of SOCCP or that of SDCP, but to our knowledge, there are no discussions on this type of algorithms under the general framework of SCCP.

In this paper, with the help of the Euclidean Jordan algebra, we analyze the strong semismoothness and Jacobian nonsingularity of a natural residual function

*Received by the editors December 5, 2006; accepted for publication (in revised form) May 19, 2008; published electronically October 22, 2008. The work was partly supported by the National Natural Science Foundation of China (10671010, 70640420143), the Singapore-MIT Alliance, and Grant RP314000-057-112 of National University of Singapore.

http://www.siam.org/journals/siopt/19-3/67677.html

†Department of Applied Mathematics, Beijing Jiaotong University, Beijing 100044, P. R. China (konglchen@126.com, nhxiu@center.njtu.edu.cn).

‡Department of Decision Sciences and Singapore-MIT Alliance, National University of Singapore, Republic of Singapore (jsun@nus.edu.sg).

(the so-called total NR-function). We also show the level-boundedness of the natural merit function of the total NR-function for SCCP under monotonicity and strict feasibility assumptions. We then construct the Chen–Mangasarian smoothing function of the natural residual for SCCP. Our work provides a theoretical and computational framework for solving general nonlinear SCCP. In particular, we derive the uniform approximation property and the Jacobian consistency of this smoothing function. These properties form a basis for establishing quadratic convergence of Newton-type algorithms. Finally, we state a globally and quadratically convergent algorithm for solving monotone SCCP that was originated from a similar algorithm of Hayashi, Yamashita, and Fukushima [15] for SOCCP. Many analytic tools we used are taken from the recent work by Sun and Sun [30], in which the differential properties of the Löwner's operator and spectral functions are studied in the space of Euclidean Jordan algebras.

This paper is organized as follows. In section 2, we briefly describe Euclidean Jordan algebra and some of its properties used in our analysis. We also derive new results on the Jacobian and the Clarke generalized Jacobian of Löwner operators. In section 3, we introduce and characterize the total NR-function for SCCP. In section 4, we present the Chen–Mangasarian smoothing function in the context of SCCPs and discuss its properties. In section 5, we introduce the regularized smoothing Newton method for SCCP and discuss its convergence.

The following notations will be used throughout this paper. Let $\mathcal{X}$ and $\mathcal{Y}$ be two finite dimensional real Euclidean spaces. For a given set $S \subseteq \mathcal{X}$, int$S$ and conv$S$ denote the interior and convex hull of S, respectively. Let dist(a, S) be $\min\{\|a-b\| : b \in S\}$ for $a \in \mathcal{X}$, where $\|\cdot\|$ is the norm on $\mathcal{X}$ induced by the inner product $\langle\cdot,\cdot\rangle$. We write $x \succeq_K y$ (respectively, $x \succ_K y$) to mean $x - y \in K$ (respectively, $x - y \in \text{int}K$) for vectors $x, y \in \mathcal{J}$. Also, we write $A \succeq B$ ($A \succ B$) to mean $A - B$ being positive semidefinite (positive definite) for operators A and B from $\mathcal{J}$ into itself. Let I be the identity operator, i.e., $Ix = x$ for all $x \in \mathcal{J}$. We say that the operator A is invertible (or nonsingular) if the equation $Ax = 0$ has a unique solution $x = 0$. For an operator A, A^T denotes the adjoint operator of A.

2. Preliminaries.

2.1. Euclidean Jordan algebras. We give a brief introduction to Euclidean Jordan algebras. Details on Euclidean Jordan algebras can be found in Koecher's lecture notes [19] and the monograph by Faraut and Korányi [10].

A *Euclidean Jordan algebra (EJA)* is a triple $(\mathcal{J}, \langle\cdot,\cdot\rangle, \circ) \triangleq \mathcal{A}$, where $(\mathcal{J}, \langle\cdot,\cdot\rangle)$ is a real n-dimensional inner product space and $(x, y) \mapsto x \circ y : \mathcal{J} \times \mathcal{J} \to \mathcal{J}$ is a bilinear mapping which satisfies the following conditions:

(i) $x \circ y = y \circ x$ for all $x, y \in \mathcal{J}$,
(ii) $x \circ (x^2 \circ y) = x^2 \circ (x \circ y)$ for all $x, y \in \mathcal{J}$ where $x^2 := x \circ x$,
(iii) $\langle x \circ y, z\rangle = \langle x, y \circ z\rangle$ for all $x, y, z \in \mathcal{J}$.

We call $x \circ y$ the *Jordan product* of x and y. In general, the Jordan product is not associative; i.e., $(x \circ y) \circ z \neq x \circ (y \circ z)$ for all $x, y, z \in \mathcal{J}$. In addition, we assume that there exists an element e (called the *identity* element) such that $x \circ e = e \circ x = x$ for all $x \in \mathcal{J}$. The following are some basic facts about Euclidean Jordan algebras.

- Given a Euclidean Jordan algebra $\mathcal{A}$, define *the set of squares* as $K := \{x^2 : x \in \mathcal{J}\}$. From Theorem III 2.1 in [10], K is a *symmetric cone* in $\mathcal{A}$. In other words, K is a self-dual closed convex cone, and for any two elements $x, y \in \text{int}K$, there exists an invertible linear transformation $\Gamma : \mathcal{J} \to \mathcal{J}$ such that $\Gamma(K) = K$ and $\Gamma(x) = y$.

- For $x \in \mathcal{J}$, let $m := m(x)$ be the smallest positive integer such that the set $\{e, x, x^2, \ldots, x^m\}$ is linearly dependent. Then m is said to be the *degree* of x, which is denoted by $\deg(x)$.
- The *rank* of $\mathcal{A}$ denoted by $\mathrm{rk}(\mathcal{A})$ is defined as $\mathrm{rk}(\mathcal{A}) \triangleq \max\{\deg(x) : x \in \mathcal{J}\}$. Let $\dim(\mathcal{J})$ denote the dimension of $\mathcal{J}$. Obviously, $\mathrm{rk}(\mathcal{A}) \le \dim(\mathcal{J})$.
- An element $c \in \mathcal{J}$ is an *idempotent* if $c^2 = c \neq 0$. An idempotent element is *primitive* if it cannot be written as a sum of two idempotents.
- A *complete system of orthogonal idempotents* in $\mathcal{A}$ is a finite set $\{c_1, c_2, \ldots, c_k\}$ of idempotents where $c_i \circ c_j = 0$ for all $i \neq j$, and $c_1 + c_2 + \cdots + c_k = e$.
- A *Jordan frame* in $\mathcal{A}$ is a complete system of orthogonal primitive idempotents. The number of elements of any Jordan frame equals the positive integer $\mathrm{rk}(\mathcal{A})$.

Example 2.1. Let $\mathbb{R}^n$ denote the space of n-dimensional real column vectors, and $\mathbb{R}^n_+$ be the nonnegative orthant. Consider $\mathbb{R}^n$ with the (usual) inner product and Jordan product defined, respectively, by $\langle x, y\rangle := \sum_{i=1}^n x_i y_i$ and $x \circ y := x * y$, where x_i denotes the ith component of x, etc., and $x * y$ denotes the componentwise product of vectors x and y. Then $(\mathbb{R}^n, \langle\cdot,\cdot\rangle, *)$ forms a Euclidean Jordan algebra with $\mathrm{rk}((\mathbb{R}^n, \langle\cdot,\cdot\rangle, *)) = \dim(\mathbb{R}^n) = n$ and $\mathbb{R}^n_+$ as its cone of squares. The identity element is the n-vector of ones, and the set $\{e_1, e_2, \ldots, e_n\}$ is the unique Jordan frame where e_i is the ith coordinate vector for $i \in \{1, 2, \ldots, n\}$.

Example 2.2. Consider $\mathbb{R}^n (n \ge 2)$ where any x is written as $x = (x_1, x_2^T)^T$ with $x_1 \in \mathbb{R}$ and $x_2 \in \mathbb{R}^{n-1}$. The inner product is the same as usual, and the Jordan product is defined by

$$x \circ y = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \circ \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} := \begin{pmatrix} \langle x, y\rangle \\ x_1 y_2 + y_1 x_2 \end{pmatrix}.$$

Then $\Lambda^n := (\mathbb{R}^n, \langle\cdot,\cdot\rangle, \circ)$ forms a Euclidean Jordan algebra, and its cone of squares (*Lorentz cone or second-order cone*) is specified by $\Lambda^n_+ := \{(x_1, x_2^T)^T : \ x_1 \ge \|x_2\|\}$, where $\|\cdot\|$ denotes the 2-norm. The identity element in Λ^n is $e = \binom{1}{0}$. The set $\{c_1, c_2\}$ is a Jordan frame given by $c_i = \frac{1}{2}\binom{1}{(-1)^i\ \omega}$ for $i = 1, 2$ with any $\omega \in \mathbb{R}^{n-1}$ satisfying $\|\omega\| = 1$.

Example 2.3. Let $\mathbb{S}^n$ denote the set of all $n \times n$ real symmetric matrices with the inner product and Jordan product defined, respectively, by $\langle X, Y\rangle := \mathrm{Trace}(XY)$ and $X \circ Y := (XY + YX)/2$. Thus $(\mathbb{S}^n, \langle\cdot,\cdot\rangle, \circ)$ forms a Euclidean Jordan algebra, and its cone of squares $\mathbb{S}^n_+$ is the set of all positive semidefinite symmetric matrices. The identity element in this setting is the identity matrix E. The set $\{E_1, E_2, \ldots, E_n\}$ is a Jordan frame where E_i is the diagonal matrix with one in the ii-entry and zeros elsewhere for $i \in \{1, 2, \ldots, n\}$.

We now review the following spectral decomposition theorem of an element in a Euclidean Jordan algebra.

THEOREM 2.4. *(Spectral Decomposition Type* II *(Theorem* III.1.2, [10]*)) Let $\mathcal{A}$ be a Euclidean Jordan algebra with rank r. Then for $x \in \mathcal{J}$ there exist a Jordan frame $\{c_1, c_2, \ldots, c_r\}$ and real numbers $\lambda_1(x), \lambda_2(x), \ldots, \lambda_r(x)$ such that*

$$x = \lambda_1(x)c_1 + \lambda_2(x)c_2 + \cdots + \lambda_r(x)c_r. \tag{2.1}$$

The numbers $\lambda_i(x)$ $(i = 1, 2, \ldots, r)$ are the eigenvalues of x. We call (2.1) *the spectral decomposition (or the spectral expansion) of x.*

Note that the Jordan frame $\{c_1, c_2, \ldots, c_r\}$ in (2.1) depends on x. We do not write this dependence explicitly for simplicity of notation. (The same for $\{b_1, b_2, \ldots, b_{\bar{r}}\}$

below.) Let $\sigma(x)$ be the set of all distinct eigenvalues of x. Then $\sigma(x)$ contains at least one element and at most r. For each $\mu_i(x) \in \sigma(x)$, denote $N_i(x) := \{j : \lambda_j(x) = \mu_i(x)\}$ and $b_i \stackrel{\Delta}{=} \sum_{j\in N_i(x)} c_j$. Then the set $\{b_i : \mu_i(x) \in \sigma(x)\}$ is a complete system of orthogonal idempotents, and its uniqueness is guaranteed by Theorem III.1.1 in [10]. Let $\bar{r}$ be the number of elements in this set. We then have the spectral decomposition of type I stated in [10]; i.e.,

$$x = \mu_1(x)b_1 + \mu_2(x)b_2 + \cdots + \mu_{\bar{r}}(x)b_{\bar{r}}.$$

Next, we recall the Peirce decomposition theorem on the space $\mathcal{J}$, where a Jordan frame $\{c_1, c_2, \ldots, c_r\}$ is fixed beforehand. In this case, define the following subspaces

$$J_{ii} \stackrel{\Delta}{=} \{x \in \mathcal{J} : x \circ c_i = x\} \quad \text{and} \quad J_{ij} \stackrel{\Delta}{=} \left\{x \in \mathcal{J} : x \circ c_i = \frac{1}{2}x = x \circ c_j\right\},\ i \neq j, \tag{2.2}$$

for $i, j \in \{1, 2, \ldots, r\}$. In the second-order cone (SOC) case, we have $J_{12} \stackrel{\Delta}{=} \{x \in \mathbb{R}^n : x_1 = 0, \langle x_2, w\rangle = 0\}$, where w is characterized by the Jordan frame as in Example 2.2.

THEOREM 2.5 (Theorem IV.2.1, [10]). *Let $\{c_1, c_2, \ldots, c_r\}$ be a given Jordan frame in a Euclidean Jordan algebra $\mathcal{A}$ of rank r. Then $\mathcal{J}$ is the orthogonal direct sum of spaces J_{ij} $(i \leq j)$. Furthermore,*

(i) $J_{ij} \circ J_{ij} \subseteq J_{ii} + J_{jj}$;
(ii) $J_{ij} \circ J_{jk} \subseteq J_{ik}$, *if* $i \neq k$;
(iii) $J_{ij} \circ J_{kl} = \{0\}$, *if* $\{i, j\} \cap \{k, l\} = \emptyset$.

For each $x \in \mathcal{J}$, we define the *Lyapunov transformation* $L(x) : \mathcal{J} \to \mathcal{J}$ by $L(x)y = x \circ y$ for all $y \in \mathcal{J}$, which is a symmetric operator in the sense that $\langle L(x)y, z\rangle = \langle y, L(x)z\rangle$ for all $y, z \in \mathcal{J}$. Meanwhile, the operator $Q(x) \stackrel{\Delta}{=} 2L^2(x) - L(x^2)$ is called the *quadratic representation* of x. We say two elements $x, y \in \mathcal{J}$ *operator commute* if $L(x)L(y) = L(y)L(x)$. Lemma X.2.2 in [10] gives the following characterization of operator commutativity.

THEOREM 2.6. *Two elements x and y of a Euclidean Jordan algebra of rank r are operator commute if and only if they share a common Jordan frame.*

Thus, for a given Jordan frame $\{c_1, c_2, \ldots, c_r\}$, it is easy to see that c_i, c_j operator commute and $L(c_i)L(c_j) = L(c_j)L(c_i)$ for any $i, j \in \{1, 2, \ldots, r\}$. So do b_i and b_j for any $i, j \in \{1, 2, \ldots, \bar{r}\}$.

2.2. The Jacobian of Löwner operators. We review differentiability and semismoothness of a vector-valued function, which was called the Löwner operator by Sun and Sun [30] in recognition of Löwner's contribution [21]. We also present some new results on the Jacobian and the Clarke generalized Jacobian of the Löwner operator, which are basic and useful in the subsequent analysis.

DEFINITION 2.7. *Let $x = \sum_{j=1}^r \lambda_j(x)c_j$ and $g : \mathbb{R} \to \mathbb{R}$ be a real-valued function. We define the Löwner operator (function) $G : \mathcal{J} \to \mathcal{J}$ as*

$$G(x) \stackrel{\Delta}{=} \sum_{j=1}^{r} g(\lambda_j(x))c_j = g(\lambda_1(x))c_1 + g(\lambda_2(x))c_2 + \cdots + g(\lambda_r(x))c_r. \tag{2.3}$$

When $g(t) = t_+ = \max\{0, t\}(t \in \mathbb{R})$, this becomes the *metric projection operator*

$$P_K(x) \stackrel{\Delta}{=} (\lambda_1(x))_+c_1 + (\lambda_2(x))_+c_2 + \cdots + (\lambda_r(x))_+c_r \tag{2.4}$$

onto the symmetric cone K. Note that $x \in K$ (respectively, $x \in \mathrm{int}K$) if and only if $\lambda_i(x) \geq 0$ (respectively, $\lambda_i(x) > 0$) $(i = 1, 2, \ldots, r)$. For any $x \in K$, we define $\sqrt{x} \stackrel{\Delta}{=} \sum_{j=1}^r \sqrt{\lambda_j(x)}c_j$ for $x \in K$.

We consider the differentiability of the Löwner operator $G(\cdot)$. Suppose that g is differentiable at $\tau_i, i = 1, 2, \ldots, r$. Define *the first divided difference* $g^{[1]}(\tau)$ of g at $\tau \stackrel{\Delta}{=} (\tau_1, \tau_2, \ldots, \tau_r)^T \in \mathbb{R}^r$ as the $r \times r$ symmetric matrix with the ijth entry given by

$$[g^{[1]}(\tau)]_{ij} = [\tau_i, \tau_j]_g \stackrel{\Delta}{=} \begin{cases} \dfrac{g(\tau_i) - g(\tau_j)}{\tau_i - \tau_j} & \text{if } \tau_i \neq \tau_j, \\ g'(\tau_i) & \text{if } \tau_i = \tau_j, \end{cases} \quad i, j = 1, 2, \ldots, r. \tag{2.5}$$

A direct implication of Theorem 3.2 in [30] is the following property of the Jacobian of the Löwner operator $G(\cdot)$.

THEOREM 2.8. *Let $x = \sum_{j=1}^{r} \lambda_j(x) c_j = \sum_{i=1}^{\bar{r}} \mu_i(x) b_i$. Then, $G(\cdot)$ is (continuously) differentiable at x if and only if for each $j \in \{1, 2, \ldots, r\}$, g is (continuously) differentiable at $\lambda_j(x)$. In this case, the Jacobian $\nabla G(x)$ is given by*

$$\nabla G(x) = 2 \sum_{i \neq j, i, j=1}^{r} [\lambda_i(x), \lambda_j(x)]_g L(c_i) L(c_j) + \sum_{i=1}^{r} g'(\lambda_i(x)) Q(c_i) \tag{2.6}$$

or equivalently

$$\nabla G(x) = 2 \sum_{i \neq j,\ i, j=1}^{\bar{r}} [\mu_i(x), \mu_j(x)]_g L(b_i) L(b_j) + \sum_{i=1}^{\bar{r}} g'(\mu_i(x)) Q(b_i). \tag{2.7}$$

Furthermore, $\nabla G(x)$ is a linear and symmetric operator from $\mathcal{J}$ into itself.

As a consequence of Theorem 2.8, we obtain the following result in the case of $\mathrm{rk}(\mathcal{A})$=$\dim(\mathcal{J})$.

COROLLARY 2.9. *Suppose that $\mathrm{rk}(\mathcal{A}) = \dim(\mathcal{J}) = n$ and $x = \sum_{j=1}^{n} \lambda_j(x) c_j = \sum_{i=1}^{\bar{n}} \mu_i(x) b_i$. If $G(\cdot)$ is (continuously) differentiable at x, then it holds that*

$$\nabla G(x) = \sum_{i=1}^{n} g'(\lambda_i(x)) L(c_i) = \sum_{i=1}^{\bar{n}} g'(\mu_i(x)) L(b_i). \tag{2.8}$$

Proof. Since $\mathrm{rk}(\mathcal{A})$=$\dim(\mathcal{J}) = n$, it follows from Theorem 3.5 in [20] that there is a unique Jordan frame $\{c_1, c_2, \ldots, c_n\}$ in $\mathcal{A}$. Thus, through Theorem 2.5, any element $h \in \mathcal{J}$ can be expressed by $h = \sum_{i=1}^{n} h_i c_i$ with $h_i \in \mathbb{R}$ $(i = 1, 2, \ldots, n)$. Therefore,

$$L(c_i)L(c_j)h = L(c_j)L(c_i)h = c_i \circ (c_j \circ h) = \begin{cases} c_i \circ (h_j c_j) = 0 & \text{if } i \neq j, \\ c_i \circ (h_i c_i) = c_i \circ h & \text{if } i = j. \end{cases}$$

This implies that $L(c_i)L(c_j) = 0$ $(i \neq j)$ and $L(c_i)L(c_i) = L(c_i)$ for any $i, j \in \{1, 2, \ldots, n\}$. Hence $Q(c_i) = L(c_i)$. Formula (2.8) is then an implication of Theorem 2.8. ∎

As an application of Corollary 2.9, we consider the Jacobian of the Löwner operator on $\mathbb{R}^n$.

Example 2.10. Suppose that $\mathcal{A} = (\mathbb{R}^n, \langle \cdot, \cdot \rangle, *)$ as in Example 2.1. Let $x = \sum_{i=1}^{n} x_i e_i$. One can easily verify that $L(e_i) = e_i e_i^T = E_i$ $(i = 1, 2, \ldots, n)$. Note that $\mathrm{rk}((\mathbb{R}^n, \langle \cdot, \cdot \rangle, *)) = \dim(\mathbb{R}^n) = n$. It is obvious via Corollary 2.9 that

$$\nabla G(x) = \sum_{i=1}^{n} g'(x_i) L(e_i) = \mathrm{Diag}\{g'(x_1), g'(x_2), \ldots, g'(x_n)\}.$$

The next theorem presents a sufficient condition which guarantees that the Jacobian $\nabla G(x)$ is positive semidefinite (respectively, positive definite). Here $\nabla G(x)$ is called *positive semidefinite* (*respectively, positive definite*) if $\langle h, \nabla G(x)h\rangle \geq 0$ for all $h \in \mathcal{J}$ (respectively, $\langle h, \nabla G(x)h\rangle > 0$ for all $0 \neq h \in \mathcal{J}$).

THEOREM 2.11. *Let $x = \sum_{j=1}^{r} \lambda_j(x)c_j$. If g is (continuously) differentiable at $\lambda_j(x)$ for each $j \in \{1, 2, \ldots, r\}$ and $g'(t) \geq 0$ for all $t \in \mathbb{R}$, then $G(\cdot)$ is (continuously) differentiable at x and the Jacobian $\nabla G(x)$ is positive semidefinite. Moreover, the Jacobian is positive definite if the condition $g'(t) \geq 0$ is replaced by $g'(t) > 0$.*

Proof. Let $x = \sum_{j=1}^{r} \lambda_j(x)c_j$. By Theorem 2.5, any element $h \in \mathcal{J}$ can be expressed by $h = \sum_{i=1}^{r} h_i c_i + \sum_{1\leq k<l\leq r} h_{kl}$ where $h_i \in \mathbb{R}$ $(i = 1, 2, \ldots, r)$ and $h_{kl} \in J_{kl}$ $(1 \leq k < l \leq r)$. Theorem 2.5 also implies that $c_j \circ \sum_{i=1}^{r} h_i c_i = h_j c_j$ and $c_j \circ \sum_{1\leq k<l\leq r} h_{kl} = \frac{1}{2}\left(\sum_{k<j} h_{kj} + \sum_{l>j} h_{jl}\right)$. It therefore holds that

$$c_j \circ h = h_j c_j + \frac{1}{2}\left(\sum_{k=1}^{j-1} h_{kj} + \sum_{l=j+1}^{r} h_{jl}\right), \tag{2.9}$$

where $\sum_{k=1}^{j-1} h_{kj} \triangleq 0$ if $j = 1$ and $\sum_{l=j+1}^{r} h_{jl} \triangleq 0$ if $j = r$. Furthermore, Theorem 2.5 implies that

$$\langle h, c_j \circ (c_i \circ h)\rangle = \langle c_j \circ h, c_i \circ h\rangle = \frac{1}{4}\langle h_{ji}, h_{ji}\rangle = \frac{1}{4}\|h_{ji}\|^2, \ \forall j \neq i, \tag{2.10}$$

and

$$Q(c_j)h = 2c_j \circ (c_j \circ h) - c_j \circ h = h_j c_j, \quad j = 1, 2, \ldots, r. \tag{2.11}$$

Meanwhile, noting that $c_j^2 = c_j$, one has $\langle h, c_j\rangle = \langle c_j \circ h, c_j\rangle = h_j\langle c_j, c_j\rangle = h_j\|c_j\|^2$. Combining this with (2.6), (2.10), and (2.11) and noting that $L(c_j)L(c_i)h = c_j\circ(c_i\circ h)$, one has

$$\begin{aligned}
\langle h, \nabla G(x)h\rangle &= \left\langle h, \sum_{j\neq i,j,i=1}^{r} 2(g^{[1]}(\lambda(x)))_{ji} c_j \circ (c_i \circ h) + \sum_{j=1}^{r} (g^{[1]}(\lambda(x)))_{jj} h_j c_j \right\rangle \\
&= \sum_{j\neq i,j,i=1}^{r} 2(g^{[1]}(\lambda(x)))_{ji}\langle h, c_j \circ (c_i \circ h)\rangle + \sum_{j=1}^{r} (g^{[1]}(\lambda(x)))_{jj} h_j \langle h, c_j\rangle \\
&= \frac{1}{2}\sum_{j\neq i,j,i=1}^{r} (g^{[1]}(\lambda(x)))_{ji}\|h_{ji}\|^2 + \sum_{j=1}^{r} (g^{[1]}(\lambda(x)))_{jj} h_j^2\|c_j\|^2.
\end{aligned}$$

If $g'(t) \geq 0$ $(g'(t) > 0)$ for all $t \in \mathbb{R}$, then by (2.5) we can easily get $(g^{[1]}(\lambda(x)))_{ji} \geq 0$ $((g^{[1]}(\lambda(x)))_{ji} > 0)$ for all $j \neq i, j, i = 1, 2, \ldots, r$. Hence, $\langle h, \nabla G(x)h\rangle \geq 0$ for all $h \in \mathcal{J}$ ($\langle h, \nabla G(x)h\rangle > 0$ for all $0 \neq h \in \mathcal{J}$) through the above equation. ∎

We proceed to study the semismoothness of the Löwner operator $G(\cdot)$. Semismoothness was originally introduced by Mifflin [24] for functionals. Qi and Sun [26] extended the concept of semismoothness to vector-valued functions and developed a systematic theory that employs semismoothness in the analysis of the superlinear convergence of Newton methods for solving systems of nondifferentiable equations.

We briefly review some concepts and results of the semismoothness from [26]. Let $F : C \subseteq \mathcal{X} \to \mathcal{Y}$ be a locally Lipschitz function on an open set C. By Rademacher's theorem, F is almost everywhere differentiable (in the sense of Fréchet) in C. Let D_F

be the set of points where F is differentiable. Let $F'(x)$ denote the *derivative* of F at $x \in D_F$ and let $\nabla F(x)$ denote the *Jacobian* of F at x, which is the adjoint operator of $F'(x)$, in the sense of $\langle y, \nabla F(x)z\rangle = \langle F'(x)y, z\rangle$ for all $y \in \mathcal{X}$ and $z \in \mathcal{Y}$. Then, the *Clarke generalized Jacobian* of F at x is defined by $\partial F(x) \stackrel{\Delta}{=} \mathrm{conv}\{\partial_B F(x)\}$, where $\partial_B F(x) \stackrel{\Delta}{=} \{\lim_{\bar{x}\to x, \bar{x}\in D_F} \nabla F(\bar{x})\}$. Observe that $\partial F(x) = \{\nabla F(x)\}$ if F is smooth (continuously differentiable) at x. We say F is *directionally differentiable* at x along the direction d if

$$F'(x,d) \stackrel{\Delta}{=} \lim_{t\downarrow 0} \frac{F(x+td)-F(x)}{t} \quad \text{exists},$$

where $F'(x,d)$ is called the *directional derivative* of F at x along the direction d; and F is directionally differentiable at x if F is directionally differentiable at x along any direction $d \neq 0$.

Employing the above concepts, we can define (strong) semismoothness of a function F.

DEFINITION 2.12. *A directionally differentiable and locally Lipschitz function $F : C \subseteq \mathcal{X} \to \mathcal{Y}$ is semismooth at $x \in C$ if $V^T d - F'(x;d) = o(\|d\|)$ for any $d \neq 0, d \in \mathcal{X}$ sufficiently small and $V \in \partial F(x+d)$. In particular, if $o(\|d\|)$ can be replaced by $O(\|d\|^2)$, F is called strongly semismooth.*

By combining Theorem 3.3 with Proposition 3.3 in [30], we have the following result on (strong) semismoothness of the Löwner operator $G(\cdot)$.

LEMMA 2.13. *Let $x = \sum_{j=1}^r \lambda_j(x)c_j$. Then $G(\cdot)$ is (strongly) semismooth at x if and only if for each $j \in \{1,2,\ldots,r\}$, g is (strongly) semismooth at $\lambda_j(x)$. In particular, the metric projection operator P_K is strongly semismooth on $\mathcal{J}$.*

We are ready to extend Theorems 2.8 and 2.11 to the case of a semismooth Löwner operator $G(\cdot)$. Let g be semismooth at τ_i $(i = 1,2,\ldots,r)$ and ∂g denote the generalized Jacobian of g in the sense of Clarke. Define *the first generalized divided difference* $g^{[1,\partial]}$ of g at τ as the set of all $r \times r$ symmetric matrices, where the ijth entry $(g^{[1,\partial]}(\tau))_{ij}$ of the element $g^{[1,\partial]}(\tau) \in g^{[1,\partial]}$ is given by a set $\{[\tau_i,\tau_j]_g\}$ for $i,j = 1,2,\ldots,r$, where

$$\{[\tau_i,\tau_j]_g\} = \begin{cases} \left\{\dfrac{g(\tau_i)-g(\tau_j)}{\tau_i-\tau_j}\right\} & \text{if } \tau_i \neq \tau_j, \\ \partial g(\tau_i) & \text{if } \tau_i = \tau_j. \end{cases}$$

THEOREM 2.14. *Let $x \in \mathcal{J}$. Then $G(\cdot)$ is (strongly) semismooth at x if and only if g is (strongly) semismooth at every eigenvalue of x. In this case, the Clarke generalized Jacobian $\partial G(x)$ satisfies*

$$\overline{\partial} G(x) \supseteq \partial G(x) \supseteq \underline{\partial} G(x)$$

with the sets $\overline{\partial} G(x)$ and $\underline{\partial} G(x)$ being given, respectively, by

$$\overline{\partial} G(x) \stackrel{\Delta}{=} \mathrm{conv}\left[\bigcup_{\{c_1,\ldots,c_r\}\in\mathcal{C}(x)} \partial_{c_1,\ldots,c_r} G(x)\right],$$

$$\underline{\partial} G(x) \stackrel{\Delta}{=} \left\{ 2 \sum_{i\neq j,\ i,j=1}^{\bar{r}} [\mu_i(x),\mu_j(x)]_g L(b_i(x))L(b_j) + \sum_{i=1}^{\bar{r}} \partial g(\mu_i(x)) Q(b_i) \right\},$$

where $\mathcal{C}(x)$ is the set consisting of all Jordan frames in the spectral decomposition type

II *of* x, *and* $\partial_{c_1,\ldots,c_r} G(x) \triangleq \{2\sum_{i\neq j,\ i,j=1}^{r}\{[\lambda_i(x),\lambda_j(x)]_g\}L(c_i)L(c_j)+\sum_{i=1}^{r}\partial g(\lambda_i(x))Q(c_i)\}$.

Proof. The first part of the theorem follows from Lemma 2.13. For the second part, we first show $\overline{\partial}G(x) \supseteq \partial G(x)$. By the definitions of $\overline{\partial}G$ and ∂G we need to prove only that $\bigcup_{\{c_1,\ldots,c_r\}\in\mathcal{C}(x)} \partial_{c_1,\ldots,c_r} G(x) \supseteq \partial_B G(x)$. Taking any $V \in \partial_B G(x)$, by the definition of $\partial_B G(x)$ there exists a vector $h \triangleq h(V) \in \mathcal{J}$ such that $V = \lim_{h\to 0, x+h\in D_G} \nabla G(x+h)$. In order to show $V \in \bigcup_{\{c_1,\ldots,c_r\}\in\mathcal{C}(x)} \partial_{c_1,\ldots,c_r} G(x)$, we proceed as follows.

Take any $\{c_1,\ldots,c_r\} \in \mathcal{C}(x)$ and let $x = \sum_{j=1}^{r}\lambda_j(x)c_j = \sum_{i=1}^{\bar{r}}\mu_i(x)b_i(x)$ with $\lambda_1(x) \geq \lambda_2(x) \geq \cdots \geq \lambda_r(x)$ and $\mu_1(x) > \mu_2(x) > \cdots > \mu_{\bar{r}}(x)$. For the above $h \in \mathcal{J}$, let $x+h \triangleq \sum_{j=1}^{r}\lambda_j(x+h)c_j(x+h)$ with $\lambda_1(x+h) \geq \lambda_2(x+h) \geq \cdots \geq \lambda_r(x+h)$. By Theorem 2.4 and the argument after it, in the sense of set convergence (see, e.g., [27]), one has

$$\lim_{h\to 0, x+h\in D_G}\{\lambda(x+h)\} = \{\lambda(x)\},$$

where $\lambda(x+h) \triangleq (\lambda_1(x+h),\lambda_2(x+h),\ldots,\lambda_r(x+h))^T$ and $\lambda(x) \triangleq (\lambda_1(x),\lambda_2(x),\ldots,\lambda_r(x))^T$. Similarly, using "lim sup" in the sense of set convergence, we have

$$\limsup_{h\to 0, x+h\in D_G}\{(c_1(x+h),c_2(x+h),\ldots,c_r(x+h))\} \subseteq \mathcal{C}(x). \tag{2.12}$$

Notice that for any $i,j = 1,2,\ldots,r$,

$$\limsup_{h\to 0, x+h\in D_G}\{[\lambda_i(x+h),\lambda_j(x+h)]_g\}\begin{cases} = \left\{\dfrac{g(\lambda_i(x))-g(\lambda_j(x))}{\lambda_i(x)-\lambda_j(x)}\right\} & \text{if } \lambda_i(x)\neq\lambda_j(x), \\ \subseteq \partial g(\lambda_i(x)) & \text{if } \lambda_i(x)=\lambda_j(x). \end{cases}$$

Thus,

$$\limsup_{h\to 0, x+h\in D_G}\{[\lambda_i(x+h),\lambda_j(x+h)]_g\} \subseteq \{[\lambda_i(x),\lambda_j(x)]_g\}. \tag{2.13}$$

Also, it holds by (2.6) that for $x+h \in D_G$,

$$\begin{aligned}\nabla G(x+h) = {} & 2\sum_{i\neq j,\ i,j=1}^{r}[\lambda_i(x+h),\lambda_j(x+h)]_g L(c_i(x+h))L(c_j(x+h)) \\ & +\sum_{i=1}^{r} g'(\lambda_i(x+h))Q(c_i(x+h)).\end{aligned}$$

This, together with (2.12), (2.13), and the continuity property of $L(x)$ and $Q(x)$, leads to

$$\begin{aligned}&\limsup_{h\to 0, x+h\in D_G}\{\nabla G(x+h)\} \\ &\subseteq \bigcup_{\{c_1,\ldots,c_r\}\in\mathcal{C}(x)}\left\{2\sum_{i\neq j,\ i,j=1}^{r}\{[\lambda_i(x),\lambda_j(x)]_g\}L(c_i)L(c_j)+\sum_{i=1}^{r}\partial g(\lambda_i(x))Q(c_i)\right\}.\end{aligned}$$

Clearly, $V = \lim_{h\to 0, x+h\in D_G}\nabla G(x+h) \in \limsup_{h\to 0, x+h\in D_G}\{\nabla G(x+h)\}$. This implies that $V \in \bigcup_{\{c_1,\ldots,c_r\}\in\mathcal{C}(x)} \partial_{c_1,\ldots,c_r} G(x)$ by the definition of $\partial_{c_1,\ldots,c_r} G(x)$.

We next prove $\partial G(x) \supseteq \underline{\partial} G(x)$. For any $W(x) \in \underline{\partial} G(x)$ with $x = \sum_{i=1}^{\bar{r}} \mu_i(x) b_i(x)$ and $\mu_1(x) > \mu_2(x) > \cdots > \mu_{\bar{r}}(x)$, by the definition of $\underline{\partial} G(x)$ we have

$$W(x) = 2 \sum_{i \neq j,\ i,j=1}^{\bar{r}} [\mu_i(x), \mu_j(x)]_g L(b_i(x)) L(b_j(x)) + \sum_{i=1}^{\bar{r}} w_i Q(b_i(x))$$

with $w_i \in \partial g(\mu_i(x))$ $(i = 1, 2, \ldots, \bar{r})$. Since g is semismooth at $\mu_i(x)$, $\partial g(\mu_i(x))$ and $\partial_B g(\mu_i(x))$ are well-defined and $\dim(\partial g(\mu_i(x))) = 1$. Let D_g be the set consisting of all the differentiable points of g. By Carathéodory theorem (see [27]), for any given $w_i \in \partial g(\mu_i(x))$ there exist $t_i \in [0, 1]$ and two subsequences $\{h_{i,0}\}$ and $\{h_{i,1}\}$ converging to 0 with $\mu_i(x) + h_{i,0}, \mu_i(x) + h_{i,1} \in D_g$ such that

$$w_{i,l_i} \triangleq \lim_{h_{i,l_i} \to 0, \mu_i(x) + h_{i,l_i} \in D_g} g'(\mu_i(x) + h_{i,l_i}) \in \partial_B g(\mu_i(x)), \quad l_i \in \{0, 1\} \tag{2.14}$$

and

$$w_i = t_i w_{i,0} + (1 - t_i) w_{i,1}. \tag{2.15}$$

Based on the set $\{h_{i,l_i} : l_i \in \{0, 1\},\ i = 1, 2, \ldots, \bar{r}\}$, we construct a set $\mathcal{H}$ by

$$\mathcal{H} \triangleq \left\{ \sum_{i=1}^{\bar{r}} h_{i,l_i} b_i(x) \ : \ l_i \in \{0, 1\} \right\}.$$

Let $l \triangleq (l_1, l_2, \ldots, l_{\bar{r}})$ and $h_l \triangleq \sum_{i=1}^{\bar{r}} h_{i,l_i} b_i(x)$ with $l_i \in \{0, 1\}$. Then the set $\mathcal{H}$ can be rewritten as $\mathcal{H} \triangleq \{h_l \ : \ l \in \{0, 1\}^{\bar{r}}\}$, which includes $2^{\bar{r}}$ elements. Meanwhile, for each element h_l, we have

$$x + h_l = \sum_{i=1}^{\bar{r}} (\mu_i(x) + h_{i,l_i}) b_i(x).$$

Moreover, taking sufficiently small $\|h_l\|$, we have $\mu_1(x) + h_{1,l_1} > \mu_2(x) + h_{2,l_2} > \cdots > \mu_{\bar{r}}(x) + h_{\bar{r},l_{\bar{r}}}$, and hence $\mu_i(x + h_l) = \mu_i(x) + h_{i,l_i}$, $b_i(x + h_l) = b_i(x)$ by the uniqueness of spectral decomposition type I. Thus, $x + h_l \in D_G$ by $\mu_i(x) + h_{i,l_i} \in D_g$, and from (2.7) and (2.14) we obtain

$$\begin{aligned}
W_l(x) &\triangleq \lim_{h_l \to 0, x + h_l \in D_G} \nabla G(x + h_l) \\
&= 2 \lim_{h_l \to 0, x + h_l \in D_G} \left[\sum_{i \neq j,\ i,j=1}^{\bar{r}} [\mu_i(x) + h_{i,l_i}, \mu_j(x) + h_{j,l_j}]_g L(b_i(x)) L(b_j(x)) \right. \\
&\qquad \left. + \sum_{i=1}^{\bar{r}} g'(\mu_i(x) + h_{i,l_i}) Q(b_i(x)) \right] \\
&= 2 \sum_{i \neq j,\ i,j=1}^{\bar{r}} [\mu_i(x), \mu_j(x)]_g L(b_i(x)) L(b_j(x)) + \sum_{i=1}^{\bar{r}} w_{i,l_i} Q(b_i(x)).
\end{aligned}$$

Therefore, $W_l(x) \in \partial_B G(x)$ for every $l \in \{0, 1\}^{\bar{r}}$. This implies that

$$\mathcal{W}(x) \triangleq \operatorname{conv}\{W_l(x) \ : \ l \in \{0, 1\}^{\bar{r}}\} \subseteq \partial G(x). \tag{2.16}$$

To prove $W(x) \in \partial G(x)$, it suffices to claim that $W(x) \in \mathcal{W}(x)$. In fact, from expressions of $W(x)$ and $W_l(x)$, it is easy to see that $w \overset{\Delta}{=} (w_1, w_2, \ldots, w_{\bar{r}})$ given above lies in the hypercube whose extreme points are defined by w_{i,l_i} with $l_i \in \{0,1\}, i = 1,2,\ldots,\bar{r}$. Hence, $W(x)$ must be a convex combination of points $\{W_l(x) : l \in \{0,1\}^{\bar{r}}\}$. The proof is completed. □

Remark 2.15. From Theorem 2.14, we easily observe that if $x \in \mathcal{J}$ has distinct eigenvalues $\lambda_1(x), \ldots, \lambda_r(x)$ and $\mathcal{C}(x)$ has an element, then $\underline{\partial} G(x) = \partial G(x) = \overline{\partial} G(x)$. However, if x has the multiple eigenvalues or $\mathcal{C}(x)$ contains many elements, the sets $\underline{\partial} G(x)$, $\partial G(x)$, and $\overline{\partial} G(x)$ may be different as the following example shows.

Let $\mathcal{A} = \Lambda^n$ $(n \geq 3)$ and $x = \sum_{i=1}^{2} \lambda_i c_i$ as in Example 2.2. Take $G(x) = P_K(x)$ where $g(t) = t_+$, and let $x = 0$. Then $\lambda_1 = \lambda_2 = 0$, and there are infinitely many Jordan frames at $x = 0$. The direct calculation yields $\overline{\partial} P_{\Lambda^n_+}(0) = \mathrm{conv}\{4[0,1]L(c_1)L(c_2) + [0,1]Q(c_1) + [0,1]Q(c_2)\}$ and $\underline{\partial} P_{\Lambda^n_+}(0) = \mathrm{conv}\{0, E\}$ where $\mathrm{conv}\{0, E\} = \{\alpha E : 0 \leq \alpha \leq 1\}$. Note that $\partial P_{\Lambda^n_+}(0) = \mathrm{conv}\{0, E, S\}$ by Proposition 4.8 in [15] where S satisfies

$$S = 4 \times \frac{1+\beta}{2} L(c_1)L(c_2) + 0 \times Q(c_1) + Q(c_2),$$

where $\frac{1+\beta}{2} \in [0,1]$. A simple calculation checks that $\underline{\partial} P_{\Lambda^n_+}(0) \subset \partial P_{\Lambda^n_+}(0) \subset \overline{\partial} P_{\Lambda^n_+}(0)$. □

Remark 2.16. Suppose that $\mathrm{rk}(\mathcal{A}) = \dim(\mathcal{J}) = n$ and $x = \sum_{j=1}^{n} \lambda_j(x) c_j = \sum_{i=1}^{\bar{n}} \mu_i(x) b_i$ as in the case of Corollary 2.9. If $G(\cdot)$ is (strongly) semismooth at x, we derive by Theorem 2.14 that $\underline{\partial} G(x) \subseteq \partial G(x) \subseteq \overline{\partial} G(x)$, where

$$\overline{\partial} G(x) = \sum_{i=1}^{n} \partial g(\lambda_i(x)) L(c_i), \quad \underline{\partial} G(x) = \sum_{i=1}^{\bar{n}} \partial g(\mu_i(x)) L(b_i).$$

Especially, when $\mathcal{A} = (\mathbb{R}^n, \langle\cdot,\cdot\rangle, *)$ as in Example 2.10 and $x = \sum_{i=1}^{n} x_i e_i = \sum_{i=1}^{\bar{n}} y_i$ $(\sum_{j \in N(i)} e_j)$, in the similar way to the second part in the proceeding proof, one has $\overline{\partial} G(x) \subseteq \partial G(x)$. Hence,

$$\overline{\partial} G(x) = \partial G(x) = \sum_{i=1}^{n} \partial g(x_i) E_i = \mathrm{Diag}\{\partial g(x_1), \ldots, \partial g(x_n)\},$$

$$\underline{\partial} G(x) = \sum_{i=1}^{\bar{n}} \partial g(y_i) \left(\sum_{j \in N(i)} E_j \right) = \mathrm{Diag}\{\partial g(y_1) I_1, \ldots, \partial g(y_{\bar{n}}) I_{\bar{n}}\},$$

where I_i is the $|N(i)| \times |N(i)|$ identity matrix for $i = 1, 2, \ldots, \bar{n}$. Moreover, letting $G(x) = P_K(x)$ and $x = 0$, we derive

$$\overline{\partial} P_{\mathbb{R}^n_+}(0) = \partial P_{\mathbb{R}^n_+}(0) = \left\{ \begin{pmatrix} [0,1] & 0 & \cdots & 0 \\ 0 & [0,1] & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & [0,1] \end{pmatrix} \right\}$$
$$\supset \{\alpha E, 0 \leq \alpha \leq 1\} = \underline{\partial} P_{\mathbb{R}^n_+}(0). \qquad \square$$

Theorem 2.14 provides an approximation to the Clarke generalized Jacobian, which can be successfully employed to prove the positive semidefiniteness of $\partial G(\cdot)$.

THEOREM 2.17. *If g is (strongly) semismooth at every eigenvalue of x and $\partial g(t) \subseteq \mathbb{R}_+$ ($\partial g(t) \subseteq \mathbb{R}_{++}$) for all $t \in \mathbb{R}$, then the function $G(\cdot)$ is (strongly) semismooth at x, and all the element $V \in \partial G(x)$ are positive semidefinite (positive definite). Moreover, when $\partial g(t) \subseteq \mathbb{R}_{++}$, there exists a scalar $\alpha(x) > 0$ such that $V \succeq \alpha(x) I \succ 0$.*

Proof. By Theorem 2.14 and the definition of $\overline{\partial} G$, it suffices to demonstrate that if $\partial g(t) \subseteq \mathbb{R}_{++}$ for all $t \in \mathbb{R}$, then for any $\{c_1, \ldots, c_r\} \in \mathcal{C}(x)$ and $V \in \partial_{c_1,\ldots,c_r} G(x)$ there is a scalar $\alpha(x)$ such that $V \succeq \alpha(x) I \succ 0$. In this case, one has $x = \sum_{i=1}^r \lambda_i(x) c_i$ and

$$V = 2 \sum_{i \neq j,\ i,j=1}^{r} \nu_{ij}(x) L(c_i) L(c_j) + \sum_{i=1}^{r} \nu_{ii}(x) Q(c_i)$$

with $\nu_{ij}(x) \in \{[\lambda_i(x), \lambda_j(x)]_g\}$. Note that $\partial g(\lambda_j(x)) \subseteq \mathbb{R}_{++}$ is a closed convex set for every $j = 1, \ldots, r$. Taking

$$\alpha(x) \triangleq \min_{i,j} \{[\lambda_i(x), \lambda_j(x)]_g\},$$

by (2.5) and the given assumptions we have $\alpha(x) > 0$ and hence $\alpha(x) I \succ 0$.

We now prove $\bar{V} \triangleq V - \alpha(x) I \succeq 0$, that is, $\langle h, \bar{V} h \rangle \geq 0$ for any $h \in \mathcal{J}$. In fact, from (2.6) with $g(\lambda) = \lambda$, we have

$$I = 2 \sum_{i \neq j,\ i,j=1}^{r} L(c_i) L(c_j) + \sum_{i=1}^{r} Q(c_i).$$

Thus,

$$\bar{V} = 2 \sum_{i \neq j,\ i,j=1}^{r} [\nu_{ij}(x) - \alpha(x)] L(c_i) L(c_j) + \sum_{i=1}^{r} [\nu_{ii}(x) - \alpha(x)] Q(c_i)$$

with $[\nu_{ij}(x) - \alpha(x)] \geq 0$ for any $i, j = 1, \ldots, r$. Modeling the proof of Theorem 2.11, we immediately derive the desired result. □

Furthermore, we can obtain the bounded property of ∂G if ∂g is a bounded set.

COROLLARY 2.18. *Under the assumptions of Theorem 2.17, for any $V \in \partial G(x)$ and scalars $a, b \in \mathbb{R}$ with $a \leq b$, there hold*

(i) *If $\partial g(t) \subseteq [a, b]$, then $aI \preceq V \preceq bI$.*

(ii) *If $\partial g(t) \subseteq (a, b)$ with $a < b$, then $aI \prec V \prec bI$.*

Proof. Let $f(t) = g(t) - at$. Note that $\partial g(t) \subseteq [a, b]$, then $\partial f(t) \subseteq \mathbb{R}_+$. By Theorem 2.17, one has $V - aI \succeq 0$ for any $V \in \partial G(x)$. On the other hand, letting $\bar{f}(t) = bt - g(t)$, one has $\partial \bar{f}(t) \subseteq \mathbb{R}_+$ and hence $bI - V \succeq 0$ for any $V \in \partial G(x)$. These two arguments show part (i). Similarly, we can verify Part (ii). □

3. The total NR-function. For problem (1.1), we define the *natural residual function* (NR-function) $\Phi_{NR} : \mathcal{J} \times \mathcal{J} \to \mathcal{J}$ by

$$\Phi_{NR}(x, y) \triangleq x - P_K(x - y), \tag{3.1}$$

and the *total NR-function* $H_{NR} : \mathcal{J} \times \mathcal{J} \to \mathcal{J} \times \mathcal{J}$ by

$$H_{NR}(x, y) \triangleq \begin{pmatrix} \Phi_{NR}(x, y) \\ F(x) - y \end{pmatrix}. \tag{3.2}$$

Moreover, we specify function $\Psi_{NR} : \mathcal{J} \times \mathcal{J} \to \mathbb{R}$ by

$$\Psi_{NR}(x,y) \triangleq \frac{1}{2}\|H_{NR}(x,y)\|^2 = \frac{1}{2}\|\Phi_{NR}(x,y)\|^2 + \frac{1}{2}\|F(x)-y\|^2. \tag{3.3}$$

From Proposition 6 in [12], we know that

$$\Phi_{NR}(x,y) = 0 \iff x \in K,\ y \in K,\ \langle x,y\rangle = 0.$$

Therefore, problem (1.1) can be reformulated as a nonsmooth system of nonlinear equations: $H_{NR}(x,y) = 0$. Based on this system, we may establish various solution methods, such as nonsmooth and smoothing Newton methods, see, e.g., [9, 17] for the case of NCP. In this paper, our aim is to present a globally and quadratically convergent regularized smoothing Newton method for SCCP. For this purpose, we need to investigate strong semismoothness of H_{NR}, nonsingularity of ∂H_{NR}, and level-boundedness of Ψ_{NR}.

First, we present a result concerning the strong semismoothness of H_{NR}. Since the proof is similar to that of Theorem 4.6 in [15], it is omitted.

THEOREM 3.1. *Let $F : \mathcal{J} \to \mathcal{J}$ be continuously differentiable. Then the function H_{NR} defined by* (3.2) *is semismooth at any $(x,y) \in \mathcal{J} \times \mathcal{J}$. Moreover, if ∇F is locally Lipschitzian, then H_{NR} is strongly semismooth at any $(x,y) \in \mathcal{J} \times \mathcal{J}$.*

Next, we address Clarke generalized Jacobian ∂H_{NR}. Let $T \in \partial H_{NR}(x,y)$ for any $(x,y) \in \mathcal{J} \times \mathcal{J}$. Then T has the following form:

$$T = \begin{pmatrix} I - V & \nabla F(x) \\ V & -I \end{pmatrix}, \tag{3.4}$$

where $V \in \partial P_K(x-y)$. Since ∂t_+ equals $\{1\}$ for $t > 0$, $[0,1]$ for $t = 0$, and $\{0\}$ for $t < 0$, by Corollary 2.18 (i) we have $0 \preceq V \preceq I$.

The nonsingularity result on T is well-known for NCP (see, e.g., [9]) or SOCCP (see, e.g., [11]). In a similar manner, we can easily show that it is still true for SCCP, which does not need a further proof. We say that $F : \mathcal{J} \to \mathcal{J}$ is *monotone* (*strongly monotone*) if for all $(x,y) \in \mathcal{J} \times \mathcal{J}$, $\langle x-y, F(x)-F(y)\rangle \geq 0$ ($\langle x-y, F(x)-F(y)\rangle \geq \varepsilon\|x-y\|^2$ with some $\varepsilon > 0$).

THEOREM 3.2. *Let $F : \mathcal{J} \to \mathcal{J}$ be continuously differentiable, and T be given by* (3.4).

(a) *If F is monotone and $0 \prec V \prec I$, then T is invertible for any $(x,y) \in \mathcal{J} \times \mathcal{J}$.*

(b) *If F is strongly monotone and $0 \preceq V \preceq I$, then T is invertible for any $(x,y) \in \mathcal{J} \times \mathcal{J}$.*

It should be noted that if V is a linear and symmetric operator from $\mathcal{J}$ into itself, then the results in this theorem are still true.

We end this section by stating a well-known result on the boundedness of the level sets $\mathrm{Lev}_\alpha(\Psi_{NR}) \triangleq \{(x,y) \in \mathcal{J} \times \mathcal{J} : \Psi_{NR}(x,y) \leq \alpha\}$ for $\alpha \in \mathbb{R}$, which can ensure that the sequence generated by a descent method for solving $\min \Psi_{NR}(x,y)$ has at least one accumulation point. For more details, see, e.g., [25, 36].

THEOREM 3.3. *Let Ψ_{NR} be defined by* (3.3). *If $F(x)$ is strongly monotone and locally Lipschitzian, then the level sets $Lev_\alpha(\Psi_{NR})$ are bounded for all $\alpha \in \mathbb{R}$.*

4. The Chen–Mangasarian smoothing function. In the previous section, we know that the total NR-function shares the strong semismoothness property because of that of the NR-function. In order to establish the desired smoothing Newton methods, we need to smoothen the NR-function and the total NR-function. This section deals with this issue.

In the literature on NCP, there are two well-known classes of the smoothing functions, i.e., the Chen–Mangasarian smoothing function and the smoothed Fischer–Burmeister function. Recently, they were successfully extended to SDCP [6, 31] and SOCCP [11]. In what follows, we first study an extension of the Chen–Mangasarian smoothing function.

DEFINITION 4.1. *Let $F : \mathcal{X} \to \mathcal{Y}$ be a nondifferentiable function. A function $F_u : \mathcal{X} \to \mathcal{Y}$ with a parameter vector $u \in \mathbb{R}^q_+$ is called a smoothing function of F if it has the following properties:*

(a) *F_u is continuously differentiable for any $u \in \mathbb{R}^q_{++}$;*

(b) *$\lim_{u \downarrow 0} F_u(x) = F(x)$ for any $x \in \mathcal{X}$, where $u \downarrow 0$ means $u \in \mathbb{R}^q_{++}, u \to 0$.*

We say F_u is a uniformly smooth approach function of F if there is a scalar $\kappa > 0$ such that

$$\|F_u(x) - F(x)\| \le \kappa \|u\|, \quad \forall u \in \mathbb{R}^q_{++}, \ \forall x \in \mathcal{X}.$$

Let $\varrho \in \mathbb{R}_{++}$. For NR-function Φ_{NR} as in (3.1), we define the *Chen–Mangasarian smoothing function* $\Phi_\varrho : \mathcal{J} \times \mathcal{J} \to \mathcal{J}$ as

$$\Phi_\varrho(x, y) = x - \Pi_\varrho(x - y), \tag{4.1}$$

where $\Pi_\varrho : \mathcal{J} \to \mathcal{J}$ is specified by $\Pi_\varrho(z) \triangleq \varrho G(z/\varrho)$ and $G \in \mathcal{CM}$. Here, $\mathcal{CM}$ denotes the set of Löwner operators defined by (2.3) with $g : \mathbb{R} \to \mathbb{R}_+$, a continuously differentiable convex function satisfying

$$\lim_{t \to -\infty} g(t) = 0, \ \lim_{t \to \infty} (g(t) - t) = 0 \ \text{ and } \ 0 < g'(t) < 1 \ \text{ for all } \ t \in \mathbb{R}. \tag{4.2}$$

Two known cases of function g are as follows: One is the CHKS function $g(t) = (\sqrt{t^2 + 4} + t)/2$, which was proposed by Chen and Harker [1], Kanzow [18], and Smale [28], and the other is the neural network function $g(t) = \ln(e^t + 1)$, which was used in neural networks [2]. Based on the above definitions and Theorem 2.4, we below derive formulae for Φ_ϱ.

PROPOSITION 4.2. *Let Φ_ϱ be given by (4.1). Then it holds that $\Phi_\varrho(x, y) = x - \varrho \sum_{i=1}^r g(\lambda_i/\varrho) c_i$ where λ_i, c_i $(i = 1, 2, \ldots, r)$ are given by $x - y = \sum_{i=1}^r \lambda_i c_i$. Moveover, one has*

$$\Phi_0(x, y) \triangleq \lim_{\varrho \downarrow 0} \Phi_\varrho(x, y) = x - P_K(x - y).$$

Proof. The first part is trivial. Note that $\lim_{\varrho \downarrow 0} \varrho g(\lambda_i/\varrho) = (\lambda_i)_+$ by (4.2). This derives that $\lim_{\varrho \downarrow 0} \Phi_\varrho(x, y) = x - \sum_{i=1}^r (\lambda_i)_+ c_i$. The second part holds by (2.4). ∎

4.1. Uniformly smooth approximation. The following proposition claims that Φ_ϱ is a uniformly smooth approximation of Φ_{NR}.

PROPOSITION 4.3. *Let Φ_ϱ be given by (4.1). Then, for any scalars $\varrho > \nu \ge 0$, we have*

$$g(0)(\varrho - \nu)e \succeq_K \Phi_\nu(x, y) - \Phi_\varrho(x, y) \succ_K 0, \quad \forall x, y \in \mathcal{J}. \tag{4.3}$$

Proof. In order to prove the proposition, we first consider the case where $\varrho > \nu > 0$. By Proposition 4.2, it is easy to verify $\Phi_\nu(x, y) - \Phi_\varrho(x, y) = \sum_{i=1}^r (\varrho g(\lambda_i/\varrho) - \nu g(\lambda_i/\nu)) c_i$ where λ_i and c_i are given by $x - y = \sum_{i=1}^r \lambda_i c_i$. Noting that for every $i = 1, 2, \ldots, r$, $0 < \varrho g(\lambda_i/\varrho) - \nu g(\lambda_i/\nu) \le g(0)(\varrho - \nu)$ by Lemma 3.1 in [34], we have

$$g(0)(\varrho - \nu)e = \sum_{i=1}^r g(0)(\varrho - \nu) c_i \succeq_K \Phi_\nu(x, y) - \Phi_\varrho(x, y) \succ_K 0. \tag{4.4}$$

This shows that (4.3) holds in the case of $\varrho > \nu > 0$, and that $-\Phi_\nu$ is monotone in $\nu > 0$ with respect to the partial ordering $\succ_K$. Taking $\nu \to 0^+$ in (4.4), one has $g(0)\varrho e \succeq_K \Phi_0(x,y) - \Phi_\varrho(x,y) \succ_K 0$. That is, (4.4) also holds for $\varrho > \nu = 0$. The proof is completed. □

4.2. Differentiability. Let $g : \mathbb{R} \to \mathbb{R}_+$ be a continuously differentiable convex function satisfying (4.2). As in [34] for the setting of NCP and in [15] for the context of SOCCP, we define for any $\varrho > 0$,

$$\gamma_\varrho(t) \triangleq \varrho g(t/\varrho), \tag{4.5}$$

$$\gamma_0(t) \triangleq \lim_{\varrho \downarrow 0} \gamma_\varrho(t) = \max\{0, t\}, \tag{4.6}$$

$$\gamma_0^+(t) \triangleq \lim_{\varrho \downarrow 0} \gamma'_\varrho(t) = \begin{cases} 0 & \text{for } t < 0, \\ g'(0) & \text{for } t = 0, \\ 1 & \text{for } t > 0. \end{cases} \tag{4.7}$$

Let $z = \sum_{j=1}^r \lambda_j(z) c_j(z)$. By $\Pi_\varrho(z) = \varrho G(z/\varrho)$ with $G \in \mathcal{CM}$, Theorem 2.8 leads to

$$\nabla \Pi_\varrho(z) = \nabla G(z/\varrho) = 2 \sum_{i \neq j, i, j=1}^r a_{ij} L(c_i(z)) L(c_j(z)) + \sum_{i=1}^r a_{ii} Q(c_i(z)), \tag{4.8}$$

where for all $i, j = 1, 2, \ldots, r$,

$$a_{ij} = [\lambda_i(z)/\varrho, \lambda_j(z)/\varrho]_g = \begin{cases} \dfrac{g(\lambda_i(z)/\varrho) - g(\lambda_j(z)/\varrho)}{\lambda_i(z)/\varrho - \lambda_j(z)/\varrho} & \text{if } \lambda_i(z) \neq \lambda_j(z), \\ g'(\lambda_i(z)/\varrho) & \text{if } \lambda_i(z) = \lambda_j(z). \end{cases}$$

By (4.5), we have $\gamma'_\varrho(t) = g'(t/\varrho)$. Therefore

$$a_{ij} = [\lambda_i(z), \lambda_j(z)]_{\gamma_\varrho} = \begin{cases} \dfrac{\gamma_\varrho(\lambda_i(z)) - \gamma_\varrho(\lambda_j(z))}{\lambda_i(z) - \lambda_j(z)} & \text{if } \lambda_i(z) \neq \lambda_j(z), \\ \gamma'_\varrho(\lambda_i(z)) & \text{if } \lambda_i(z) = \lambda_j(z). \end{cases} \tag{4.9}$$

By (4.2) and (4.9), one has $0 < a_{ij} < 1$. Thus, by Corollary 2.18 (ii), it holds $I \succ \nabla \Pi_\varrho(z) \succ 0$. In summary, we have the following conclusion.

PROPOSITION 4.4. *The function Π_ϱ is continuously differentiable, and $I \succ \nabla \Pi_\varrho(z) \succ 0$.*

Furthermore, by applying Theorem 2.8 and the chain rule, we immediately obtain the differential property of the Chen–Mangasarian smoothing function Φ_ϱ, which does not need a proof.

PROPOSITION 4.5. *For any $\varrho > 0$, the Chen–Mangasarian smoothing function Φ_ϱ, defined by* (4.1)*, is continuously differentiable and its Jacobian is given by*

$$\nabla \Phi_\varrho(x,y) = \begin{pmatrix} I - \nabla \Pi_\varrho(x-y) \\ \nabla \Pi_\varrho(x-y) \end{pmatrix} = \begin{pmatrix} I - \nabla G((x-y)/\varrho) \\ \nabla G((x-y)/\varrho) \end{pmatrix}.$$

4.3. Jacobian consistency. Like strong semismoothness, Jacobian consistency plays an important role in establishing rapid convergence of smoothing Newton methods. This concept was originally introduced by Chen, Qi, and Sun [8] for variational

inequalities, and was recently used by Hayashi, Yamashita, and Fukushima [15] analyzing the regularized smoothing method for SOCCP, where their Jacobian consistency contains two parameters. We state a more general definition as follows.

DEFINITION 4.6. *Suppose that $F : \mathcal{X} \to \mathcal{Y}$ is a continuous function and ∂F exists. Let F_u be a smoothing function of F. We say that F_u satisfies the Jacobian consistency if*

$$\lim_{u \downarrow 0} \operatorname{dist}(\nabla F_u(x), \partial F(x)) = 0, \quad \text{for any } x \in \mathcal{X}. \tag{4.10}$$

To show Jacobian consistency of the Chen–Mangasarian smoothing function Φ_ϱ, we first look at the function $\Pi_\varrho(z)$. Define $b_{ij} \stackrel{\Delta}{=} \lim_{\varrho \downarrow 0} a_{ij}$ for all $i, j = 1, 2, \ldots, r$. From (4.5)–(4.7) and (4.9), we derive that

$$b_{ij} = \begin{cases} \frac{\gamma_0(\lambda_i(z)) - \gamma_0(\lambda_j(z))}{\lambda_i(z) - \lambda_j(z)} & \text{if } \lambda_i(z) \neq \lambda_j(z), \\ \gamma_0^+(\lambda_i(z)) & \text{if } \lambda_i(z) = \lambda_j(z). \end{cases} \tag{4.11}$$

Obviously, by (4.2), $0 \leq b_{ij} \leq 1$. By the direct calculation, one has

$$\lim_{\varrho \downarrow 0} \nabla \Pi_\varrho(z) = 2 \sum_{i \neq j, i, j=1}^{r} b_{ij} L(c_i(z)) L(c_j(z)) + \sum_{i=1}^{r} b_{ii} Q(c_i(z)). \tag{4.12}$$

Rewriting z as $z = \sum_{i=1}^{\bar{r}} \mu_i(z) b_i(z)$, from Theorem 2.8 we deduce

$$\nabla \Pi_\varrho(z) = 2 \sum_{i \neq j,\ i, j=1}^{\bar{r}} [\mu_i(z), \mu_j(z)]_{\gamma_\varrho} L(b_i(z)) L(b_j(z)) + \sum_{i=1}^{\bar{r}} \gamma_\varrho'(\mu_i(z)) Q(b_i(z)).$$

In a similar manner as in (4.12), we derive that

$$\lim_{\varrho \downarrow 0} \nabla \Pi_\mu(z) = 2 \sum_{i \neq j,\ i, j=1}^{\bar{r}} [\mu_i(z), \mu_j(z)]_{\gamma_0} L(b_i(z)) L(b_j(z)) + \sum_{i=1}^{\bar{r}} \gamma_0^+(\mu_i(z)) Q(b_i(z)).$$

Take $\partial_\Pi^0(z) \stackrel{\Delta}{=} \lim_{\varrho \downarrow 0} \nabla \Pi_\mu(z)$. It follows from Theorem 2.14 that $\partial_\Pi^0(z) \in \underline{\partial} P_K(z) \subseteq \partial P_K(z)$. Summarizing the preceding argument, we have the following.

LEMMA 4.7. *Let $\partial_\Pi^0(z) = \lim_{\varrho \downarrow 0} \nabla \Pi_\varrho(z)$. Then $\partial_\Pi^0(z) \in \partial P_K(z)$ for any $z \in \mathcal{J}$. Thus Π_ϱ satisfies the Jacobian consistency.*

Combining Lemma 4.7 with Proposition 4.5, the Jacobian consistency of Φ_ϱ is immediate.

PROPOSITION 4.8. *Φ_ϱ satisfies the Jacobian consistency.*

In the end of this section, we further consider the function g satisfying both (4.2) and the following

$$g(t) - t = g(-t), \quad \forall t \in \mathbb{R}. \tag{4.13}$$

For instance, $(\sqrt{t^2 + 4} + t)/2$ and $\ln(e^t + 1)$ are such two functions. Can we get a more specific result than Proposition 4.8 in this case? To settle this question, we need the following lemma from [15].

LEMMA 4.9 (Lemma 4.10, [15]). *Let g be a continuously differentiable convex function satisfying* (4.2) *and* (4.13). *Let γ_ϱ, γ_0, and γ_0^+ be given by* (4.5)–(4.7). *Then it holds that*

(a) $\gamma_\varrho(t) - \gamma_0(t) = \gamma_\varrho(-t) - \gamma_0(-t)$ *for any* $t \in \mathbb{R}$;
(b) $\left|\gamma_\varrho'(t) - \gamma_0^+(t)\right| = \left|\gamma_\varrho'(|t|) - \gamma_0^+(|t|)\right|$ *for any* $t \in \mathbb{R}$;
(c) $\left|\gamma_\varrho'(0) - \gamma_0^+(0)\right| = 0 < \left|\gamma_\varrho'(t_2) - \gamma_0^+(t_2)\right| \le \left|\gamma_\varrho'(t_1) - \gamma_0^+(t_1)\right|$ *for any* $t_i \in \mathbb{R}(i = 1, 2)$ *such that* $0 < |t_1| \le |t_2|$.

For $z = \sum_{j=1}^r \lambda_j(z)c_j(z)$, let $N(z)$ be the index set specified by $N(z) \triangleq \{i : \lambda_i(z) \neq 0\}$. Define the function $\tilde{\lambda} : \mathcal{J} \to \mathbb{R}_+$ by

$$\tilde{\lambda}(z) \triangleq \begin{cases} \min_{i\in N(z)} |\lambda_i(z)| & \text{for } N(z) \neq \emptyset, \\ 0 & \text{for } N(z) = \emptyset. \end{cases} \tag{4.14}$$

Obviously, $\tilde{\lambda}(z) = 0$ if and only if $z = 0$. When $z \neq 0$, by (4.5) and the continuous differentiability of g, there is a scalar $\varsigma \in (0, \tilde{\lambda}(z))$ such that $\gamma_\varrho'(\varsigma) = \frac{\gamma_\varrho(\tilde{\lambda}(z)) - \gamma_\varrho(0)}{\tilde{\lambda}(z)}$; meanwhile, noting that g is convex, one has $\gamma_\varrho'(t) \le \frac{\gamma_\varrho(\tilde{\lambda}(z)) - \gamma_\varrho(0)}{\tilde{\lambda}(z)}$ for any $t \in (0, \varsigma)$. So, in the case of $z \neq 0$, there exists a positive integer l such that $\frac{1}{2^l}\tilde{\lambda}(z) \in (0, \varsigma)$.

Based on the preceding argument, we define the function $\lambda^\star : \mathcal{J} \to \mathbb{R}_+$ by

$$\lambda^\star(z) \triangleq \begin{cases} \frac{1}{2^l}\tilde{\lambda}(z) & \text{for } N(z) \neq \emptyset, \\ 0 & \text{for } N(z) = \emptyset, \end{cases} \tag{4.15}$$

where l is the smallest positive integer such that

$$\gamma_\varrho'\left(\frac{1}{2^l}\tilde{\lambda}(z)\right) \le \frac{\gamma_\varrho\left(\tilde{\lambda}(z)\right) - \gamma_\varrho(0)}{\tilde{\lambda}(z)}. \tag{4.16}$$

Then $\lambda^\star(z)$ is well-defined and $0 < \lambda^\star(z) < \tilde{\lambda}(z)$. Thus, it holds by Lemma 4.9 (c) that

$$\begin{aligned} \left|\gamma_\varrho'(\lambda_i(z)) - \gamma_0^+(\lambda_i(z))\right| &\le \left|\gamma_\varrho'\left(\tilde{\lambda}(z)\right) - \gamma_0^+\left(\tilde{\lambda}(z)\right)\right| \\ &\le \left|\gamma_\varrho'\left(\lambda^\star(z)\right) - \gamma_0^+\left(\lambda^\star(z)\right)\right|, \ i = 1, 2, \ldots, r. \end{aligned} \tag{4.17}$$

Now we are ready to claim that $\Pi_\varrho(z)$ not only satisfies the Jacobian consistency but also has the stronger Jacobian property.

THEOREM 4.10. *Let* $\partial_\Pi^0(z) = \lim_{\varrho\downarrow 0} \nabla\Pi_\varrho(z)$. *Suppose* g *is a continuously differentiable convex function satisfying* (4.2) *and* (4.13). *Let* $\gamma_\varrho, \gamma_0, \gamma_0^+$, *and* $\lambda^\star$ *be given by* (4.5)–(4.7) *and* (4.15), *respectively. Then there exists a scalar* $\bar{M} > 0$ *such that*

$$\left\|\nabla\Pi_\varrho(z) - \partial_\Pi^0(z)\right\| \le \bar{M}\left|\gamma_\varrho'\left(\lambda^\star(z)\right) - \gamma_0^+\left(\lambda^\star(z)\right)\right|, \quad \forall \varrho \in \mathbb{R}_{++}, \ \forall\, z \in \mathcal{J}.$$

Proof. Let $z = \sum_{j=1}^r \lambda_j(z)c_j(z)$. Then from (4.8) and (4.12) we obtain

$$\nabla\Pi_\varrho(z) - \partial_\Pi^0(z) = 2\sum_{i\neq j, i,j=1}^r (a_{ij} - b_{ij})L(c_i(z))L(c_j(z)) + \sum_{i=1}^r (a_{ii} - b_{ii})Q(c_i(z)).$$

To prove the theorem, it is enough to show $|a_{ij} - b_{ij}| \le |\gamma_\varrho'(\lambda^\star(z)) - \gamma_0^+(\lambda^\star(z))|$ for every $i, j = 1, 2, \ldots, r$. We consider below two cases.

Case (i): $0 = \lambda_i(z) < |\lambda_j(z)|$. By (4.9) and (4.11), the direct calculation yields

$$\begin{aligned}|a_{ij} - b_{ij}| &= \left| \frac{\gamma_\varrho(0) - \gamma_\varrho(\lambda_j(z))}{0 - \lambda_j(z)} - \frac{\gamma_0(0) - \gamma_0(\lambda_j(z))}{0 - \lambda_j(z)} \right| \\ &= \left| \frac{\gamma_\varrho(\lambda_j(z)) - \gamma_\varrho(0)}{\lambda_j(z)} - 1 \right| \\ &= 1 - \frac{\gamma_\varrho(\lambda_j(z)) - \gamma_\varrho(0)}{\lambda_j(z)} \\ &\le 1 - \gamma'_\varrho\left(\lambda^\star(z)\right) \\ &= \left|\gamma'_\varrho\left(\lambda^\star(z)\right) - \gamma_0^+\left(\lambda^\star(z)\right)\right|,\end{aligned}$$

where the second equality follows from the fact $\frac{\gamma_0(0)-\gamma_0(\lambda_j(z))}{0-\lambda_j(z)} = 1$ by (4.6), the third one from $0 < \frac{\gamma_\varrho(\lambda_j(z))-\gamma_\varrho(0)}{\lambda_j(z)} = \frac{g(\lambda_j(z)/\varrho)-g(0)}{\lambda_j(z)/\varrho} < 1$ by (4.2), the inequality from (4.15), and the last equality from $\gamma_0^+\left(\lambda^\star(z)\right) = 1$ by (4.15) and (4.7).

Case (ii): Otherwise, one has $|a_{ij} - b_{ij}| \le \left|\gamma'_\varrho\left(\tilde{\lambda}(z)\right) - \gamma_0^+\left(\tilde{\lambda}(z)\right)\right|$, whose proof is perfectly similar to that in [15] and is omitted for brevity. □

5. Regularized smoothing function and algorithm. Based on the proceeding results, we shall develop the Chen–Mangasarian class of regularized smoothing functions for SCCP, and derive the regularized smoothing Newton method for solving the monotone SCCP.

For the given F in (1.1) and a parameter $\varepsilon > 0$, we define a new function $F_\varepsilon : \mathcal{J} \to \mathcal{J}$ as

$$F_\varepsilon(x) \triangleq F(x) + \varepsilon x. \tag{5.1}$$

Again, define functions $H_{\varrho,\varepsilon} : \mathcal{J} \times \mathcal{J} \to \mathcal{J} \times \mathcal{J}$ and $\Psi_{\varrho,\varepsilon} : \mathcal{J} \times \mathcal{J} \to \mathbb{R}$ by

$$H_{\varrho,\varepsilon}(x,y) \triangleq \begin{pmatrix} \Phi_\varrho(x,y) \\ F_\varepsilon(x) - y \end{pmatrix}, \tag{5.2}$$

$$\Psi_{\varrho,\varepsilon}(x,y) \triangleq \frac{1}{2}\|H_{\varrho,\varepsilon}(x,y)\|^2 = \frac{1}{2}\|\Phi_\varrho(x,y)\|^2 + \frac{1}{2}\|F_\varepsilon(x) - y\|^2. \tag{5.3}$$

Then, $H_{\varrho,\varepsilon}$ is a smoothing approximation of the regularized SCCP involving F_ε with $\varepsilon > 0$. Obviously, if F is monotone, then F_ε is strongly monotone for any $\varepsilon > 0$. In addition, if F is also locally Lipschitzian, then $\Psi_{\varrho,\varepsilon}$ is level-bounded for any $\varrho \ge 0$ and $\varepsilon > 0$ via Theorem 3.3.

The proposed method applies the Newton algorithm to the system $H_{\varrho,\varepsilon}(x,y) = 0$ with ϱ and ε properly adjusted at each iteration, so that a solution of the original SCCP is eventually obtained by taking the limits as $\varrho \downarrow 0$ and $\varepsilon \downarrow 0$.

For this purpose, we deal with $H_{\varrho,\varepsilon}$. From Proposition 4.5, we obtain

$$\nabla H_{\varrho,\varepsilon}(x,y) = \begin{pmatrix} I - \nabla\Pi_\varrho(x-y) & \nabla F(x) + \varepsilon I \\ \nabla\Pi_\varrho(x-y) & -I \end{pmatrix}, \tag{5.4}$$

where $\nabla\Pi_\varrho(\cdot)$ is specified by (4.8).

From (5.4) and Proposition 4.4, one can easily get the nonsingularity of $\nabla H_{\varrho,\varepsilon}$. The proof is omitted.

THEOREM 5.1. *Let $F : \mathcal{J} \to \mathcal{J}$ be continuously differentiable. For parameters $\varrho > 0$ and $\varepsilon > 0$, let $\Phi_\varrho(x,y), F_\varepsilon(x)$, and $H_{\varrho,\varepsilon}(x,y)$ be defined by* (4.1), (5.1), *and*

(5.2), *respectively. If* F *is monotone, then* $\nabla H_{\varrho,\varepsilon}$, *given by* (5.4), *is invertible for any* $(x,y) \in \mathcal{J} \times \mathcal{J}$.

In view of (5.4), we also deduce the Jacobian consistency of $H_{\varrho,\varepsilon}$.

THEOREM 5.2. *Let* $F : \mathcal{J} \to \mathcal{J}$ *be continuously differentiable. For parameters* $\varrho > 0$ *and* $\varepsilon > 0$, *let* $\Phi_\varrho(x,y), F_\varepsilon(x)$, *and* $H_{\varrho,\varepsilon}(x,y)$ *be defined by* (4.1), (5.1), *and* (5.2), *respectively. Then* $H_{\varrho,\varepsilon}$ *satisfies the Jacobian consistency.*

Proof. It holds by (5.4) and $\partial^0_\Pi(z) = \lim_{\varrho \downarrow 0} \nabla \Pi_\varrho(z)$ that

$$\partial^0_\Pi H(x,y) \triangleq \lim_{(\varrho,\varepsilon)\downarrow(0,0)} \nabla H_{\varrho,\varepsilon}(x,y) = \begin{pmatrix} I - \partial^0_\Pi(x-y) & \nabla F(x) \\ \partial^0_\Pi(x-y) & -I \end{pmatrix}. \tag{5.5}$$

This implies from (3.4) and Lemma 4.7 that $\partial^0_\Pi H(x,y) \in \partial H_{NR}(x,y)$ for any $(x,y) \in \mathcal{J} \times \mathcal{J}$. The desired conclusion holds obviously. □

Furthermore, applying Theorems 4.10 and 5.2, we estimate the upper bound of the distance $\operatorname{dist}(\nabla H_{\varrho,\varepsilon}(x,y), \partial H_{NR}(x,y))$.

THEOREM 5.3. *Let* $F : \mathcal{J} \to \mathcal{J}$ *be continuously differentiable, and* g *be a continuously differentiable convex function satisfying* (4.2) *and* (4.13). *Suppose* γ_ϱ, γ_0, *and* γ_0^+ *are given by* (4.5)–(4.7), *and let* $\lambda^\star$ *be defined by* (4.15). *Then, there exists a scalar* $M > 0$ *such that*

$$\operatorname{dist}(\nabla H_{\varrho,\varepsilon}(x,y), \partial H_{NR}(x,y)) \le M(|\gamma'_\varrho(\lambda^\star(x-y)) - \gamma_0^+(\lambda^\star(x-y))| + \varepsilon),$$

for any $\varrho > 0, \varepsilon \ge 0$ *and any* $(x,y) \in \mathcal{J} \times \mathcal{J}$.

Proof. By (5.4), (5.5), and the fact $\partial^0_\Pi H(x,y) \in \partial H_{NR}(x,y)$, one has for any $\varrho > 0, \varepsilon \ge 0$, and any $(x,y) \in \mathcal{J} \times \mathcal{J}$,

$$\begin{aligned} \operatorname{dist}(\nabla H_{\varrho,\varepsilon}(x,y), \partial H_{NR}(x,y)) &\le ||\nabla H_{\varrho,\varepsilon}(x,y) - \partial^0_\Pi H(x,y)|| \\ &\le \tilde{M}(\|\nabla \Pi_\varrho(x-y)) - \partial^0_\Pi(x-y))\| + \varepsilon) \\ &\le \tilde{M}(\bar{M}|\gamma'_\varrho(\lambda^\star(x-y)) - \gamma_0^+(\lambda^\star(x-y))| + \varepsilon), \end{aligned}$$

where $\tilde{M}$ in the second inequality is a positive scalar, the third follows from Theorem 4.10. The desired holds immediately. □

In the end of this paper, we describe an algorithm which is a word-for-word extension of the one by Hayashi, Yamashita, and Fukushima [15] for SOCCP, and state the corresponding convergence theorem, which can be obtained by Theorems 5.1–5.3 and following the proof of Theorem 4.13 in [15].

ALGORITHM Set $w \triangleq (x,y)$ and $w^{(k)} \triangleq (x^{(k)}, y^{(k)})$. Choose $\eta, \rho \in (0,1), \bar{\eta} \in (0, \eta], \sigma \in (0, 1/2), \kappa > 0$, and $\hat{\kappa} > 0$.

Step 0 Choose $w^{(0)} \in \mathcal{J} \times \mathcal{J}$ and $\beta_0 \in (0, \infty)$. Let $\varrho_0 \triangleq \|H_{NR}(w^{(0)})\|$ and $\varepsilon_0 \triangleq \|H_{NR}(w^{(0)})\|$. Set $k \triangleq 0$.

Step 1 Terminate if $\|H_{NR}(w^{(k)})\| = 0$.

Step 2

Step 2.0 Set $v^{(0)} \triangleq w^{(0)}$ and $j \triangleq 0$.

Step 2.1 Find a vector $\hat{d}^{(j)}$ such that

$$H_{\varrho_k,\varepsilon_k}\left(v^{(j)}\right) + \nabla H_{\varrho_k,\varepsilon_k}\left(v^{(j)}\right)^T \hat{d}^{(j)} = 0.$$

Step 2.2 If $\|H_{\varrho_k,\varepsilon_k}(v^{(j)} + \hat{d}^{(j)})\| \le \beta_k$, then let $w^{(k+1)} \triangleq v^{(j)} + \hat{d}^{(j)}$ and go to Step 3. Otherwise, go to Step 2.3.

Step 2.3 Find the smallest nonnegative integer m such that

$$\Psi_{\varrho_k,\varepsilon_k}\left(v^{(j)}+\rho^m\hat{d}^{(j)}\right)\le(1-2\sigma\rho^m)\Psi_{\varrho_k,\varepsilon_k}\left(v^{(j)}\right).$$

Let $m_j\triangleq m$, $\tau_j\triangleq\rho^{m_j}$, and $v^{(j+1)}\triangleq v^{(j)}+\tau_j\hat{d}^{(j)}$.

Step 2.4 If $\|H_{\varrho_k,\varepsilon_k}(v^{(j+1)})\|\le\beta_k$, then let $w^{(k+1)}\triangleq v^{(j+1)}$ and go to Step 3. Otherwise, set $j\triangleq j+1$ and go back to Step 2.1.

Step 3 Update the parameters as follows:

$$\begin{aligned}\varrho_{k+1}:&=\min\left\{\kappa\left\|H_{NR}\left(w^{(k+1)}\right)\right\|^2,\varrho_0\bar{\eta}^{k+1},\bar{\varrho}\Big(\lambda^\star\left(x^{(k+1)}-y^{(k+1)}\right),\hat{\kappa}\right.\\&\qquad\left.\left\|H_{NR}\left(w^{(k+1)}\right)\right\|\Big)\right\},\\\varepsilon_{k+1}:&=\min\left\{\kappa\left\|H_{NR}\left(w^{(k+1)}\right)\right\|^2,\varepsilon_0\bar{\eta}^{k+1}\right\},\\\beta_{k+1}:&=\beta_0\eta^{k+1},\end{aligned}$$

where $\lambda^\star$ is given by (4.15), and $\bar{\varrho}(t,\delta)$ is determined so that $|\gamma'_\varrho(t)-\gamma_0^+(t)|<\delta$ for any $\varrho\in(0,\bar{\varrho}(t,\delta))$.

Step 4 Set $k\triangleq k+1$. Go back to Step 1.

Note that by (4.14)–(4.16) it is not hard to calculate $\lambda^\star$ for NCP, SOCCP, and SDCP cases.

THEOREM 5.4. *Let $F:\mathcal{J}\to\mathcal{J}$ be a continuously differentiable and monotone function, and $\{w^{(k)}\}$ be a sequence generated by the Algorithm. If the solution set of SCCP*(1.1) *is nonempty and bounded, then $\{w^{(k)}\}$ is bounded, and every accumulation point is a solution of SCCP*(1.1). *In addition, if ∇F is locally Lipschitzian and every accumulation point of $\{\nabla H_{\varrho_k,\varepsilon_k}(w^{(k)})\}$ is nonsingular, then the sequence $\{w^{(k)}\}$ converges to a solution w^* of SCCP*(1.1) *quadratically.*

Acknowledgments. The authors are very grateful to the anonymous referees for their constructive comments and valuable suggestions. The authors thank Defeng Sun and Levent Tunçel for helpful discussions.

REFERENCES

[1] B. CHEN AND P. T. HARKER, *A non-interior-point continuation method for linear complementarity problems*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1168–1190.

[2] C. CHEN AND O. L. MANGASARIAN, *Smoothing methods for convex inequalities and linear complementarity problems*, Math. Program., 71 (1995), pp. 51–69.

[3] J.-S. CHEN AND P. TSENG, *An unconstrained smooth minimization reformulation of the second-order cone complementarity problem*, Math. Program. Ser. B, 104 (2005), pp. 293–327.

[4] X. CHEN AND H. D. QI, *Cartesian P-property and its applications to the semidefinite linear complementarity problem*, Math. Program., 106 (2006), pp. 177–201.

[5] X. CHEN, H. D. QI, AND P. TSENG, *Analysis of nonsmooth symmetric matrix functions with applications to semidefinite complementarity problems*, SIAM J. Optim., 13 (2003), pp. 960–985.

[6] X. CHEN AND P. TSENG, *Non-interior continuation methods for solving semidefinite complementarity problems*, Math. Program., 95 (2003), pp. 431–474.

[7] X. D. CHEN, D. SUN, AND J. SUN, *Complementarity functions and numerical experiments on some smoothing Newton methods for second-order-cone complementarity problems*, Comput. Optim. Appl., 25 (2003), pp. 39–56.

[8] X. J. CHEN, L. QI, AND D. SUN, *Global and superlinear convergence of the smoothing Newton method and its application to general box constrained variational inequalities*, Math. Comput., 67 (1998), pp. 519–540.

[9] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems, Volume* I *and* II, Springer-Verlag, New York, 2003.
[10] J. Faraut and A. Korányi, *Analysis on Symmetric Cones*, Oxford University Press, New York, 1994.
[11] M. Fukushima, Z. Q. Luo, and P. Tseng, *Smoothing functions for second-order cone complementarity problems*, SIAM J. Optim., 12 (2001), pp. 436–460.
[12] M. S. Gowda, R. Sznajder, and J. Tao, *Some P-properties for linear transformations on Euclidean Jordan algebras*, Linear Algebra Appl., 393 (2004), pp. 203–232.
[13] M. S. Gowda and R. Sznajder, *Automorphism invariance of P and GUS properties of linear transformations on Euclidean Jordan algebras*, Math. Oper. Res., 31 (2006), pp. 109–123.
[14] M. S. Gowda and R. Sznajder, *Some global uniqueness and solvability results for linear complementarity problems over symmetric cones*, SIAM J. Optim., 18 (2007), pp. 461–481.
[15] S. Hayashi, N. Yamashita, and M. Fukushima, *A combined smoothing and regularization method for monotone second-order cone complementarity problems*, SIAM J. Optim., 15 (2005), pp. 593–615.
[16] Z. H. Huang and J. Han, *Non-interior continuation method for solving the monotone semidefinite complementarity problem*, Appl. Math. Optim., 47 (2003), pp. 195–211.
[17] G. Isac, *Topological Methods in Complementarity Theory*, Kluwer Academic Publishers, Dordrecht, 2000.
[18] C. Kanzow, *Some noninterior continuation methods for linear complementarity problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 851–868.
[19] M. Koecher, *The Minnesota Notes on Jordan Algebras and Their Applications*, edited and annotated by A. Brieg and S. Walcher, Springer, Berlin, 1999.
[20] L. C. Kong and N. H. Xiu, *On uniqueness of the Jordan frame in Euclidean Jordan algebras*, J. Beijing Jiaotong University, 3 (2007), pp. 54–57.
[21] K. Löwner, *Über monotone matrixfunctionen*, Math. Z., 38 (1934), pp. 177–216.
[22] Y. Liu, L. Zhang, and Y. Wang, *Some properties of a class of merit functions for symmetric cone complementarity problems*, Asia-Pacific J. Oper. Res., 23 (2006), pp. 473–496.
[23] M. Malik and S. R. Mohan, *Cone complementarity problems with finite solution sets*, Oper. Res. Lett., 34 (2006), pp. 121–126.
[24] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Cont. Optim., 15 (1977), pp. 959–972.
[25] J. M. Peng and M. Fukushima, *A hybrid Newton method for solving the variational inequality problem via the D-gap function*, Math. Program., 86 (1999), pp. 367–386.
[26] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Program., 58 (1993), pp. 353–367.
[27] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer, Berlin, 2004.
[28] S. Smale, *Algorithms for solving equations*, in Proceedings of the International Congress of Mathematicians, American Mathematical Society, Providence, RI, 1987, pp. 172–195.
[29] D. Sun and J. Sun, *Semismooth matrix valued functions*, Math. Oper. Res., 27 (2002), pp. 150–169.
[30] D. Sun and J. Sun, *Löwner's operator and spectral functions on Euclidean Jordan algebras*, Math. Oper. Res., 33 (2008), pp. 421–445.
[31] D. Sun and J. Sun, *Strong semismoothness of the Fischer-Burmeister SDC and SOC complementarity functions*, Math. Program., 103 (2005), pp. 575–582.
[32] J. Sun, D. Sun, and L. Qi, *A squared smoothing Newton method for nonsmooth matrix equations and its applications in semidefinite optimization problems*, SIAM J. Optim., 14 (2004), pp. 783–806.
[33] J. Tao and M. S. Gowda, *Some P-properties for nonlinear transformations on Euclidean Jordan algebras*, Math. Oper. Res., 30 (2005), pp. 985–1004.
[34] P. Tseng, *Analysis of a non-interior continuation method based on Chen–Mangasarian smoothing functions for complementarity problems*, in Reformulation-Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, M. Fukushima, L. Qi, eds., Kluwer Academic, Boston, 1999, pp. 381–404.
[35] Y. Xia and J. M. Peng, *A continuation method for the linear second-order cone complementarity Problem*, Computational Science and Its Applications-ICCSA 2005, Vol. 4, Proceedings Lecture Notes in Computer Science 3483, Springer, Berlin, 2005, pp. 290–300.
[36] N. Yamashita and M. Fukushima, *On the level-boundedness of the natural residual function for variational inequality problems*, Pac. J. Optim., 1 (2005), pp. 625–630.
[37] A. Yoshise, *Interior point trajectories and a homogeneous model for nonlinear complementarity problems over symmetric cones*, SIAM J. Optim., 17 (2006), pp. 1129–1153.

SIAM J. OPTIM.
Vol. 19, No. 3, pp. 1048–1072

VARIATIONAL ANALYSIS OF PSEUDOSPECTRA*

ADRIAN S. LEWIS† AND C. H. JEFFREY PANG†

Abstract. The ϵ-pseudospectrum of a square matrix A is the set of eigenvalues attainable when A is perturbed by matrices of spectral norm not greater than ϵ. The pseudospectral abscissa is the largest real part of such an eigenvalue, and the pseudospectral radius is the largest absolute value of such an eigenvalue. We find conditions for the pseudospectrum to be Lipschitz continuous in the set-valued sense and hence find conditions for the pseudospectral abscissa and the pseudospectral radius to be Lipschitz continuous in the single-valued sense. Our approach illustrates diverse techniques of variational analysis. The points at which the pseudospectrum is not Lipschitz (or more properly, does not have the Aubin property) are exactly the critical points of the resolvent norm, which in turn are related to the coalescence points of pseudospectral components.

Key words. pseudospectrum, variational analysis, Lipschitz multifunction, Aubin property, nonsmooth analysis, normal cone

AMS subject classifications. 15A18, 49K40, 65F15, 65K10, 90C31

DOI. 10.1137/070681521

1. Introduction.

Analysis using eigenvalues is prevalent in many different areas of applied mathematics. As we consider perturbations to an $n \times n$ complex matrix A with spectrum $\Lambda(A)$, we are led to study the ϵ-pseudospectrum $\Lambda_\epsilon : M^n \rightrightarrows \mathbb{C}$, which is a set-valued map defined by

$$\Lambda_\epsilon(A) = \{z \mid \exists E \in M^n \text{ such that } \|E\| \leq \epsilon, z \in \Lambda(A+E)\},$$

where M^n is the space of matrices of size $n \times n$. A well-known equivalent formulation, assuming $\|\cdot\| = \|\cdot\|_2$ as we do throughout, is

$$\Lambda_\epsilon(A) = \{z \mid \underline{\sigma}(A - zI) \leq \epsilon\},$$

where $\underline{\sigma}(A)$ denotes the smallest singular value of the matrix A. As discussed extensively in [22], the function $z \mapsto (zI - A)^{-1}$ is called the *resolvent* of the matrix A. Thus the pseudospectra of A are just upper-level sets of the *resolvent norm* function $n_A : \mathbb{C} \backslash \Lambda(A) \to \mathbb{R}_+$ defined by

$$n_A(z) := \left\|(zI - A)^{-1}\right\| = \frac{1}{\underline{\sigma}(A - zI)}.$$

Pseudospectra may be more informative than eigenvalues in applications where matrices are nonnormal [22, 13].

The continuity of the spectrum is well known [14]. One immediate question is whether continuity extends to Λ_ϵ. Since Λ_ϵ is a set-valued map, we ask whether we have continuity in the Hausdorff metric, and it is known that the answer is yes [17, Theorem 2.3.7].

*Received by the editors January 31, 2007; accepted for publication (in revised form) May 23, 2008; published electronically October 22, 2008.
http://www.siam.org/journals/siopt/19-3/68152.html
†School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY 14853 and Center for Applied Mathematics, Cornell University, Ithaca, NY 14853 (aslewis@orie.cornell.edu, cp229@cornell.edu). The research of the first author is supported in part by National Science Foundation grant DMS-0504032.

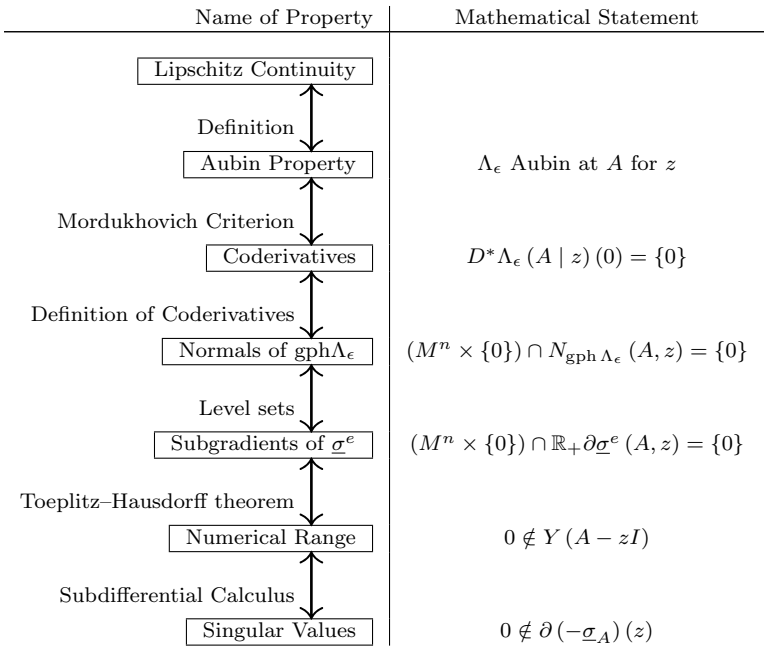


FIG. 1. *Equivalences of properties summarized in Theorem* 5.2.

Does the pseudospectrum mapping Λ_ϵ have stronger continuity properties? One of the main contributions of this paper is to find conditions under which the map Λ_ϵ is Lipschitz continuous. The ingredients of our analysis are variational-analytic techniques from the last couple of decades, as described in Rockafellar and Wets [21], Clarke et al. [10], and Mordukhovich [20]. In particular, we should note that there are technical details involved in the generalization of Lipschitz continuity to set-valued maps. Our proof (of the main results in Theorem 5.2 and Proposition 6.3) may be described loosely by Figure 1. The reader may find the schematic outline helpful as the argument proceeds.

For the moment, we remark on the notation

$$\underline{\sigma}_A(z) = \underline{\sigma}^e(A, z) = \frac{1}{n_A(z)}$$

and $Y(A - zI)$, which refers to the set of the inner products of associated left and right singular vectors (see page 1050). N refers to the normal cone, ∂ refers to the subdifferential and D^* refers to the coderivative. We expand more on the notation of Figure 1 (see page 1051).

In Figure 1, the six properties on the right on A and z are equivalent. For a given matrix A, we call points z not satisfying these equivalent properties "resolvent-critical" because they are smooth or nonsmooth critical points of the norm of the resolvent n_A. When the multiplicity of the smallest singular value of $A - zI$ is one, this property is equivalent to z being a "degenerate point" (in the sense of [4, Definition 4.5, Corrigendum]) or not "regular" in the sense of [5, Definition 4.4]. Points not resolvent-critical are exceptional for several aspects of pseudospectra, notably the quadratic convergence of the pseudospectral abscissa algorithm in [5].

As well as our main result equating the absence of the Aubin property with resolvent-criticality, we derive a variety of other properties of resolvent-critical points proving, in particular, that points where pseudospectral components coalesce as ϵ grows are resolvent-critical.

As an application of the Lipschitz continuity of $\Lambda_\epsilon : M^n \rightrightarrows \mathbb{C}$, we find conditions for the Lipschitz continuity (in the single-valued sense) and the strict differentiability of the pseudospectral abscissa $\alpha_\epsilon : M^n \to \mathbb{R}$, and the pseudospectral radius $\rho_\epsilon : M^n \to \mathbb{R}_+$ defined by

$$\begin{aligned} \alpha_\epsilon(A) &:= \max\{\mathrm{Re}\,(\lambda) \mid \lambda \in \Lambda_\epsilon(A)\}, \\ \rho_\epsilon(A) &:= \max\{|\lambda| \mid \lambda \in \Lambda_\epsilon(A)\}. \end{aligned}$$

We write $MSV : M^n \rightrightarrows \mathbb{C}^n \times \mathbb{C}^n$, with

$$\begin{aligned} MSV(A) := \{(u,v) \mid\ & u, v \text{ minimal left and} \\ & \text{right singular vectors of } A\}. \end{aligned}$$

In the above definition of MSV, u, v are *minimal left and right singular vectors* of A if they are unit vectors satisfying

$$\begin{aligned} \underline{\sigma}(A)\, u &= Av \\ \text{and } \underline{\sigma}(A)\, v &= A^H u, \end{aligned}$$

where A^H is the Hermitian transpose of A. A key tool in our analysis is the set

$$Y(A) := \left\{v^H u \mid (u,v) \in MSV(A)\right\}.$$

We prove that the set $Y(A - zI)$ is the subgradient set at z of the function $-\underline{\sigma}_A : \mathbb{C} \to \mathbb{R}_-$, where $\underline{\sigma}_A(z) = \underline{\sigma}(A - zI)$.

Related to Λ_ϵ is the mapping $\Lambda_\epsilon^c : M^n \rightrightarrows \mathbb{C}$ defined by $\Lambda_\epsilon^c(A) = \{z \mid \underline{\sigma}(A - zI) \geq \epsilon\}$. This mapping turns out to be easier to analyze because $-\underline{\sigma}(\cdot)$ has the property of subdifferential regularity (as defined in [21]) except at where it is zero. We show that the normal cone $N_{\Lambda_\epsilon^c(A)}(\bar{z})$ is $\mathbb{R}_+(Y(A - \bar{z}I))$. This establishes a link between the variational properties of $-\underline{\sigma}_A$ and Λ_ϵ^c, and the Aubin property.

Notation. For future reference, Tables 1 and 2 summarize the mappings that appear throughout the paper.

Unless otherwise stated, our notation follows [21]. See also the table of notation in [21, page 725]. The term "regular" means subdifferentially regular in the sense of [21, Definition 7.25]. Table 2 summarizes the symbols we use.

The "^{H}" in A^H and v^H represent the Hermitian transpose of a matrix or vector, while the "*" in L^* represents the adjoint of the linear operator L. Note that D^* stands for the coderivative instead. The real inner product on $A, B \in M^n$ is defined by $\mathrm{Re\ tr}\,\left(A^H B\right)$.

Outline. The paper is organized as follows. Section 2 studies the continuity properties of the pseudospectra Λ_ϵ and its "complement" Λ_ϵ^c via more general feasible-set mappings. In sections 3, 4, and 5, we prove the main result that Λ_ϵ has the Aubin property at A for z if and only if $0 \notin Y(A - zI)$, with section 3 containing general results on variational analysis and the singular value decomposition, section 4 performing subdifferential calculus, and section 5 finishing the proof of the main result.

TABLE 1
Summary of definitions.

Name/domain/range	Definition
$\bar{\sigma} : M^n \to \mathbb{R}_+$	$\bar{\sigma}(A)$ is maximum singular value of A
$\underline{\sigma} : M^n \to \mathbb{R}_+$	$\underline{\sigma}(A)$ is minimum singular value of A
$\underline{\sigma}^e : M^n \times \mathbb{C} \to \mathbb{R}_+$	$\underline{\sigma}^e(A, z) = \underline{\sigma}(A - zI)$
$\underline{\sigma}_A : \mathbb{C} \to \mathbb{R}_+$	$\underline{\sigma}_A(z) = \underline{\sigma}(A - zI)$
$\Lambda_\epsilon : M^n \rightrightarrows \mathbb{C}$	$\Lambda_\epsilon(A) = \{z \mid \underline{\sigma}(A - zI) \leq \epsilon\}$
$\Lambda : M^n \rightrightarrows \mathbb{C}$	$\Lambda(A) = \Lambda_0(A) = \{\text{eigenvalues of } A\}$
$\Lambda_\epsilon^c : M^n \rightrightarrows \mathbb{C}$	$\Lambda_\epsilon^c(A) = \{z \mid \underline{\sigma}(A - zI) \geq \epsilon\}$
$\alpha_\epsilon : M^n \to \mathbb{R}$	$\alpha_\epsilon(A) = \max_{z \in \Lambda_\epsilon(A)} \operatorname{Re} z$
$\rho_\epsilon : M^n \to \mathbb{R}_+$	$\rho_\epsilon(A) = \max_{z \in \Lambda_\epsilon(A)} \lvert z \rvert$
$W : M^n \rightrightarrows \mathbb{C}$	Numerical range/ field of values[15, Definition 1.1.1]
$MSV : M^n \rightrightarrows \mathbb{C}^n \times \mathbb{C}^n$	See Definition 3.2
$Y : M^n \rightrightarrows \mathbb{C}$	See Definition 3.2

TABLE 2
Summary of definitions.

Symbol	Explanation	Reference from [21]
$\hat{\partial}$	regular subgradient set	Definition 8.3
∂	subgradient set	Definition 8.3
∂^∞	horizon subgradient set	Definition 8.3
$\hat{N}$	regular normal cone	Definition 6.3
N	normal cone	Definition 6.4
osc	outer semicontinuous	Definition 5.4
isc	inner semicontinuous	Definition 5.4
pos	positive hull	section 3G
$\operatorname{lip} S(\cdot \mid \cdot)$	graphical modulus	Definition 9.36
$\operatorname{lip}_\infty S(\cdot)$	Lipschitz modulus	Definition 9.28
limsup	(set) outer limit	Formula 5(1)
liminf	(set) inner limit	Formula 5(1)
$D^* S(\cdot \mid \cdot)$	coderivative	Definition 8.33
$\lvert\cdot\rvert^+$	outer norm	Formula 9(4)
$\mathbf{d}(\cdot, \cdot)$	Pompieu-Hausdorff distance	Example 4.13
$\operatorname{lev}_{\leq\alpha} f$	Level sets: $\{x \mid f(x) \leq \alpha\}$	section 1B
conv	convex hull	section 1E
bdry	boundary of a set	
$\mathbb{B}$	unit ball $\{x \mid \lvert x\rvert \leq 1\}$	

In section 6, we show how the Lipschitz constant for the map Λ_ϵ can be calculated. Section 7 gives conditions for the Lipschitz continuity and strict differentiability of the pseudospectral abscissa α_ϵ and the pseudospectral radius ρ_ϵ. Finally, we present properties of resolvent-critical points in section 8. We prove, in particular, that the points at which the components of $\Lambda_\epsilon(A)$ coalesce as ϵ grows are resolvent-critical, and we pose some questions about resolvent-critical points.

2. Feasible-set mappings and continuity of pseudospectra. The pseudospectral mapping $\Lambda_\epsilon : M^n \rightrightarrows \mathbb{C}$ has two inputs: $\epsilon \in \mathbb{R}_+$ and the matrix in the argument of $\Lambda_\epsilon(\cdot)$. As $\mathbb{R}_+$ is one-dimensional, the variation of $\Lambda_\epsilon(A)$ for a fixed matrix A and variable ϵ is easier to visualize, as implemented in EigTool [24]. Some attractive results in this direction have been obtained in [7, 8, 18, 1, 17] and elsewhere. By contrast, in this work we study how Λ_ϵ behaves for a fixed ϵ and a varying matrix argument, primarily taking a more abstract and systematic approach than [6].

We study pseudospectra using the language of set-valued analysis as described in the monograph [21]. We take the definition of inner semicontinuity and outer semicontinuity in [21, section 5B].

In the next two propositions, let $f : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}^m$ be a continuous function, and let $T : \mathbb{R}^d \rightrightarrows \mathbb{R}^n$ be the mapping defined by

$$T(w) = \{x \mid f(x, w) \in D\}, \tag{2.1}$$

where D is a closed set.

PROPOSITION 2.1. *T is outer semicontinuous.*

Proof. We just need to check that T has a closed graph (by [21, Theorem 5.7]), which is easy. □

Note that the ϵ-pseudospectrum can be written as

$$\begin{aligned}\Lambda_\epsilon(A) &= \{z \mid \underline{\sigma}^e(A, z) \le \epsilon\} \\ &= \{z \mid \underline{\sigma}^e(A, z) \in (-\infty, \epsilon]\}.\end{aligned}$$

If we apply Proposition 2.1, we can deduce that Λ_ϵ is outer semicontinuous. In a similar manner, Λ^c_ϵ, defined by $\Lambda^c_\epsilon(A) = \{z \mid \underline{\sigma}^e(A, z) \ge \epsilon\}$, is also outer semicontinuous.

Turning to inner semicontinuity, we begin with a technical result.

PROPOSITION 2.2. *Let*

$$Q := \mathrm{cl}\{x \mid f(x, \bar{w}) \in \mathrm{int}(D)\},$$

so $Q \subset T(\bar{w})$. We have

(a) $Q \subset \liminf_{w \to \bar{w}} T(w) \subset T(\bar{w})$.

In the case where $m = 1$:

(b) *If $D = (-\infty, \alpha]$, then*

$$\begin{aligned}Q = \{&x \mid f(x, \bar{w}) = \alpha \text{ , } x \text{ is not a local minimizer of } f(\cdot, \bar{w})\} \\ &\cup \{x \mid f(x, \bar{w}) < \alpha\}.\end{aligned}$$

(c) *If $D = [\alpha, \infty)$, then*

$$\begin{aligned}Q = \{&x \mid f(x, \bar{w}) = \alpha \text{ , } x \text{ is not a local maximizer of } f(\cdot, \bar{w})\} \\ &\cup \{x \mid f(x, \bar{w}) > \alpha\}.\end{aligned}$$

(d) *If $\alpha > 0$, f is positively homogeneous (that is, $\lambda f(\cdot) = f(\lambda \cdot)$ for $\lambda > 0$) and either $D = (-\infty, \alpha]$ or $D = [\alpha, \infty)$, then $Q = \liminf_{w \to \bar{w}} T(w)$.*

Proof. Property (a) is easy and standard. See, for example, the techniques in [2, 16].

Statements (b) and (c) are clear by the definition of Q, so we proceed to prove statement (d) for the case $D = (-\infty, \alpha]$. (The case $D = [\alpha, \infty)$ is similar and is omitted.) From statement (a), we just need to prove that if $\bar{x} \notin Q$, then $\bar{x} \notin \liminf_{w \to \bar{w}} T(w)$. Suppose that $\bar{x} \notin Q$. We need to consider only $\bar{x} \in T(\bar{w})$, so we can assume that $\bar{x}$ is a minimizer of $f(\cdot, \bar{w})$ and $f(\bar{x}, \bar{w}) = \alpha$. Then there is a neighborhood $\mathbb{B}_\delta(\bar{x})$ about $\bar{x}$ such that $f(x, \bar{w}) \ge f(\bar{x}, \bar{w}) = \alpha$ if $x \in \mathbb{B}_\delta(\bar{x})$. If $\|x - \bar{x}\| < \delta/2$, then

$$\left\| \frac{1}{1 + \frac{1}{j}} x - \bar{x} \right\| < \delta \text{ if } j \text{ is large.}$$

This means that

$$\begin{aligned} f\left(x, \left(1+\frac{1}{j}\right)\bar{w}\right) &= \left(1+\frac{1}{j}\right) f\left(\frac{1}{1+\frac{1}{j}}x, \bar{w}\right) \\ &\geq \left(1+\frac{1}{j}\right)\alpha \quad \left(\text{because } \left\|\left(\frac{1}{1+\frac{1}{j}}\right)x - \bar{x}\right\| < \delta\right) \\ &> \alpha, \end{aligned}$$

which implies that $\mathbb{B}_{\delta/2}(\bar{x}) \cap T((1+\frac{1}{j})\bar{w}) = \emptyset$ if j is large enough. So for the sequence $(1+\frac{1}{j})\bar{w} \to \bar{w}$ as $j \to \infty$, we cannot find a subsequence x_j such that $x_j \in T((1+\frac{1}{j})\bar{w})$ and $x_j \to \bar{x}$, and this means that $\bar{x} \notin \liminf_{w\to\bar{w}} T(w)$. ∎

The following corollary is immediate from the definition of inner semicontinuity.

COROLLARY 2.3. *If $T(\bar{w}) = Q$, then T is continuous at $\bar{w}$. Furthermore, if f is positively homogeneous, then the converse holds as well.*

Proof. The mapping T is continuous if and only if it is both inner and outer semicontinuous. Apply the last two propositions. ∎

Now that we have established conditions for outer and inner semicontinuity for feasible-set mappings, we shall study the continuity of the pseudospectrum Λ_ϵ and Λ_ϵ^c. Let us consider the case $\epsilon = 0$ first. The map $\Lambda_0^c : M^n \rightrightarrows \mathbb{C}$ is not interesting as $\Lambda_0^c(A) = \mathbb{C}$ for all matrices A. We are then led to consider the spectrum $\Lambda_0 = \Lambda$, which is well known to be continuous [14, Appendix D].

To extend to $\epsilon > 0$, we may apply Propositions 2.1 and 2.2, combined with the fact that $\underline{\sigma}_A(\cdot)$ has no local minimum other than at the eigenvalues [22, Theorem 2.4(i)], to prove the following result. This result is not new and can be found, for example, in [17, Corollary 2.3.8].

PROPOSITION 2.4. *$\Lambda_\epsilon : M^n \rightrightarrows \mathbb{C}$ is continuous for $\epsilon \geq 0$.*

For $\Lambda_\epsilon^c : M^n \rightrightarrows \mathbb{C}$, we obtain the following using Proposition 2.2(d).

PROPOSITION 2.5. *$\Lambda_\epsilon^c : M^n \rightrightarrows \mathbb{C}$ is outer semicontinuous, but it is inner semicontinuous at a matrix A if and only if there is no local maximizer $\bar{z}$ to $\underline{\sigma}_A : \mathbb{C} \to \mathbb{R}_+$, with $\underline{\sigma}_A(\bar{z}) = \epsilon$.*

Example 2.6. The mapping Λ_ϵ^c is not continuous at some points. For a concrete example of the noncontinuity of Λ_ϵ^c, consider the point $0 \in \Lambda_1^c(\bar{A})$, where $\bar{A} = \operatorname{diag}(1, -1, i, -i)$ and $\epsilon = 1$. Here $\Lambda_1(\bar{A})$ consists of the union of balls of radius 1 around the diagonal entries, and so we observe that 0 is a local maximum of $\underline{\sigma}_{\bar{A}}$. This exhibits an example of the discontinuity of Λ_1^c as $\liminf_{A\to\bar{A}} \Lambda_1^c(A) \subsetneq \Lambda_1^c(\bar{A})$.

Next, we consider Lipschitz continuity. First, we define the Pompieu–Hausdorff distance.

DEFINITION 2.7 (see [21, Example 4.13]). *For $C, D \subset \mathbb{R}^n$ closed and nonempty, the* Pompieu–Hausdorff distance $\mathbf{d}(C, D)$ *is defined as*

$$\mathbf{d}(C, D) := \inf\{\eta \geq 0 \mid C \subset D + \eta\mathbb{B}, D \subset C + \eta\mathbb{B}\}.$$

Lipschitz continuity is thus defined as follows.

DEFINITION 2.8 (see [21, Definitions 9.26, 9.28]). *A mapping $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is* Lipschitz continuous *if it is nonempty–closed-valued and there exists $\kappa \in \mathbb{R}_+$, a Lipschitz constant, such that $\mathbf{d}(S(x), S(x')) \leq \kappa|x - x'|$ for all $x, x' \in \mathbb{R}^n$, or*

$$S(x') \subset S(x) + \kappa|x' - x|\,\mathbb{B} \text{ for all } x, x' \in \mathbb{R}^n.$$

The infimum of all κ such that there exists a neighborhood V of $\bar{x}$ such that

$$S(x') \subset S(x) + \kappa |x' - x| \mathbb{B} \text{ for all } x, x' \in V$$

is the Lipschitz modulus *for S at $\bar{x}$ and is denoted by* $\operatorname{lip}_\infty S(\bar{x})$.

The Aubin property, which is a localized Lipschitz property, is defined as follows.

DEFINITION 2.9 (see [21, Definition 9.36] Aubin property and graphical modulus). *A mapping $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ has the* Aubin property at $\bar{x}$ for $\bar{u}$, *where $\bar{x} \in \mathbb{R}^n$ and $\bar{u} \in S(\bar{x})$, if* $\operatorname{gph} S$ *is locally closed at $(\bar{x}, \bar{u})$ and there are neighborhoods V of $\bar{x}$ and W of $\bar{u}$, and a constant $\kappa \in \mathbb{R}_+$ such that*

$$S(x') \cap W \subset S(x) + \kappa |x' - x| \mathbb{B} \text{ for all } x, x' \in V.$$

The graphical modulus *of S at $\bar{x}$ for $\bar{u}$, denoted by* $\operatorname{lip} S(\bar{x} \mid \bar{u})$, *is the infimum of all such κ that satisfy the formula above.*

If the function f in the feasible-set mapping in formula (2.1) in page 1052 is smooth, we understand the Aubin Property quite well through [21, Example 9.51]. If $D = (-\infty, \bar{\alpha}]$, we can also analyze the nonsmooth case. In what follows, ∂ and ∂^∞ denote, respectively, the subgradient set and the horizon subgradient set [21, Definition 8.3].

Assumptions (a), (b), and (c) in the result below are standard for computing normals to level sets (see, for example, [21, Proposition 10.3].) Assumption (d) is needed to apply a chain rule.

THEOREM 2.10. *Consider the set-valued map $C : \mathbb{R}^d \rightrightarrows \mathbb{R}^n$ defined via a level set representation*

$$C(p) = \{x \mid F(x, p) \leq \bar{\alpha}\},$$

with $F : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$. Suppose that

(a) $F(\bar{x}, \bar{p}) = \bar{\alpha}$,

(b) $(0, 0) \notin \partial F(\bar{x}, \bar{p})$,

(c) *F is regular at $(\bar{x}, \bar{p})$,*

(d) $(0, y_2) \in \partial^\infty F(\bar{x}, \bar{p}) \implies y_2 = 0$.

Then C has the Aubin property at $\bar{p}$ for $\bar{x}$ if and only if $0 \notin \partial F_{\bar{p}}(\bar{x})$, where $F_{\bar{p}} : \mathbb{R}^n \to \mathbb{R}$ is defined by $F_{\bar{p}}(x) := F(x, \bar{p})$. In this case,

$$\operatorname{lip} C(\bar{p} \mid \bar{x}) = \max_{\substack{(a,b) \in N_{\operatorname{gph} C}(\bar{p}, \bar{x}) \\ \|b\| = 1}} \|a\|,$$

If $F(\bar{x}, \bar{p}) < \bar{\alpha}$, then C has the Aubin property at $\bar{p}$ for $\bar{x}$, with $\operatorname{lip} C(\bar{p} \mid \bar{x}) = 0$.

Proof. The Mordukhovich criterion [21, Theorem 9.40] tells us that C has the Aubin property at $\bar{p}$ for $\bar{x}$ if and only if $D^* C(\bar{p} \mid \bar{x})(0) = \{0\}$, where D^* denotes the coderivative [21, Definition 8.33]. This holds if and only if

$$(z, 0) \in N_{\operatorname{gph} C}(\bar{p}, \bar{x}) \text{ implies } z = 0. \tag{2.2}$$

This property is equivalent to

$$(0, z) \in N_{\operatorname{gph} C^{-1}}(\bar{x}, \bar{p}) \text{ implies } z = 0.$$

Conditions (a), (b), and (c) allow us to conclude that

$$N_{\operatorname{gph} C^{-1}}(\bar{x}, \bar{p}) = (\operatorname{pos} \partial F(\bar{x}, \bar{p})) \cup \partial^\infty F(\bar{x}, \bar{p}) \tag{2.3}$$

through a result on level sets [21, Proposition 10.3], or

$$(0,z) \in (\text{pos } \partial F(\bar{x},\bar{p})) \cup \partial^{\infty} F(\bar{x},\bar{p}) \text{ implies } z = 0,$$

and by condition (d), this is in turn equivalent to

$$(0,z) \in \text{pos } \partial F(\bar{x},\bar{p}) \text{ implies } z = 0. \tag{2.4}$$

We define $L_{\bar{p}} : \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^d$ by $L_{\bar{p}}(x) = (x,\bar{p})$. The adjoint $L^*_{\bar{p}} : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}^n$ is given by $L^*_{\bar{p}}(x,p) = x$. We have $F_{\bar{p}} = F \circ L_{\bar{p}}$, and so by a chain rule [21, Theorem 10.6] and condition (d), $\partial F_{\bar{p}}(\bar{x}) = L^*_{\bar{p}} \partial F(\bar{x},\bar{p})$. Thus

$$\begin{aligned} \partial F_{\bar{p}}(\bar{x}) &= L^*_{\bar{p}} \partial F(\bar{x},\bar{p}) \\ &= \{y \mid \exists z \text{ such that } (y,z) \in \partial F(\bar{x},\bar{p})\}. \end{aligned}$$

If $0 \in \partial F_{\bar{p}}(\bar{x})$, then there exists z such that $(0,z) \in \partial F(\bar{x},\bar{p})$, but condition (b) implies $z \neq 0$, which contradicts statement (2.4). If $0 \notin \partial F_{\bar{p}}(\bar{x})$, this means that there is no z such that $(0,z) \in \partial F(\bar{x},\bar{p})$ and implies statement (2.4). So $0 \notin \partial F_{\bar{p}}(\bar{x})$ is equivalent to C not having the Aubin property at $\bar{p}$ for $\bar{x}$ as claimed.

The calculation of the value lip $C(\bar{p} \mid \bar{x})$ follows from the definition of the coderivative $D^*C(\bar{p} \mid \bar{x})$ and its relation to the normal cone through the Mordukhovich criterion. If $F(\bar{x},\bar{p}) < \bar{\alpha}$, then the normal cone is $\{(0,0)\}$, giving us the required value of lip $C(\bar{p} \mid \bar{x})$. ∎

To obtain the Lipschitz modulus from the graphical modulus, one may use [21, Theorem 9.38], but Proposition 6.2 is sufficient for our purposes in this paper.

In sections 3 to 6, we will be using the general principle illustrated in Theorem 2.10 to study where the pseudospectrum Λ_ϵ has the Aubin property and also to illustrate how this can identify where Λ_ϵ is Lipschitz continuous and give a value of the Lipschitz constant.

One may immediately try to apply Theorem 2.10 to show that Λ_ϵ has the Aubin property for A at z. In this case, $p = A$, $x = z$, and so $C(p) = \Lambda_\epsilon(A)$, $F(x,p) = \underline{\sigma}(A - zI) = \underline{\sigma}^e(A,z)$. However, $\underline{\sigma}^e$ is not a regular function, but this can be overcome by studying $-\underline{\sigma}^e$ instead, which is regular if $A - zI$ is nonsingular. This is what we will do in the analysis that follows.

3. General results. First, we are interested in finding out whether the functions $-\underline{\sigma}^e$ and $\frac{1}{\underline{\sigma}^e}$ enjoy similar regularity properties so that we can deduce properties of $\underline{\sigma}^e$. We recall a result on the reciprocals of functions.

PROPOSITION 3.1 (see [20, Corollary 1.111(iii)]). *For any function $h : \mathbb{R}^n \to \mathbb{R}$ at z where $h(z) > 0$, we have $\partial h(z) = h(z)^2 \, \partial\left(-\frac{1}{h}\right)(z)$, and h is regular at z if and only if $-\frac{1}{h}$ is regular there.*

The set of minimal singular vectors of A, $MSV(A)$, is defined below.

DEFINITION 3.2. *For a matrix A, the left and right singular vectors corresponding to the smallest singular value of A are the pairs $(u,v) \in \mathbb{C}^n \times \mathbb{C}^n$, $\|u\| = \|v\| = 1$, which appear in the appropriate columns of U and V in some singular value decomposition $A = USV^H$ of A. We refer to u and v as* minimal singular vectors*, and we denote the set of pairs of minimal singular vectors of A as $MSV(A)$. Furthermore, we define $Y : M^n \rightrightarrows \mathbb{C}$ by*

$$Y(A) := \left\{v^H u \mid (u,v) \in MSV(A)\right\}.$$

An equivalent definition given in the introduction is to have pairs of unit vectors (u,v) satisfying the equations $\underline{\sigma}(A)\,u = Av$ and $\underline{\sigma}(A)\,v = A^H u$.

The following result summarizes a complete characterization of left and right minimal singular vectors when we have one particular singular value decomposition, which is helpful for the case where the smallest singular value is multiple.

PROPOSITION 3.3. *Consider a matrix $A \in M^n$ with singular value decomposition (for unit vectors u_j, v_j)*

$$A = \sum_{j=1}^{n} \sigma_j u_j v_j^H = USV^H,$$

where $\sigma_1 = \sigma_2 = \cdots = \sigma_m < \sigma_j$ for all $j > m$. Define matrices $\bar{U} = (u_1\, u_2 \cdots u_m)$ and $\bar{V} = (v_1\, v_2 \cdots v_m)$. Then

$$MSV(A) = \left\{ \left(\bar{U}q, \bar{V}q\right) \mid q \in \mathbb{C}^m, \|q\| = 1 \right\}$$

if A is invertible (in other words, $\sigma_1 > 0$) and

$$MSV(A) = \left\{ \left(\bar{U}q_1, \bar{V}q_2\right) \mid q_1, q_2 \in \mathbb{C}^m, \|q_1\| = \|q_2\| = 1 \right\}$$

if A is singular.

Proof. The equations $Av = \underline{\sigma}(A)\, u$ and $A^H u = \underline{\sigma}(A)\, v$ require u to be an eigenvector for AA^H and v to be an eigenvector for $A^H A$, and so they lie in the subspaces spanned by the columns of $\bar{U}$ and $\bar{V}$, respectively. We have assumed that these columns are placed at the left of U and V. Then let $v = \bar{V}q$. As we want a v of unit length, we must have $\|q\| = 1$. Since A is invertible, $\underline{\sigma} := \underline{\sigma}(A) > 0$, and so

$$u = \frac{1}{\underline{\sigma}} Av = \frac{1}{\underline{\sigma}} USV^H \bar{V}q = \frac{1}{\underline{\sigma}} US \begin{pmatrix} I \\ 0 \end{pmatrix} q = U \begin{pmatrix} I \\ 0 \end{pmatrix} q = U \begin{pmatrix} q \\ 0 \end{pmatrix} = \bar{U}q.$$

Thus $MSV(A) \subset \left\{ \left(\bar{U}q, \bar{V}q\right) \mid q \in \mathbb{C}^m, \|q\| = 1 \right\}$. The reverse is straightforward.

If A is singular, then as before, $u = \bar{U}q_1$ and $v = \bar{V}q_2$ for some unit vectors q_1, q_2. It is evident that u and v satisfy the relations $\underline{\sigma}(A)\, u = Av$ and $\underline{\sigma}(A)\, v = A^H u$, so we are done. ∎

The significance of $Y(A)$ will become clear later in sections 4 and 5. We first show a result on $Y(A)$.

PROPOSITION 3.4. *If A is invertible, then $Y(A)$ is convex.*

Proof. We make the observation that the set $Y(A)$ can be determined as follows. Let $\bar{U}$ and $\bar{V}$ be as described in Proposition 3.3. The numerical range of a matrix $B \in M^n$ is the set $\left\{ v^H B v \mid v \in \mathbb{C}^n, \|v\| = 1 \right\}$, denoted by $W(B)$, and is convex by the Toeplitz–Hausdorff theorem [15, Property 1.2.2]. Then

$$\begin{aligned} Y(A) &= \left\{ v^H u \mid (u, v) \in MSV(A) \right\} \\ &= \left\{ q^H \bar{V}^H \bar{U} q \mid \|q\| = 1 \right\} \text{ (by Proposition 3.3)} \\ &= W\left(\bar{V}^H \bar{U}\right), \text{ the numerical range of } \bar{V}^H \bar{U}, \end{aligned}$$

establishing the convexity of $Y(A)$. ∎

For singular matrices A, $Y(A)$ need not be convex. Take, for example, the singular value decomposition

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

With this matrix,

$$Y(A) = \left\{ q_1 \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} q_2 \mid q_1, q_2 \in \mathbb{C}, |q_1| = |q_2| = 1 \right\}$$
$$= \{ q \in \mathbb{C} \mid |q| = 1 \},$$

which is not convex.

4. Subdifferential calculus. This section collects some results about subdifferential calculus involving $\underline{\sigma}^e : M^n \times \mathbb{C} \to \mathbb{R}_+$, where $\underline{\sigma}^e(A, z) = \underline{\sigma}(A - zI)$. As suggested in Figure 1, there is a link between the subdifferential $\partial \underline{\sigma}^e(A, z)$ and normal cone $N_{\operatorname{gph} \Lambda_\epsilon}(A, z)$ for $\underline{\sigma}^e(A, z) = \epsilon$. Before we can apply the appropriate theorems in [21], we have to calculate $\partial \underline{\sigma}^e(A, z)$, establish regularity properties, and characterize whether $0 \in \partial \underline{\sigma}^e(A, z)$.

When the smallest singular value is simple, $\underline{\sigma}$ and $\underline{\sigma}^e$ are analytic, as the next lemmas assert.

We remind the reader that the spaces M^n and $M^n \times \mathbb{C}$ have (real) inner products defined by

$$\langle A, B \rangle = \operatorname{Re} \operatorname{tr} \left(A^H B \right) \text{ for } A, B \in M^n$$

and

$$\langle (X, x), (Y, y) \rangle = \operatorname{Re} \left(\operatorname{tr} \left(X^H Y \right) + x^H y \right) \text{ for } X, Y \in M^n \text{ and } x, y \in \mathbb{C}.$$

LEMMA 4.1. *If the invertible matrix A has a simple smallest singular value, then the function $\underline{\sigma} : M^n \to \mathbb{R}_+$ is real-analytic at A, with gradient*

$$\nabla \underline{\sigma}(A) = uv^H$$

for any $(u, v) \in MSV(A)$.

The proof for the above lemma is standard (for example, [4, Theorem 7.1]), while the lemma below follows by noticing that $\underline{\sigma}^e = \underline{\sigma} \circ L$ and applying the chain rule, where $L : M^n \times \mathbb{C} \to \mathbb{C}$ is defined by $L(A, z) = A - zI$.

LEMMA 4.2. *If $z \notin \Lambda(A)$ and $A - zI$ has a simple smallest singular value, then the function $\underline{\sigma}^e : M^n \times \mathbb{C} \to \mathbb{R}_+$ is real-analytic at (A, z), with gradient*

$$\nabla \underline{\sigma}^e(A, z) = \left(uv^H, -v^H u \right)$$

for any $(u, v) \in MSV(A - zI)$.

The next two results are generalizations of Lemmas 4.1 and 4.2 to the nonsmooth case, and they calculate the subgradients needed in the main result in section 5.

PROPOSITION 4.3. *Suppose $z \notin \Lambda(A)$. Then*

$$\partial \left(-\underline{\sigma}^e \right)(A, z) = conv \left\{ \left(-uv^H, v^H u \right) \mid (u, v) \in MSV(A - zI) \right\}.$$

Furthermore, $-\underline{\sigma}^e$ is regular at (A, z) and globally Lipschitz.

Proof. We consider the functions

$$\bar{\sigma}^e : M^n \times \mathbb{C} \to \mathbb{R}_+, \ \iota : M^n \to M^n \text{ and } L : M^n \times \mathbb{C} \to M^n$$

defined by

$$\bar{\sigma}^e(A, z) = \bar{\sigma} \left((A - zI)^{-1} \right), \ \iota(B) = B^{-1} \text{ and } L(A, z) = A - zI.$$

That is, $\bar{\sigma}^e = \bar{\sigma} \circ \iota \circ L$. To evaluate the subdifferential of this function, we apply a chain rule [21, Theorem 10.6]. Given a matrix B, we seek to evaluate $\nabla (\iota \circ L)(A,z)^*(B)$, which is, by the chain rule, $\nabla L(A,z)^*(\nabla \iota(A - zI)^*(B))$.

As $\bar{\sigma}$ is everywhere Lipschitz, $\partial^\infty \bar{\sigma}(\iota \circ L(A,z)) = \{0\}$. Furthermore, since $\bar{\sigma}$ is convex, it is regular at $\iota \circ L(A,z)$, and so the conditions for [21, Theorem 10.6] are satisfied.

It is easy to check the identity $L^*(B) = (B, -\mathrm{tr}B)$. (Note that L is linear so $\nabla L = L$ and $\nabla L^* = L^*$.) Using the binomial expansion

$$(M + \Delta)^{-1} = M^{-1} - M^{-1}\Delta M^{-1} + o(\Delta),$$

it follows that $\nabla \iota(M)(B) = -M^{-1}BM^{-1}$, so $\nabla \iota(M)^*(B) = -M^{-H}BM^{-H}$ follows easily.

Next, we evaluate $\partial \bar{\sigma}^e(A,z)$. Let the singular value decomposition of $(A - zI)$ be USV^H. Then the singular value decomposition of $(A - zI)^{-1}$ is $VS^{-1}U^H$, and $(A - zI)^{-H} = US^{-1}V^H$. So

$$\partial \bar{\sigma}^e(A,z) = \nabla L(A,z)^* \nabla \iota(A - zI)^* \partial \bar{\sigma}\left((A - zI)^{-1}\right).$$

We know that

$$\partial \bar{\sigma}(B) = \mathrm{conv}\left\{uv^H \mid \|u\| = \|v\| = 1, Bv = \bar{\sigma}(B)u, B^H u = \bar{\sigma}(B)v\right\}.$$

(See, for example, [23].) Therefore,

$$\partial \bar{\sigma}\left((A - zI)^{-1}\right) = \mathrm{conv}\left\{vu^H \mid (u,v) \in MSV(A - zI)\right\}.$$

Then for any $(u,v) \in MSV(A - zI)$, we have

$$\begin{aligned}
\nabla L(A,z)^* \nabla \iota(A - zI)^* \left(vu^H\right) &= \nabla L(A,z)^* \left(-US^{-1}V^H vu^H US^{-1}V^H\right) \\
&= \underline{\sigma}(A - zI)^{-2} \nabla L(A,z)^* \left(-uv^H\right) \\
&= \underline{\sigma}(A - zI)^{-2} \left(-uv^H, \mathrm{tr}\ \left(uv^H\right)\right) \\
&= \underline{\sigma}(A - zI)^{-2} \left(-uv^H, v^H u\right),
\end{aligned}$$

and so

$$\partial \bar{\sigma}^e(A,z) = \underline{\sigma}(A - zI)^{-2} \mathrm{conv}\left\{\left(-uv^H, v^H u\right) \mid (u,v) \in MSV(A - zI)\right\}.$$

By Proposition 3.1, we conclude that

$$\begin{aligned}
\partial\left(-\underline{\sigma}^e\right)(A,z) &= \partial\left(-\frac{1}{\bar{\sigma}^e}\right)(A,z) \\
&= \bar{\sigma}^e(A,z)^{-2} \partial \bar{\sigma}^e(A,z) \\
&= \mathrm{conv}\left\{\left(-uv^H, v^H u\right) \mid (u,v) \in MSV(A - zI)\right\}.
\end{aligned}$$

The function $-\underline{\sigma}^e$ is regular at (A,z) because $\bar{\sigma}$ is regular and both the chain rule [21, Theorem 10.6] and Proposition 3.1 guarantee the preservation of regularity. Also, the function $-\underline{\sigma}^e$ is globally Lipschitz because $-\underline{\sigma}^e = -\underline{\sigma} \circ L$ is the composition of two globally Lipschitz functions. □

From the definition of $\Lambda_\epsilon(A) = \{z \mid \underline{\sigma}_A(z) \leq \epsilon\}$, where $\underline{\sigma}_A : \mathbb{C} \to \mathbb{R}_+$ is defined by $\underline{\sigma}_A(z) = \underline{\sigma}(A - zI)$, it is clear that the functions $\underline{\sigma}$ and $\underline{\sigma}_A$ figure prominently in the study of pseudospectra. The following two results can be seen as nonsmooth analogues of [4, Theorem 7.1 and Corollary 7.2]. Even though $\underline{\sigma}$ and $\underline{\sigma}_A$ are not necessarily smooth, we are able to prove that $-\underline{\sigma}$ and $-\underline{\sigma}_A$ are regular and calculate their subgradients.

PROPOSITION 4.4. *The function $-\underline{\sigma}$ is regular at every nonsingular matrix $A \in M^n$, with*

$$\partial(-\underline{\sigma})(A) = -\operatorname{conv}\left\{uv^H \mid (u,v) \in MSV(A)\right\}.$$

Proof. Define $L_{M^n} : M^n \to M^n \times \mathbb{C}$ by $L_{M^n}(A) = (A, 0)$, so we have $-\underline{\sigma}_A = (-\underline{\sigma}^e) \circ L_{M^n}$. Clearly L_{M^n} is smooth, with $\nabla L_{M^n} = I \times \mathbf{0}$ at all points. $(\nabla L_{M^n})^* : M^n \times \mathbb{C} \to M^n$ is just the natural projection. Thus, by appealing to [21, Theorem 10.6] and Proposition 4.3, we get what we need. ∎

PROPOSITION 4.5. *For a matrix A, consider the function $\underline{\sigma}_A : \mathbb{C} \to \mathbb{R}_+$ defined by $\underline{\sigma}_A(z) = \underline{\sigma}(A - zI)$. If $z \notin \Lambda(A)$, then*

$$\partial(-\underline{\sigma}_A)(z) = Y(A - zI),$$

and $-\underline{\sigma}_A$ is regular at z and globally Lipschitz.

Proof. The proof is similar to the proof above, but we work through the details for completeness. We note $-\underline{\sigma}_A = (-\underline{\sigma}^e) \circ L_A$, where $L_A : \mathbb{C} \to M^n \times \mathbb{C}$, $L_A(z) = (A, z)$. Clearly L_A is smooth, with $\nabla L_A = \mathbf{0} \times I$ at all points. Furthermore, $(\nabla L_A)^* : M^n \times \mathbb{C} \to \mathbb{C}$ is just the natural projection. Thus, by appealing to a chain rule [21, Theorem 10.6] and Proposition 4.3, we have

$$\begin{aligned}\partial(-\underline{\sigma}_A)(z) &= (\nabla L_A)^* \partial(-\underline{\sigma}^e)(A, z)\\ &= Y(A - zI).\end{aligned}$$

As in Proposition 4.3, $\underline{\sigma}_A$ is globally Lipschitz because it is a composition of two globally Lipschitz functions. ∎

We note that the assumptions that $A - zI$ is nonsingular in Proposition 4.3 and A is nonsingular in Proposition 4.4 cannot be dropped in the proposition below.

PROPOSITION 4.6. *If $z \in \Lambda(A)$, then $-\underline{\sigma}^e$ is not regular at (A, z). Similarly, $-\underline{\sigma}$ is not regular at A if A is singular.*

Proof. Take $\bar{U}$ and $\bar{V}$ to the matrices corresponding to the minimal left and right singular vectors of $A - zI$ in the statement of Proposition 3.3. For small $\epsilon > 0$, we have

$$\begin{aligned}-\underline{\sigma}^e\left(A + \epsilon\bar{U}\bar{V}^H, z\right) &= -\underline{\sigma}^e(A, z) - \epsilon\\ \text{and } -\underline{\sigma}^e\left(A - \epsilon\bar{U}\bar{V}^H, z\right) &= -\underline{\sigma}^e(A, z) - \epsilon.\end{aligned}$$

Hence if $(B, x) \in \hat{\partial}(-\underline{\sigma}^e)(A, z)$, we have

$$\begin{aligned}-\underline{\sigma}^e\left(A \pm \epsilon\bar{U}\bar{V}^H, z\right) &\geq -\underline{\sigma}^e(A, z) + \left\langle (B, x), \left(\pm\epsilon\bar{U}\bar{V}^H, 0\right)\right\rangle + o(\epsilon)\\ \implies -\epsilon &\geq \epsilon\left\langle (B, x), \left(\pm\bar{U}\bar{V}^H, 0\right)\right\rangle + o(\epsilon).\end{aligned}$$

Dividing by ϵ throughout and taking limits as $\epsilon \downarrow 0$, we have

$$\begin{aligned}&-1 \geq \left\langle (B, x), \left(\pm\bar{U}\bar{V}^H, 0\right)\right\rangle\\ \implies &-2 \geq \left\langle (B, x), \left(\bar{U}\bar{V}^H, 0\right)\right\rangle + \left\langle (B, x), \left(-\bar{U}\bar{V}^H, 0\right)\right\rangle = 0,\end{aligned}$$

which is obviously a contradiction. This means that $\hat{\partial}\left(-\underline{\sigma}^e\right)(A,z) = \emptyset$. To show that $\partial\left(-\underline{\sigma}^e\right)(A,z) \neq \emptyset$, we note that for small $\epsilon > 0$, we have

$$\left(-u_1 v_1^H, v_1^H u_1\right) \in \hat{\partial}\left(-\underline{\sigma}^e\right)\left(A + \epsilon \bar{U}\bar{V}^H, z\right)$$

by Proposition 4.3, where the minimal left and right singular vectors u_1, v_1 are defined in the statement of Proposition 3.3. Taking $\epsilon \downarrow 0$, this ensures that $(-u_1 v_1^H, v_1^H u_1) \in \partial(-\underline{\sigma}^e)(A,z)$, and thus $\partial\left(-\underline{\sigma}^e\right)(A,z) \neq \emptyset$. Since $\partial\left(-\underline{\sigma}^e\right)$ and $\hat{\partial}\left(-\underline{\sigma}^e\right)$ differ and appealing to [21, Corollary 8.11], $-\underline{\sigma}^e$ is not regular at (A,z). The proof for $-\underline{\sigma}$ is similar. ∎

PROPOSITION 4.7. *The resolvent norm $n_A : \mathbb{C} \to \mathbb{R}$ defined by $n_A(z) = \|(zI - A)^{-1}\|$ is regular at every point where $z \notin \Lambda(A)$, with*

$$\partial n_A(z) = n_A(z)^2\, Y(A - zI).$$

Proof. From the identity $n_A = 1/\underline{\sigma}_A$ and Propositions 3.1 and 4.5, we note the following calculations:

$$\begin{aligned}\partial n_A(z) =& n_A(z)^2\, \partial\left(-\frac{1}{n_A}\right)(z)\\ =& n_A(z)^2\, \partial\left(-\underline{\sigma}_A\right)(z)\\ =& n_A(z)^2\, Y(A - zI).\end{aligned}$$ ∎

This motivates the following definition.

DEFINITION 4.8. *A point $z \in \mathbb{C}$ is* resolvent-critical *for a square matrix A if either $z \in \Lambda(A)$ or $0 \in Y(A - zI)$.*

Thus resolvent-critical points that are not eigenvalues are simply critical points of the resolvent norm n_A (in the nonsmooth sense). Recall that, for a locally Lipschitz function f, $\partial^\circ f(x)$, the convex hull of $\partial f(x)$, is the *Clarke subdifferential* of f at x and that $\bar{x}$ is *Clarke-critical* if $0 \in \partial^\circ f(\bar{x})$. Since $\underline{\sigma}_A$ is globally Lipschitz, the following holds as well.

THEOREM 4.9. *For a given matrix A, the following are equivalent:*

(1) *z is resolvent-critical.*

(2) *z is Clarke-critical for $-\underline{\sigma}_A$.*

(3) *z is Clarke-critical for $\underline{\sigma}_A$.*

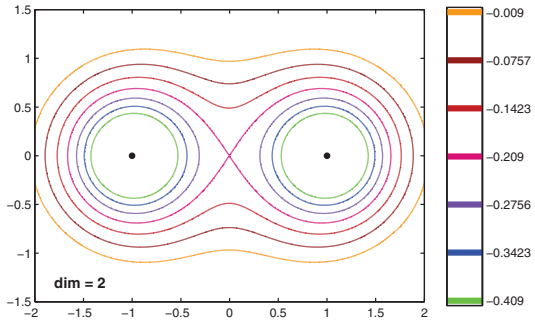
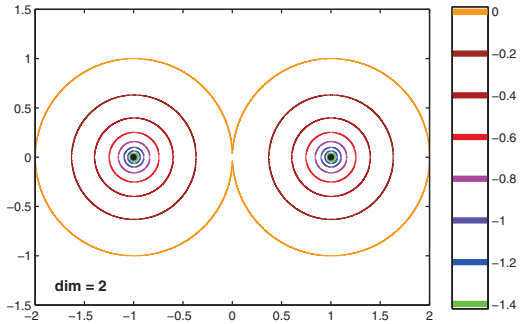
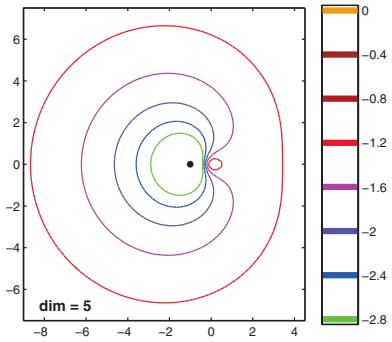
Proof. Since $\underline{\sigma}_A$ is Lipschitz, we have $\partial^\circ\left(-\underline{\sigma}_A\right)(z) = -\partial^\circ \underline{\sigma}_A(z)$ by [9, Proposition 2.3.1]. This means that (2) and (3) are equivalent.

Next we prove that (1) implies (2). If z is resolvent-critical, then either z is an eigenvalue of A or $0 \in \partial\left(-\underline{\sigma}_A\right)(z)$. In the second case, z is Clarke-critical for $-\underline{\sigma}_A$ because $\partial\left(-\underline{\sigma}_A\right)(z) \subset \partial^\circ\left(-\underline{\sigma}_A\right)(z)$. In the first case, z is a maximizer of $-\underline{\sigma}_A$, and so z is Clarke-critical.

Lastly, we prove that (2) implies (1). If z is not resolvent-critical, then z is not an eigenvalue, and $0 \notin \partial\left(-\underline{\sigma}_A\right)(z)$. But $\partial\left(-\underline{\sigma}_A\right)(z) = \partial^\circ\left(-\underline{\sigma}_A\right)(z)$ by the regularity of $-\underline{\sigma}_A$ at z, so we are done. ∎

Example 4.10. Table 3 shows some examples where 0 is a resolvent-critical point of A. (In the third example, the resolvent-critical point is close to 0 but not exactly at 0.) These plots were obtained with EigTool [24]. The curves represent the boundaries of the pseudospectra $\Lambda_\epsilon(A)$ for $\epsilon = 10^\alpha$, where α is the number corresponding to the line generated by EigTool in the legend on the right. The third example is found in [12].

TABLE 3
Examples of pseudospectra for Example 4.10.

A	Diagram
$\begin{pmatrix} 1 & 1 \\ 0 & -1 \end{pmatrix}$	Smooth Saddle 
$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$	Nonsmooth Saddle 
$-\begin{pmatrix} 1 & 5 & 5^2 & 5^3 & 5^4 \\ 0 & 1 & 5 & 5^2 & 5^3 \\ 0 & 0 & 1 & 5 & 5^2 \\ 0 & 0 & 0 & 1 & 5 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$	Local minimum of n_A 

We also have an alternative proof to [4, Theorem 9.2] after the remark below.

Remark 4.11. The set

$$G(z) = \left\{ v^H (A - zI) v \mid v \in V(z), \|v\| = 1 \right\},$$

where the subspace $V(z) \subset \mathbb{C}^n$ is spanned by all right singular vectors of $A - zI$ as defined in [4, Section 9], is equal to $\underline{\sigma}(A - zI) Y(A - zI)$.

PROPOSITION 4.12. *If $\bar{z}$ is not resolvent-critical and $\underline{\sigma}_A(\bar{z}) = \epsilon$, then the set $\Lambda_\epsilon^c(A)$ is Clarke regular at $\bar{z}$, with normal cone $N_{\Lambda_\epsilon^c(A)}(\bar{z}) = \operatorname{pos}(Y(A - \bar{z}I))$.*

Proof. This involves applying Proposition 4.5 to a result on level sets [21, Proposition 10.3]. ∎

The conditions below on $\partial\underline{\sigma}^e(A,z)$ and $\partial(-\underline{\sigma}^e)(A,z)$ are needed in a manner similar to condition (b) in Theorem 2.10 in the proof of our main result.

PROPOSITION 4.13. *The condition* $(0,0) \in \partial\underline{\sigma}^e(A,z)$ *holds if and only if* $z \in \Lambda(A)$. *Also, if* $z \notin \Lambda(A)$, *then* $(0,0) \notin \partial(-\underline{\sigma}^e)(A,z)$.

Proof. If $\underline{\sigma}^e(A,z) = 0$, then (A,z) is a local minimizer, and thus $(0,0) \in \partial\underline{\sigma}^e(A,z)$. On the other hand, if $\underline{\sigma}^e(A,z) > 0$, we need to prove that $(0,0) \notin \partial\underline{\sigma}^e(A,z)$. We try to evaluate $\hat{\partial}\underline{\sigma}^e(A,z)$. From Proposition 4.3, we know that at points where the multiplicity of the singular value $\underline{\sigma}(A - zI)$ is greater than one, $\underline{\sigma}^e$ is not differentiable. By [21, Corollary 9.21], $\hat{\partial}\underline{\sigma}^e(A,z) = \emptyset$ at these points. For points where the multiplicity of the singular value is one, the norm calculation tells us that the only point in $\hat{\partial}\underline{\sigma}^e(A,z)$ has norm at least one; the only element in $\hat{\partial}\underline{\sigma}^e(A,z)$ is of the form $\left(uv^H, -v^H u\right)$, and the matrix part already contributes one to the norm. So it is impossible that $(0,0) \in \partial\underline{\sigma}^e(A,z)$.

Next, we move on to $\partial(-\underline{\sigma}^e)(A,z)$. Take $\bar{U}, \bar{V}$ to be the matrix corresponding to the left and right singular vectors of $A - zI$ in the sense of Proposition 3.3. Note that $\left(\bar{U}\bar{V}^H, 0\right)$ represents a direction of linear descent, as

$$-\underline{\sigma}^e\left(A + \epsilon\bar{U}\bar{V}^H, z\right) = -\underline{\sigma}^e(A,z) - \epsilon$$

for small ϵ, and so we have $(0,0) \notin \hat{\partial}(-\underline{\sigma}^e)(A,z)$. Due to regularity (Proposition 4.3), we have $(0,0) \notin \partial(-\underline{\sigma}^e)(A,z)$. ∎

Despite the fact that $\underline{\sigma}^e$ is not regular, we are still able to calculate the subdifferential $\partial\underline{\sigma}^e(A,z)$.

PROPOSITION 4.14. *If* $z \notin \Lambda(A)$, *then*

$$\partial\underline{\sigma}^e(A,z) = \left\{\left(uv^H, -v^H u\right) \mid (u,v) \in MSV(A - zI)\right\}.$$

Proof. We observe that

$$\begin{aligned}\partial\underline{\sigma}^e(A,z) &\subset -\partial(-\underline{\sigma}^e)(A,z)\\ &= \operatorname{conv}\left\{\left(uv^H, -v^H u\right) \mid (u,v) \in MSV(A - zI)\right\}\end{aligned}$$

by [21, Corollary 9.21] and Proposition 4.3. Next, note that if $(B,w) \in \partial\underline{\sigma}^e(A,z)$, then

$$(B,w) \in \operatorname{conv}\left\{\left(uv^H, -v^H u\right) \mid (u,v) \in MSV(A - zI)\right\},$$

and so we may write $(B,w) = \sum_{i=1}^k \lambda_i\left(u_i v_i^H, -v_i^H u_i\right)$ for a convex combination of left and right singular vectors u_i, v_i corresponding to the smallest singular value. But since the 2-norm is a strictly convex norm, $\|B\| < 1$ if $k > 1$ and (u_i, v_i)'s are not complex multiples each other. We take a closer look: (B,w) can be written as a limit of $(B_i, w_i) = \nabla\underline{\sigma}^e(A_i, z_i)$, where $(A_i, z_i) \to (A,z)$ by [21, Corollary 9.21]. Since $\|B_i\| = 1$, it follows that $\|B\| = 1$.

With this, we conclude that $(B,w) = (uv^*, -v^*u)$ for some $(u,v) \in MSV(A - zI)$, and so

$$\partial\underline{\sigma}^e(A,z) \subset \left\{\left(uv^H, -v^H u\right) \mid (u,v) \in MSV(A - zI)\right\}.$$

To prove the other containment, note that, for any $(u,v) \in MSV(A - zI)$, we have

$$\begin{aligned}\hat{\partial}\underline{\sigma}^e\left(A - \delta uv^H, z\right) &= \left\{\nabla\underline{\sigma}^e\left(A - \delta uv^H, z\right)\right\}\\ &= \left\{\left(uv^H, -v^H u\right)\right\}\end{aligned}$$

for $0 < \delta < \epsilon$ by Lemma 4.2. Taking limits as $\delta \downarrow 0$, we have $\left(uv^H, -v^H u\right) \in \partial \underline{\sigma}^e (A, z)$, which completes the proof. ∎

5. Main result. Before proving our main result, we make a statement about the normal cones $N_{\operatorname{gph} \Lambda_\epsilon^c} (A, z)$ and $N_{\operatorname{gph} \Lambda_\epsilon} (A, z)$. We make use of properties that we have established in section 4 to establish the link between level sets and normal vectors.

PROPOSITION 5.1. *If* $\epsilon = \underline{\sigma}^e (A, z) > 0$, *then*

$$N_{\operatorname{gph} \Lambda_\epsilon^c} (A, z) = \operatorname{pos} \operatorname{conv} \left\{ \left(-uv^H, v^H u\right) \mid (u, v) \in MSV (A - zI) \right\},$$
$$N_{\operatorname{gph} \Lambda_\epsilon} (A, z) = \operatorname{pos} \left\{ \left(uv^H, -v^H u\right) \mid (u, v) \in MSV (A - zI) \right\}.$$

Proof. Apply a result on level sets [21, Proposition 10.3], Proposition 4.13, and the fact that $-\underline{\sigma}^e$ is Lipschitz to get

$$N_{\operatorname{gph} \Lambda_\epsilon^c} (A, z) = \operatorname{pos} \left(\partial \left(-\underline{\sigma}^e\right) (A, z)\right).$$

Next, apply Proposition 4.3 to deduce the first result.

By [21, Proposition 10.3] and Proposition 4.14, we have

$$\begin{aligned} N_{\operatorname{gph} \Lambda_\epsilon} (A, z) &\subset \operatorname{pos} \partial \underline{\sigma}^e (A, z) \\ &= \operatorname{pos} \left\{ \left(uv^H, -v^H u\right) \mid (u, v) \in MSV (A - zI) \right\}. \end{aligned}$$

Furthermore, if $\underline{\sigma} (A - zI)$ is simple, then $\underline{\sigma}^e$ is smooth and regular at (A, z) by Lemma 4.2, and so the above inclusion holds with equality.

For the opposite containment, take any $(u, v) \in MSV (A - zI)$. Consider the pair

$$(A_\delta, z_\delta) := \left((1 + \delta) A - \epsilon \delta u v^H, (1 + \delta) z\right) \text{ for small } \delta > 0.$$

At these points, $\underline{\sigma}^e$ is smooth (and thus regular) because the singular value is of multiplicity one with corresponding singular vectors (u, v), and $\underline{\sigma}^e (A_\delta, z_\delta) = \epsilon$. Thus

$$\left(uv^H, -v^H u\right) \in \hat{N}_{\operatorname{gph} \Lambda_\epsilon} \left((1 + \delta) A - \epsilon \delta u v^H, (1 + \delta) z\right).$$

Taking $\delta \downarrow 0$, we see that $\left(uv^H, -v^H u\right) \in N_{\operatorname{gph} \Lambda_\epsilon} (A, z)$. Since $N_{\operatorname{gph} \Lambda_\epsilon} (A, z)$ is a cone, we have the formula for $N_{\operatorname{gph} \Lambda_\epsilon} (A, z)$ as claimed. ∎

The following is the main result that summarizes the links between Figure 1 in the introduction.

THEOREM 5.2. *Consider a point* $z \notin \Lambda (A)$. *Let* $\epsilon = \underline{\sigma}^e (A, z)$. *Then the following are equivalent:*

(1) z *is not resolvent-critical for* A.
(2) Λ_ϵ^c *has the Aubin property at* A *for* z.
(3) Λ_ϵ *has the Aubin property at* A *for* z.

Proof. For the purposes of the proof, we introduce several other properties:

(4) $(M^n \times \{0\}) \cap N_{\operatorname{gph} \Lambda_\epsilon^c} (A, z) = \{0\}$.
(5) $D^* \Lambda_\epsilon^c (A \mid z) (0) = \{0\}$.
(6) $(M^n \times \{0\}) \cap N_{\operatorname{gph} \Lambda_\epsilon} (A, z) = \{0\}$.
(7) $D^* \Lambda_\epsilon (A \mid z) (0) = \{0\}$.

Properties (4) and (5) are equivalent because $\alpha \in D^* \Lambda_\epsilon^c (A \mid z) (\beta)$ if and only if $(\alpha, -\beta) \in N_{\operatorname{gph} \Lambda_\epsilon^c} (A, z)$ by the definition of coderivatives [21, Definition 8.33].

Properties (5) and (2) are equivalent by the Mordukhovich Criterion [21, Theorem 9.40]. The same goes for properties (6), (7), and (3).

Next, we show the equivalence of properties (1) and (4). We apply Proposition 5.1 to reduce property (4) to

$$(M^n \times \{0\}) \cap \operatorname{pos conv} \left\{ \left(-uv^H, v^H u\right) \mid (u,v) \in MSV\,(A - zI) \right\} = \{0\}\,.$$

$(1 \Rightarrow 4)$ Suppose that z is not resolvent-critical, that is, $0 \notin Y\,(A - zI)$, and yet property (4) fails. Then there is some nonzero pair with second coordinate (the one in $\mathbb{C}$) zero lying in

$$\operatorname{pos conv} \left\{ \left(-uv^H, v^H u\right) \mid (u,v) \in MSV\,(A - zI) \right\}.$$

This means that there is a convex combination of pairs $\left(-uv^H, v^H u\right)$ such that their second coordinate is zero. Then $0 \in Y\,(A - zI)$ (appealing to Proposition 3.4), a contradiction.

$(1 \Leftarrow 4)$ If property (1) fails, there are minimal left and right singular vectors u, v such that $v^H u = 0$, and then $\left(-uv^H, v^H u\right)$ is a nonzero element in

$$(M^n \times \{0\}) \cap \operatorname{pos conv} \left\{ \left(-uv^H, v^H u\right) \mid (u,v) \in MSV\,(A - zI) \right\}.$$

So we have proved the equivalence of properties (1) and (4). We proceed to prove the equivalence of properties (1) and (6). We lose regularity, but nevertheless, the proof still looks similar.

$(1 \Rightarrow 6)$ We prove $(4 \Rightarrow 6)$. If $0 \notin Y\,(A - zI)$, then $(M^n \times \{0\}) \cap N_{\operatorname{gph} \Lambda_\epsilon^c}\,(A, z) = \{0\}$. But Proposition 5.1 gives

$$\begin{aligned} \{0\} &\subset (M^n \times \{0\}) \cap N_{\operatorname{gph} \Lambda_\epsilon}\,(A, z) \\ &\subset (M^n \times \{0\}) \cap -N_{\operatorname{gph} \Lambda_\epsilon^c}\,(A, z) \\ &= \{0\}\,. \end{aligned}$$

$(1 \Leftarrow 6)$. If property (1) fails, there are minimal left and right singular vectors u, v such that $v^H u = 0$, and thus $\left(uv^H, -v^H u\right)$ is a nonzero element in $(M^n \times \{0\}) \cap N_{\operatorname{gph} \Lambda_\epsilon}\,(A, z)$. ∎

When we consider fixing the matrix A and increasing ϵ, it is natural to ask whether the map $\epsilon \mapsto \Lambda_\epsilon\,(A)$ is Lipschitz.

PROPOSITION 5.3. *Given $z \in \mathbb{C}$, the map $\epsilon \mapsto \Lambda_\epsilon\,(A)$ has the Aubin property at $\underline{\sigma}_A\,(z)$ for z if and only if $0 \notin \partial \underline{\sigma}_A\,(z)$, whereas the map $\epsilon \mapsto \Lambda_\epsilon^c\,(A)$ has the Aubin property at $\underline{\sigma}_A\,(z)$ for z if and only if $0 \notin \partial\,(-\underline{\sigma}_A)\,(z)$ (or equivalently, assuming $z \notin \Lambda\,(A)$, z is not resolvent-critical for A).*

Proof. A straightforward application of [21, Theorem 9.41(b)] on $\underline{\sigma}_A$ gives us $0 \notin \partial \underline{\sigma}_A\,(z)$ if and only if the map $\epsilon \mapsto \operatorname{lev}_{\leq \epsilon} \underline{\sigma}_A = \Lambda_\epsilon\,(A)$ has the Aubin property at ϵ for z, which is the first part of what we seek to prove. The second part is similar, using Proposition 4.5. ∎

A particular example worked out in full detail exploiting this is highlighted in [6].

It is natural to ask whether there are any differences between Theorem 5.2 and the two parts of Proposition 5.3, and it comes down to comparing $\partial\,(-\underline{\sigma}_A)$ and $\partial \underline{\sigma}_A$. In general, if z is not an eigenvalue of A,

$$-\partial \underline{\sigma}_A\,(z) \subset \partial\,(-\underline{\sigma}_A)\,(z) = Y\,(A - zI)$$

by Proposition 4.5 and [21, Corollary 9.21], but the inclusion can be strict. Consider the matrix $\bar{A} = \operatorname{diag}(1, -1, i, -i)$ in Example 2.6. Here,

$$\partial\left(-\underline{\sigma}_{\bar{A}}\right)(0) = \{a + bi \mid |a| + |b| \leq 1\},$$

so 0 is resolvent-critical while $\partial\underline{\sigma}_{\bar{A}}(0) = \{1, -1, i, -i\}$.

6. Lipschitz continuity of pseudospectra. The results in the last section study the Aubin property of the pseudospectra Λ_ϵ. The next natural step is to evaluate the graphical modulus and investigate the Lipschitz continuity of Λ_ϵ.

If $\underline{\sigma}(A - zI) = \epsilon > 0$, then from Proposition 5.1 and the definition of the coderivative, we can deduce the formula for $D^*\Lambda_\epsilon^c(A \mid z)(c)$. To keep the expressions compact, we understand that (u_i, v_i) ranges over $MSV(A - zI)$ whenever u_i, v_i appear in the formulas below. We have

$$\begin{aligned}
&D^*\Lambda_\epsilon^c(A \mid z)(c)\\
&= \left\{-k\sum_i \lambda_i u_i v_i^H \mid c = -k\sum_i \lambda_i v_i^H u_i, \sum_i \lambda_i = 1, \lambda_i \geq 0, k \geq 0\right\}\\
&= \begin{cases} \left\{c\frac{\sum_i \lambda_i u_i v_i^H}{\sum_i \lambda_i v_i^H u_i} \mid \sum_i \lambda_i v_i^H u_i \neq 0\right\} & \text{if } c \neq 0,\\ \operatorname{pos}\left\{\sum_i \lambda_i u_i v_i^H \mid \sum_i \lambda_i v_i^H u_i = 0\right\} & \text{if } c = 0, \end{cases}
\end{aligned}$$

and

$$\begin{aligned}
&D^*\Lambda_\epsilon(A \mid z)(c)\\
&= \left\{kuv^H \mid c = kv^H u, k \geq 0, (u, v) \in MSV(A - zI)\right\}\\
&= \begin{cases} \left\{c\frac{uv^H}{v^H u} \mid (u, v) \in MSV(A - zI), v^H u \neq 0\right\} & \text{if } c \neq 0,\\ \operatorname{pos}\left\{uv^H \mid (u, v) \in MSV(A - zI), v^H u = 0\right\} & \text{if } c = 0. \end{cases}
\end{aligned}$$

We can then calculate the graphical moduli for Λ_ϵ and Λ_ϵ^c in the theorem below.

THEOREM 6.1. *We have the following graphical moduli:*

$$\operatorname{lip}\Lambda_\epsilon(A \mid z) = \begin{cases} 1/d(0, Y(A - zI)) & \text{if } \underline{\sigma}(A - zI) = \epsilon,\\ 0 & \text{if } \underline{\sigma}(A - zI) < \epsilon, \end{cases}$$

$$\operatorname{lip}\Lambda_\epsilon^c(A \mid z) = \begin{cases} 1/d(0, Y(A - zI)) & \text{if } \underline{\sigma}(A - zI) = \epsilon,\\ 0 & \text{if } \underline{\sigma}(A - zI) > \epsilon. \end{cases}$$

(Here, we interpret $1/0 = +\infty$*.)*

Proof. It is clear that if $\underline{\sigma}(A - zI) < \epsilon$, then (A, z) lies in the interior of $\operatorname{gph}\Lambda_\epsilon$, so $N_{\operatorname{gph}\Lambda_\epsilon}(A, z) = \{(0, 0)\}$, and so

$$\operatorname{lip}\Lambda_\epsilon(A \mid z) = |D^*\Lambda_\epsilon(A \mid z)|^+ = 0.$$

Similarly, $\operatorname{lip}\Lambda_\epsilon^c(A \mid z) = 0$ if $\underline{\sigma}(A - zI) > \epsilon$.

If $\underline{\sigma}(A - zI) = \epsilon$ and $0 \in Y(A - zI)$, then Λ_ϵ and Λ_ϵ^c do not have the Aubin property at A for z, and so

$$\operatorname{lip}\Lambda_\epsilon(A \mid z) = \operatorname{lip}\Lambda_\epsilon^c(A \mid z) = \infty.$$

By the Mordukhovich criterion [21, Theorem 9.40] and the definition of outer norms [21, Section 9D], we have $\operatorname{lip}\Lambda_\epsilon^c(A \mid z)$ to be

$$\sup_{c \neq 0} \sup_{d \in D^*\Lambda_\epsilon^c(A|z)(c)} \frac{\|d\|}{|c|},$$

or, in other words, the infimum of all κ such that

$$d \in D^* \Lambda_\epsilon^c (A \mid z)(c) \implies \|d\| \leq \kappa |c|. \tag{6.1}$$

In view of the formula for $D^* \Lambda_\epsilon^c (A \mid z)$, formula (6.1) is equivalent to

$$\left\| \sum \lambda_i u_i v_i^H \right\| \leq \kappa \left| \sum \lambda_i v_i^H u_i \right| \tag{6.2}$$

for all $(u_i, v_i) \in MSV(A - zI), \lambda_i \geq 0, \sum \lambda_i = 1$. To prove that $\operatorname{lip} \Lambda_\epsilon^c (A \mid z) = 1/d(0, Y(A - zI))$, it remains to prove that formula (6.2) is equivalent to

$$\kappa \geq 1/d(0, Y(A - zI)). \tag{6.3}$$

Suppose that κ satisfies formula (6.2). Then for $y \in Y(A - zI)$, we have some $(u, v) \in MSV(A - zI)$ such that $y = v^H u$. Then

$$\begin{aligned} \kappa |y| &= \kappa \left| v^H u \right| \\ &\geq \left\| u v^H \right\| \\ &= 1. \end{aligned}$$

Formula (6.3) follows. Next, suppose that κ satisfies formula (6.3). If $(u_i, v_i) \in MSV(A - zI)$, $\lambda_i \geq 0$ and $\sum \lambda_i = 1$, we have $\sum \lambda_i v_i^H u_i \in Y(A - zI)$ by the convexity of $Y(A - zI)$. Thus

$$\begin{aligned} \left\| \sum \lambda_i u_i v_i^H \right\| &\leq \sum \lambda_i \left\| u_i v_i^H \right\| \\ &= 1 \\ &\leq \kappa \left| \sum \lambda_i v_i^H u_i \right|. \end{aligned}$$

Formula (6.2) follows, and so $\operatorname{lip} \Lambda_\epsilon^c (A \mid z) = 1/d(0, Y(A - zI))$. Similar and simpler calculations give us $\operatorname{lip} \Lambda_\epsilon (A \mid z) = 1/d(0, Y(A - zI))$. ∎

We next turn to the Lipschitz constant for the pseudospectral mapping Λ_ϵ. We want to find $\operatorname{lip}_\infty \Lambda_\epsilon (\bar{A})$, the Lipschitz modulus of the pseudospectral map at $\bar{A}$. For a set-valued map $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$, we are able to calculate $\operatorname{lip}_\infty S(\bar{x})$ from the graphical modulus easily with the following formula.

PROPOSITION 6.2 (see [20, Theorem 1.42]). *If $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is outer semicontinuous and S is locally bounded at $\bar{x}$, then*

$$\operatorname{lip}_\infty S(\bar{x}) = \max_{y \in S(\bar{x})} \operatorname{lip} S(\bar{x} \mid y).$$

Thus the Lipschitz constants for Λ_ϵ are easily obtained.

PROPOSITION 6.3. *The following expressions are equal:*

(i) $\operatorname{lip}_\infty \Lambda_\epsilon (A)$,

(ii) $\max_{z \in \Lambda_\epsilon(A)} \{\operatorname{lip} \Lambda_\epsilon (A \mid z)\}$,

(iii) $\max_{z : \underline{\sigma}(A - zI) = \epsilon} \{1/d(0, Y(A - zI))\}$,

(iv) $\max_z \left\{ 1/\left| v^H u \right| \mid (u, v) \in MSV(A - zI), \underline{\sigma}(A - zI) = \epsilon \right\}$.

Proof. The expressions (i) and (ii) are equal by Proposition 6.2 and the fact that Λ_ϵ is compact and locally bounded. Then expressions (ii) and (iii) are equal by Theorem 6.1, and expression (iv) is just an expansion of the definition of $Y(\cdot)$ applied to expression (iii). ∎

7. Pseudospectral abscissa and pseudospectral radius. In this section we apply our results on Lipschitz continuity of pseudospectra to reexplore earlier work on the pseudospectral abscissa and pseudospectral radius in [4, 5, 19, 22].

DEFINITION 7.1. *Define the* ϵ-pseudospectral abscissa $\alpha_\epsilon : M^n \to \mathbb{R}$ *by*

$$\alpha_\epsilon(A) = \max_{z \in \Lambda_\epsilon(A)} \operatorname{Re}(z),$$

and the ϵ-pseudospectral radius $\rho_\epsilon : M^n \to \mathbb{R}_+$ *by*

$$\rho_\epsilon(A) = \max_{z \in \Lambda_\epsilon(A)} |z|.$$

Note that if $\epsilon > 0$, then $\rho_\epsilon(A) > 0$. We shall establish the continuity properties of α_ϵ and ρ_ϵ. We begin with another routine piece of theory on parametric minimization.

COROLLARY 7.2 (to [21, Corollary 10.14]). *Suppose that* $F : \mathbb{R}^m \rightrightarrows \mathbb{R}^n$ *is outer semicontinuous and maps to compact sets. Define* $p : \mathbb{R}^m \to \mathbb{R}$ *and* $P : \mathbb{R}^m \rightrightarrows \mathbb{R}^n$ *below by*

$$p(u) = \min_{x \in F(u)} g(x), \quad P(u) = \arg\min_{x \in F(u)} g(x),$$

where the lower semicontinuous function $g : \mathbb{R}^n \to \mathbb{R}$ *is differentiable at all points in* $P(\bar{u})$ *for some given* $\bar{u} \in \mathbb{R}^m$*. Then* p *is*

(a) *Lipschitz continuous around* $\bar{u}$ *if* F *has the Aubin property at* $\bar{u}$ *for* $\bar{x}$ *for all* $\bar{x} \in P(\bar{u})$*, with*

$$\operatorname{lip} p(\bar{u}) \le \max\{|y| : y \in S\} < \infty,$$

where $S = \{y \mid \bar{x} \in P(\bar{u}), y \in D^*F(\bar{u} \mid \bar{x})(\nabla g(\bar{x}))\}$*;*

(b) *strictly differentiable at* $\bar{u}$ *with* $\nabla p(\bar{u}) = \bar{y}$ *if* $S = \{\bar{y}\}$.

Proof. Let $f : \mathbb{R}^n \times \mathbb{R}^m \to \bar{\mathbb{R}}$ be defined by

$$f(x,u) = \delta_{\operatorname{gph} F}(u,x) + g(x) = \begin{cases} g(x) & \text{if } x \in F(u), \\ \infty & \text{otherwise.} \end{cases}$$

Then

$$p(u) = \inf_x f(x,u), \quad P(u) = \arg\min_x f(x,u).$$

Since F is outer semicontinuous, $\operatorname{gph} F$ is closed, so f is proper and lower semicontinuous.

Next, we prove f is level bounded in x locally uniformly in u. That is, for each $\bar{u} \in \mathbb{R}^m$ and $\alpha \in \mathbb{R}$, there is a neighborhood V of $\bar{u}$ along with a bounded set $B \subset \mathbb{R}^n$ such that $\{x \mid f(x,u) \le \alpha\} \subset B$ for all $u \in V$. Note that $f(x,u) \le \alpha$ means that $x \in F(u)$ and $g(x) \le \alpha$. Since F is outer semicontinuous, choose V such that $F(u) \subset F(\bar{u}) + \mathbb{B}$ for all $u \in V$, by the characterization of outer semicontinuity in [21, Proposition 5.12]. The set B can be chosen to be $F(\bar{u}) + \mathbb{B}$, and we are done.

Following the notation in [21, Corollary 10.13], for any $\bar{x} \in P(\bar{u})$,

$$\begin{aligned} M(\bar{x},\bar{u}) &:= \{y \mid (0,y) \in \partial f(\bar{x},\bar{u})\} \\ &= \{y \mid (y,0) \in \partial \delta_{\operatorname{gph} F}(\bar{u},\bar{x}) + \{(0, \nabla g(\bar{x}))\}\} \\ &\qquad \text{(by [21, Exercise 8.8(c)])} \\ &= \{y \mid (y, -\nabla g(\bar{x})) \in N_{\operatorname{gph} F}(\bar{u},\bar{x})\} \\ &\qquad \text{(by [21, Exercise 8.14])} \\ &= D^*F(\bar{u} \mid \bar{x})(\nabla g(\bar{x})) \\ &\qquad \text{(by [21, Definition 8.33]).} \end{aligned}$$

Also,

$$\begin{aligned} M_\infty(\bar{x},\bar{u}) &:= \{y \mid (0,y) \in \partial^\infty f(\bar{x},\bar{u})\} \\ &= \{y \mid (y,0) \in \partial^\infty \delta_{\mathrm{gph} F}(\bar{u},\bar{x})\} \\ &= \{y \mid (y,0) \in N_{\mathrm{gph}\, F}(\bar{u},\bar{x})\} \\ &= D^* F(\bar{u} \mid \bar{x})(0). \end{aligned}$$

This means that $Y_\infty(\bar{u}) := \bigcup_{\bar{x}\in P(\bar{u})} M_\infty(\bar{x},\bar{u}) = \{0\}$, so part (a) of [21, Corollary 10.14] applies. Furthermore, $Y(\bar{u})$, where $Y(\cdot)$ is defined in [21, Corollary 10.13], is

$$\begin{aligned} Y(\bar{u}) &:= \bigcup_{\bar{x}\in P(\bar{u})} M(\bar{x},\bar{u}) \\ &= \bigcup_{\bar{x}\in P(\bar{u})} D^* F(\bar{u} \mid \bar{x})(\nabla g(\bar{x})), \end{aligned}$$

and so

$$\begin{aligned} \mathrm{lip}\, p(\bar{u}) &\le \max_{y\in Y(\bar{u})} |y| \\ &= \max\{|y| : \bar{x} \in P(\bar{u}), y \in D^* F(\bar{u} \mid \bar{x})(\nabla g(\bar{x}))\} < \infty. \end{aligned}$$

The rest of the claim follows by [21, Corollary 10.14]. □

The continuity of α_ϵ and ρ_ϵ can be proved by the following proposition when the conditions for Lipschitz continuity are absent. The proof is routine.

PROPOSITION 7.3. *Suppose that $F : \mathbb{R}^m \rightrightarrows \mathbb{R}^n$ is continuous and maps to compact sets. If p, P, and g are defined as in Corollary 7.2 with g continuous, then p is continuous and P is outer semicontinuous.*

As a consequence of Corollary 7.2, we obtain the following result.

COROLLARY 7.4. *The pseudospectral abscissa α_ϵ and pseudospectral radius ρ_ϵ are Lipschitz continuous at a matrix A if $\mathrm{lip}_\infty \Lambda_\epsilon(A) < \infty$, with Lipschitz constants bounded above by $\mathrm{lip}_\infty \Lambda_\epsilon(A)$.*

Proof. Following the notation in Corollary 7.2, take $F = \Lambda_\epsilon$ and $g(x) = \langle -1, x\rangle$. Then $\alpha_\epsilon = -p$, and we obtain

$$\begin{aligned} \mathrm{lip}\, \alpha_\epsilon(A) &\le \max\{|y| : y \in D^*\Lambda_\epsilon(A \mid z)(-1) \\ &\qquad , z \in \Lambda_\epsilon(A), \mathrm{Re}(z) = \alpha_\epsilon(A)\} \\ &= \max\{1/d(0, \mathbb{R}_- \cap Y(A - zI)) : \\ &\qquad z \in \Lambda_\epsilon(A), \mathrm{Re}(z) = \alpha_\epsilon(A)\} \end{aligned}$$

using our derivative computation before Theorem 6.1. If we take $g(x) = -|x|$ instead, then $\rho_\epsilon = -p$, and

$$\begin{aligned} \mathrm{lip}\, \rho_\epsilon(A) &\le \max\left\{|y| : y \in D^*\Lambda_\epsilon(A \mid z)\left(-\frac{z}{|z|}\right), \right. \\ &\qquad \left. z \in \Lambda_\epsilon(A), |z| = \rho_\epsilon(A)\right\} \\ &= \max\left\{1/d\left(0, \mathbb{R}_+\left(\frac{z}{|z|}\right) \cap Y(A - zI)\right) : \right. \\ &\qquad \left. z \in \Lambda_\epsilon(A), |z| = \rho_\epsilon(A)\right\}. \end{aligned}$$

The upper bounds for lip $\alpha_\epsilon(A)$ and lip $\rho_\epsilon(A)$ obtained above are both not greater than $\text{lip}_\infty \Lambda_\epsilon(A)$ by Proposition 6.3, and so we are done. ∎

8. Resolvent-critical points. Resolvent-critical points are crucial throughout our analysis. They are also, for example, explicitly excluded in the analysis of the quadratic convergence of the algorithm for finding the pseudospectral abscissa in [5]. We investigate their properties further.

PROPOSITION 8.1. *All resolvent-critical points lie in the numerical range of A.*

Proof. Suppose that z is resolvent-critical. Then there exists a right singular vector v of $(A - zI)$ such that $v^H(A - zI)v = 0$, which implies that $v^H Av = zv^H v = z$ if $|v| = 1$. This means that z lies in the numerical range of A. ∎

PROPOSITION 8.2. *For ϵ large enough such that $\Lambda_\epsilon(A)$ contains the numerical range of A, $W(A)$, in its interior, the map $\Lambda_\epsilon : M^n \rightrightarrows \mathbb{C}$ is strictly continuous at A for any point in $\Lambda_\epsilon(A)$, and thus Lipschitz continuous at a neighborhood of A. For α_ϵ and ρ_ϵ to be Lipschitz continuous, we just need the interior of* $\operatorname{conv}\Lambda_\epsilon(\mathrm{A})$ *to contain $W(A)$.*

Proof. For the first part, if $\Lambda_\epsilon(A)$ contains $W(A)$ in its interior, then the points in the boundary of Λ_ϵ are not resolvent-critical by the previous result. Apply Proposition 6.3.

For the second part, by the proof of Corollary 7.4, it suffices to show that if z satisfies $\text{Re}\, z = \alpha_\epsilon(A)$ and $\underline{\sigma}(A - zI) = \epsilon$, then $z \notin W(A)$. But if z satisfies these conditions, then $z \in \operatorname{conv}\Lambda_\epsilon(A)$. The same goes for ρ_ϵ. ∎

Remark 8.3. In Table 3 in page 1061, the third example of a 5×5 matrix illustrates that a resolvent-critical can lie outside the convex hull of the spectrum of A. There is a resolvent-critical point close to 0, but the convex hull of the eigenvalues is just $\{-1\}$.

With all that we have done so far, the following is a natural consequence of [3, Corollary 8].

COROLLARY 8.4 (to [3, Corollary 8]). *Given a matrix A, the set of resolvent-critical values $\{\underline{\sigma}_A(z) \mid z$ resolvent critical for $A\}$ is finite.*

Proof. This is just the (semialgebraic) set of Clarke-critical values of $\underline{\sigma}_A$ by Theorem 4.9, which is finite by [3, Corollary 8]. ∎

With the above result, we arrive at the following appealing result.

COROLLARY 8.5. *Given a matrix A, the mappings Λ_ϵ, α_ϵ, and ρ_ϵ are Lipschitz around A for all but finitely many $\epsilon \geq 0$, so, in particular, for all small $\epsilon > 0$.*

Proof. This is a direct consequence of Theorem 5.2 and Corollaries 8.4 and 7.4. ∎

Remark 8.6. The conditions that guarantee Lipschitz continuity of the pseudospectral abscissa α_ϵ in the result above are much more general than the conditions in [4, Corollary 8.3]. Firstly, we do not need the assumption that active eigenvalues are nonderogatory made in [4, Corollary 8.3], and our current result holds for all but finitely many ϵ.

Here is another general observation on resolvent-critical points.

THEOREM 8.7. *For a fixed A, the set of resolvent-critical points is compact, semialgebraic with empty interior, and contains eigenvalues as isolated points.*

Proof. Denote the set of resolvent-critical points by S_A. The set S_A is bounded by Proposition 8.1. It is clear that S_A is semialgebraic. As $\underline{\sigma}_A$ is Lipschitz, $\partial^\circ(-\underline{\sigma}_A)$ has closed graph by [9, Proposition 2.1.5(b)], and thus S_A is closed.

Suppose that S_A does not have empty interior. Note that $\underline{\sigma}_A$ has to be constant on a component by Corollary 8.4, and this would mean that $\underline{\sigma}_A$ is constant on a set of

nonempty interior, which contradicts the fact that $\underline{\sigma}_A$ cannot have minimizers other than at the eigenvalues of A [4, Theorem 4.2]. Thus S_A has empty interior.

Lastly, S_A can be written as a union of curves and points in $\mathbb{C}$. If an eigenvalue, say $\bar{z}$, is not an isolated point in S_A, then it is on some curve. This would mean that $\underline{\sigma}_A$ is zero on a curve, which contradicts the fact that $\underline{\sigma}_A$ is zero only on the set of eigenvalues, which is a finite set. Thus all eigenvalues are isolated in S_A. ∎

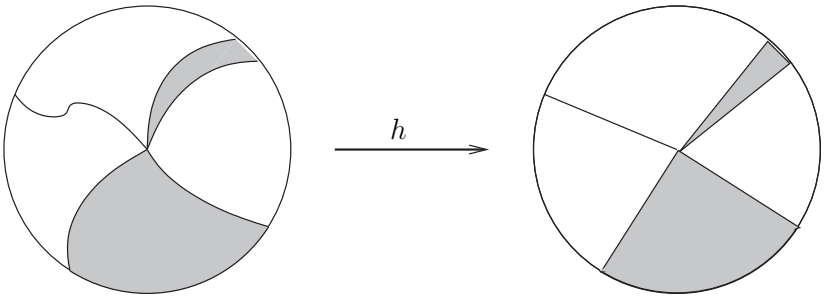
We call $\Lambda'_\epsilon(A) = \{z \mid \underline{\sigma}(A - zI) < \epsilon\}$ the *strict pseudospectrum* of A. The set $\Lambda'_\epsilon(A)$ consists of at most n components (since each component must contain an eigenvalue [22]), and the number of components is clearly a decreasing function of ϵ. There will be some points $\bar{z} \in \mathbb{C}$ where some components meet as ϵ increases. If $\Lambda'_\epsilon(A)$ has n components and $\bar{z}$ lies on the boundary of two components of $\Lambda'_\epsilon(A)$, then the distance between A and the set of matrices with repeated eigenvalues is ϵ and is attained by some matrix $\bar{A}$ having $\bar{z}$ as a repeated eigenvalue (see [1, Theorem 5.1]): It turns out that such points are resolvent-critical as the next theorem will show, generalizing [1, Proposition 4.10].

THEOREM 8.8. *If $\bar{z}$ is a common boundary point of two or more distinct components of $\Lambda'_\epsilon(A)$, then $\bar{z}$ is a resolvent-critical point.*

Proof. To reduce notation, let us assume that $\bar{z} = 0$. The rest of the proof will follow by a translation. We look at the structure of $\Lambda_\epsilon(A)$ around 0, where $\epsilon > 0$. Since $\Lambda_\epsilon(A)$ is semialgebraic, $\Lambda_\epsilon(A)$ is locally conic about 0 by [11, Theorem 4.10]. That is, there is an $r > 0$ and a semialgebraic homeomorphism

$$h : \Lambda_\epsilon(A) \cap r\mathbb{B} \to [0,1](\Lambda_\epsilon(A) \cap r(\text{bdry } \mathbb{B}))$$

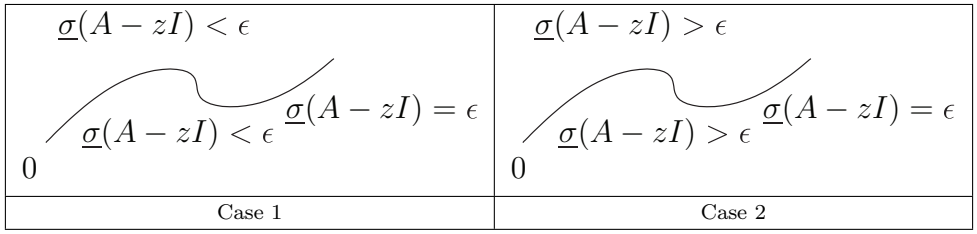
between the two spaces. Since $(\Lambda_\epsilon(A) \cap r(\text{bdry } \mathbb{B}))$ is a finite union of arcs, it follows that the boundary of $\Lambda_\epsilon(A) \cap r\mathbb{B}$ would consist of curves which start from 0 and end at somewhere on $r(\text{bdry } \mathbb{B})$. The diagram below illustrates this.



We use a proof by contradiction. Suppose that 0 is not resolvent-critical. Then $0 \notin Y(A)$, and by Proposition 4.12, $\Lambda^c_\epsilon(A)$ is Clarke regular at 0, with normal cone $N_{\Lambda^c_\epsilon(A)}(0) = \mathbb{R}_+ Y(A)$. Note that $N_{\Lambda^c_\epsilon(A)}(0)$ is pointed, otherwise $0 \in Y(A)$, contradicting the assumption that 0 is not resolvent-critical.

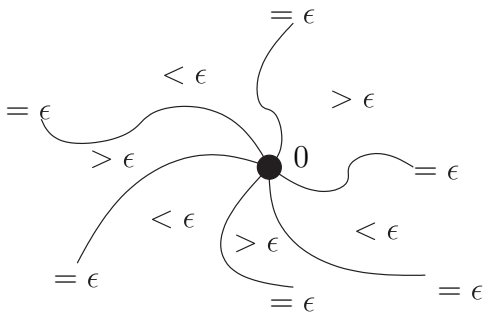
The set $\{z \mid \underline{\sigma}(A - zI) = \epsilon\}$ is semialgebraic and has empty interior since the only local minimizers of $\underline{\sigma}_A$ are eigenvalues of A [4, Theorem 4.2], and so it is a union of smooth curves. We now prove that the curves are boundaries of both $\Lambda_\epsilon(A)$ and $\Lambda^c_\epsilon(A)$. By considering the sign of $\underline{\sigma}_A - \epsilon$ on either side of such a curve, we distinguish three cases. In the following diagram, both Case 1 and Case 2 cannot hold, because the local maxima and local minima of $\underline{\sigma}_A$ are resolvent-critical, and this would make 0

resolvent-critical as well, since the set of resolvent-critical points is closed by Theorem 8.7.

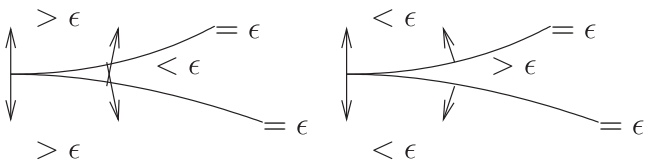


Case 1 | Case 2

Therefore, the general diagram would be as below, with the value of $\underline{\sigma}_A$ alternating above and below ϵ as we circle the origin, crossing the curves where $\underline{\sigma}_A = \epsilon$.



Two different arcs cannot be tangent at 0 since $N_{\Lambda_\epsilon^c(A)}(0)$ will otherwise not be pointed, as the diagrams below show.



Since $\Lambda_\epsilon^c(A)$ is Clarke regular at 0, its tangent cone $T_{\Lambda_\epsilon^c(A)}(0)$ is convex, so the picture above can contain only one sector where $\underline{\sigma}_A > \epsilon$. It now follows that 0 cannot be the boundary point of two components of $\Lambda_\epsilon'(A)$. This completes the proof. ∎

If we can prove the following about the pseudospectral abscissa α_ϵ, then we can conclude that the pseudospectral abscissa is Lipschitz continuous.

CONJECTURE 8.9. *The points where the pseudospectral abscissa α_ϵ are attained are not resolvent-critical.*

A natural question to ask after Theorem 8.7 is the following.

CONJECTURE 8.10. *The number of resolvent-critical points is finite.*

Acknowledgments. We wish to thank Michael Overton for discussions about much of section 8, in particular, leading to Corollary 8.5. We also thank Diethard Klatte and two anonymous referees for a wide range of comments and suggestions, which greatly improved the article.

REFERENCES

[1] R. Alam and S. Bora, *On sensitivity of eigenvalues and eigendecompositions of matrices*, Linear Algebra Appl., 396 (2005), pp. 273–301.

[2] B. Bank, J. Guddat, D. Klatte, B. Kummer, and K. Tammer, *Nonlinear Parametric Optimization*, Akademie-Verlag, Berlin, 1982.

[3] J. Bolte, A. Daniilidis, A. S. Lewis, and M. Shiota, *Clarke critical values of subanalytic Lipschitz continuous functions*, Ann. Polon. Math., 87 (2005), pp. 13–25.

[4] J. V. Burke, A. S. Lewis, and M. L. Overton, *Optimization and pseudospectra, with applications to robust stability*, SIAM J. Matrix Anal. Appl., 25 (2003), pp. 80–104. Corrigendum available online from www.cs.nyu.edu/cs/faculty/overton/papers/pseudo_corrigendum.html.

[5] J. V. Burke, A. S. Lewis, and M. L. Overton, *Robust stability and a criss-cross algorithm for pseudospectra*, IMA J. Numer. Anal., 23 (2003), pp. 359–375.

[6] J. V. Burke, A. S. Lewis, and M. L. Overton, *Convexity and Lipschitz behavior of small pseudospectra*, SIAM J. Matrix Anal. Appl., 29 (2007), pp. 586–595.

[7] J. V. Burke, A. S. Lewis, and M. L. Overton, *Spectral conditioning and pseudospectral growth*, Numer. Math., 107 (2007), pp. 27–37.

[8] F. Chaitin-Chatelin, A. Harrabi, and A. Ilahi, *About Hölder condition numbers and the stratification diagram for defective eigenvalues*, Math. Comput. Simulation, 54 (2000), pp. 397–402.

[9] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983. Reprinted by SIAM, Philadelphia, 1990.

[10] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. Wolenski, *Nonsmooth Analysis and Control Theory*, Springer, New York, 1998.

[11] M. Coste, *An Introduction to O-Minimal Geometry*, http://perso.univ-rennes1.fr/michel.coste/articles.html.

[12] J. W. Demmel, *A counterexample for two conjectures about stability*, IEEE Trans. Automat. Control, 32 (1987), pp. 340–342.

[13] M. Embree and L. N. Trefethen, *Pseudospectra Gateway*, http://web.comlab.ox.ac.uk/projects/pseudospectra/.

[14] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.

[15] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, 1991.

[16] P. Huard, *Background to point-to-set maps in mathematical programming*, Math. Program. Stud., 10 (1979), pp. 1–28.

[17] M. Karow, *Geometry of Spectral Value Sets*, Ph.D. Thesis, University of Bremen, Bremen, Germany, 2003.

[18] M. Karow, *Eigenvalue condition numbers and a formula of Burke, Lewis and Overton*, Electron. J. Linear Algebra, 15 (2006), pp. 143–153.

[19] E. Mengi and M. Overton, *Algorithms for the computation of the pseudospectral radius and the numerical radius of a matrix*, IMA J. Numer. Anal., 25 (2005), pp. 648–669.

[20] B. S. Mordukhovich, *Variational Analysis and Generalized Differentiation*, Vols I and II, Springer, New York, 2006.

[21] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer, Berlin, 1998.

[22] L. N. Trefethen and M. Embree, *Spectra and Pseudospectra*, Princeton University Press, Princeton, NJ, 2005.

[23] G. A. Watson, *Characterization of the subdifferential of some matrix norms*, Linear Algebra Appl., 170 (1992), pp. 33–45.

[24] T. G. Wright, *EigTool: A graphical tool for nonsymmetric eigenproblems*, http://www.comlab.ox.ac.uk/pseudospectra/eigtool/.

SIAM J. OPTIM.
Vol. 19, No. 3, pp. 1073–1092

ON THE SECOND-ORDER FEASIBILITY CONE: PRIMAL-DUAL REPRESENTATION AND EFFICIENT PROJECTION*

ALEXANDRE BELLONI† AND ROBERT M. FREUND‡

Abstract. We study the second-order feasibility cone $\mathcal{F} = \{y \in \mathbb{R}^n : \|My\| \le g^T y\}$ for given data (M, g). We construct a new representation for this cone and its dual based on the spectral decomposition of the matrix $M^T M - gg^T$. This representation is used to efficiently solve the problem of projecting an arbitrary point $x \in \mathbb{R}^n$ onto $\mathcal{F}$: $\min_y\{\|y - x\| : \|My\| \le g^T y\}$, which aside from theoretical interest also arises as a necessary subroutine in the rescaled perceptron algorithm. We develop a method for solving the projection problem to an accuracy ε, whose computational complexity is bounded by $O(mn^2 + n \ln\ln(1/\varepsilon) + n \ln\ln(1/\min\{\text{width}(\mathcal{F}), \text{width}(\mathcal{F}^*)\}))$ operations. Here $\text{width}(\mathcal{F})$ and $\text{width}(\mathcal{F}^*)$ denote the width of $\mathcal{F}$ and $\mathcal{F}^*$, respectively. We also perform computational tests that indicate that the method is extremely efficient in practice.

Key words. second-order cone, convex cone, projection, computational complexity, Newton method

AMS subject classifications. 90C60, 90C51, 90C25, 49M15, 49M29

DOI. 10.1137/06067198X

1. Introduction and main results. Our notation is as follows: let K^* denote the dual of a convex cone $K \subset \mathbb{R}^k$, i.e., $K^* := \{z \in \mathbb{R}^k : z^T y \ge 0 \text{ for all } y \in K\}$. A convex cone K is *regular* if it is closed, has nonempty interior, and contains no lines, in which case K^* is also regular; see Rockafellar [7]. Define the standard second-order cone in $\mathbb{R}^k$ to be $\mathcal{Q}^k := \{y \in \mathbb{R}^k : \|(y_1, \ldots, y_{k-1})\| \le y_k\}$, where $\|\cdot\|$ denotes the Euclidean norm. Let $B(y, r)$ denote the Euclidean ball of radius r centered at y.

Given data $(M, g) \in (\mathbb{R}^{m\times n}, \mathbb{R}^n)$, our interest lies in the second-order feasibility cone

$$\mathcal{F} := \left\{y \in \mathbb{R}^n : \|My\| \le g^T y\right\} = \left\{y \in \mathbb{R}^n : \left(My, g^T y\right) \in \mathcal{Q}^{m+1}\right\}$$

and its dual cone $\mathcal{F}^*$.

We first take care of some trivial cases. When $\text{rank}(M) = 0$, it follows trivially that $\mathcal{F} = \{y \in \mathbb{R}^n : g^T y \ge 0\}$, whereby $\mathcal{F}$ is either a half-space or all of $\mathbb{R}^n$, depending on whether $g \ne 0$ or $g = 0$, respectively. When $\text{rank}(M) = 1$, $M = fc^T$ for some f, c, and $\|My\| = \|f\|\|c^T y|$ for any y. This implies that $\mathcal{F} = \{y \in \mathbb{R}^n : (g - \|f\|c)^T y \ge 0,\ (g + \|f\|c)^T y \ge 0\}$, and hence $\mathcal{F}$ is the intersection of either one or two half-spaces. We dispose of these trivial cases by making the following assumption about the data.

Assumption 1. $\text{rank}(M) \ge 2$ and $g \ne 0$.

We now describe our main representation result for $\mathcal{F}$ and $\mathcal{F}^*$. It is elementary to establish that $M^T M - gg^T$ has at most one negative eigenvalue, and we can write its eigendecomposition as $M^T M - gg^T = QDQ^T$, where Q is orthonormal ($Q^{-1} = Q^T$) and D is the diagonal matrix of eigenvalues. For notational convenience we denote D_i and Q_i as the ith diagonal component of D and the ith column of Q, respectively. By

*Received by the editors October 10, 2006; accepted for publication (in revised form) May 7, 2008; published electronically October 31, 2008. This research has been partially supported through the Singapore-MIT Alliance and an IBM Ph.D. Fellowship.

http://www.siam.org/journals/siopt/19-3/67198.html

†Fuqua School of Business, Duke University, Durham, NC 27708-0120 (abn5@duke.edu).

‡MIT Sloan School of Management, Cambridge, MA 02142 (rfreund@mit.edu).

reordering the columns of Q, we can presume that $D_1 \geq \cdots \geq D_n$ and $D_1, \ldots, D_{n-1} \geq 0$. By choosing either Q_n or $-Q_n$, we can further presume that $g^T Q_n \geq 0$. We implicitly assume Q and D can be computed to within machine precision (in the relative sense) in $O(mn^2)$ operations, consistent with computational practice.

Our interest lies mainly in the case when $\mathcal{F}$ is a regular cone, so we will hypothesize that $\mathcal{F}$ is a regular cone for the remainder of this section. This hypothesis implies that $n-1 \leq \text{rank}(M) \leq \min\{n, m\}$. (We indicate how to amend our results and proofs to relax this hypothesis at the end of sections 2 and 3.) Our main representation result is as follows.

THEOREM 1. *Suppose that $\mathcal{F}$ is a regular cone. Then $D_1, \ldots, D_{n-1} > 0 > D_n$, and*

(i) $\mathcal{F} = \{y : y^T Q D Q^T y \leq 0,\ y^T Q_n \geq 0\}$;
(ii) $\mathcal{F}^* = \{z : z^T Q D^{-1} Q^T z \leq 0,\ z^T Q_n \geq 0\}$;
(iii) *if $y \in \mathcal{F}$ and $\alpha \geq 0$, then $z := -\alpha Q D Q^T y \in \mathcal{F}^*$. Furthermore, if $y \in \partial\mathcal{F}$, then $z \in \partial\mathcal{F}^*$ and $z^T y = 0$;*
(iv) *if $z \in \mathcal{F}^*$ and $\alpha \geq 0$, then $y := -\alpha Q D^{-1} Q^T z \in \mathcal{F}$. Furthermore, if $z \in \partial\mathcal{F}^*$, then $y \in \partial\mathcal{F}$ and $z^T y = 0$.*

Note that (i) and (ii) of Theorem 1 describe easily computable representations of $\mathcal{F}$ and $\mathcal{F}^*$ that have the same computational structure, in that checking membership in each cone uses similar data, operations, etc., in a manner that is symmetric between the dual cones. Parts (iii) and (iv) indicate that the same matrices in (i) and (ii) can be used constructively to map points on the boundary of one cone to their orthogonal counterpart in the dual cone.

Remark 1 (geometry of $\mathcal{F}$ and $\mathcal{F}^*$). Examining (i) and the property that $D_n < 0$, the orthonormal transformation $y \to s := Q^T y$ maps $\mathcal{F}$ onto the axes-aligned ellipsoidal cone $\mathcal{S} := \{s \in \mathbb{R}^n : \sqrt{\sum_{j=1}^{n-1} D_i s_i^2} \leq \sqrt{|D_n|} s_n\}$ so that $\mathcal{F}$ is the image of $\mathcal{S}$ under Q, $\mathcal{F} = \{y : \sqrt{\sum_{i=1}^{n-1} D_i (Q_i^T y)^2} \leq \sqrt{|D_n|} Q_n^T y\}$, and $\mathcal{F}^* = \{z : \sqrt{\sum_{i=1}^{n-1} (1/D_i)(Q_i^T z)^2} \leq \sqrt{1/|D_n|} Q_n^T z\}$. This establishes that $\mathcal{F}$ is indeed simply an ellipsoidal cone whose axes are the eigenvectors of Q with dilations corresponding to the eigenvalues of $M^T M - g g^T$. From this perspective, the representation of $\mathcal{F}^*$ via (ii) makes natural geometric sense. Also, the central axis of both $\mathcal{F}$ and $\mathcal{F}^*$ is the ray $\{\alpha Q_n : \alpha \geq 0\}$. Last of all, note that $-\mathcal{F} = \{y : y^T Q D Q^T y \leq 0,\ y^T Q_n \leq 0\}$ and $-\mathcal{F}^* = \{z : z^T Q D^{-1} Q^T z \leq 0,\ z^T Q_n \leq 0\}$.

It turns out that the eigendecomposition of $M^T M - g g^T = Q D Q^T$, while useful both conceptually and algorithmically (as we shall see), is not even necessary for the above representation of $\mathcal{F}$ and $\mathcal{F}^*$. Indeed, Theorem 1 can alternatively be stated replacing $Q D Q^T$ and $Q D^{-1} Q^T$ by $M^T M - g g^T$ and $(M^T M - g g^T)^{-1}$. Under the further hypothesis that $\text{rank}(M) = n$, the theorem can be restated as follows.

COROLLARY 1. *Suppose that $\mathcal{F}$ is a regular cone and* $\text{rank}(M) = n$. *Then*

(i) $\mathcal{F} = \{y : \sqrt{y^T (M^T M) y} \leq g^T y\}$;
(ii) $\mathcal{F}^* = \{z : \sqrt{z^T (M^T M)^{-1} z} \leq \frac{g^T (M^T M)^{-1} z}{\sqrt{g^T (M^T M)^{-1} g - 1}}\}$;
(iii) *if $y \in \mathcal{F}$ and $\alpha \geq 0$, then $z := -\alpha (M^T M - g g^T) y \in \mathcal{F}^*$. Furthermore, if $y \in \partial\mathcal{F}$, then $z \in \partial\mathcal{F}^*$ and $z^T y = 0$;*
(iv) *if $z \in \mathcal{F}^*$ and $\alpha \geq 0$, then $y := -\alpha [(M^T M)^{-1} - \frac{(M^T M)^{-1} g g^T (M^T M)^{-1}}{g^T (M^T M)^{-1} g - 1}] z \in \mathcal{F}$. Furthermore, if $z \in \partial\mathcal{F}^*$, then $y \in \partial\mathcal{F}$ and $z^T y = 0$.*

The proofs of Theorem 1 and Corollary 1 are presented in section 2, along with proofs that all of the stated quantities are well defined: in particular, D^{-1} exists and $g^T (M^T M)^{-1} g - 1 > 0$ under the given hypotheses.

These representation results are used to solve the following dual pair of optimization problems, where $x \in \mathbb{R}^n$ is a *given* point:

$$\begin{array}{lll} \mathcal{P}: \ t^* := \min_y & \|y - x\| & \mathcal{D}: \ t^* := \max_z \quad -x^T z \\ \text{s.t.} & y \in \mathcal{F}, & \text{s.t.} \quad \|z\| \le 1 \\ & & \quad\ z \in \mathcal{F}^*. \end{array} \tag{1}$$

The problem $\mathcal{P}$ is the classical projection problem onto the cone $\mathcal{F}$, whose solution is the point in $\mathcal{F}$ closest to x, and strong duality is easily established for this pair of problems. The problem $\mathcal{D}$ arises as a necessary subroutine in the rescaled perceptron algorithm in [2]: the subroutine needs to efficiently solve $\mathcal{D}$ using $x = x^k$ that arises at each outer iteration k of the algorithm. It is this latter problem that motivated our interest in efficiently representing $\mathcal{F}^*$ and solving both $\mathcal{P}$ and $\mathcal{D}$. Notice that $\mathcal{P}/\mathcal{D}$ involve intersections of a Euclidean ball and a second-order feasibility cone. This dual pair of problems is therefore a modest generalization of the trust region problem of optimizing a quadratic function over a Euclidean ball, for which Ye [10] showed how to combine a binary search and Newton's method to obtain double-logarithmic complexity. Using the representation results above and extending ideas from [10], we develop an algorithm for solving (1) in section 3. The complexity of the algorithm depends on the *widths* of the cones $\mathcal{F}$ and $\mathcal{F}^*$, where the width τ_K of a cone K is defined to be the radius of the largest ball contained in K that is centered at unit distance from the origin:

$$\tau_K := \max_{y,r} \{r : B(y, r) \subset K, \ \|y\| \le 1\}.$$

It readily follows from Theorem 1 that the widths of $\mathcal{F}$ and $\mathcal{F}^*$ are simple functions of the largest and smallest positive eigenvalues and the negative eigenvalue of $M^T M - gg^T$, and it is straightforward to derive the following:

$$\tau_{\mathcal{F}} = \sqrt{\frac{|D_n|}{|D_n| + D_1}} \quad \text{and} \quad \tau_{\mathcal{F}^*} = \sqrt{\frac{1/|D_n|}{1/|D_n| + 1/D_{n-1}}}.$$

The main complexity result, which is proved in section 3, is as follows.

THEOREM 2. *Suppose that $\mathcal{F}$ is a regular cone, and $x \in \mathbb{R}^n$ satisfying $\|x\| = 1$ is given. Then feasible solutions (y, z) of $(\mathcal{P}, \mathcal{D})$ satisfying a duality gap of at most σ are computable in $O(mn^2 + n \ln\ln(1/\sigma) + n \ln\ln(1/\min\{\tau_{\mathcal{F}}, \tau_{\mathcal{F}^*}\}))$ operations.*

In section 4, we complement this theoretical computational complexity bound with experimental computational results that indicate that the method is also extremely efficient in practice.

Last of all, we note that the hypothesis that $\mathcal{F}$ is regular can be removed with no loss of strength of the results herein but with substantial expositional overhead. The case when $\mathcal{F}$ is nonregular is discussed at the end of sections 2 and 3.

2. Proofs of representation results. Recall the eigendecomposition of $M^T M - gg^T = QDQ^T$, with $D_1 \ge \cdots \ge D_n$. A simple dimension argument establishes that $M^T M - gg^T$ has at most one negative eigenvalue, whereby $D_1, \ldots, D_{n-1} \ge 0$. By choosing either Q_n or $-Q_n$, we can ensure that $g^T Q_n \ge 0$. In preparation for the proof of Theorem 1, we first prove some preliminary results.

PROPOSITION 1. *Suppose that* **int** $\mathcal{F} \ne \emptyset$. *Then $D_n < 0$, and there exists y satisfying $\|My\| < g^T y$.*

Proof. We first suppose that there exists $\bar{y}$ that satisfies $\|M\bar{y}\| < g^T\bar{y}$. In this case it easily follows that $0 > \bar{y}^T(M^TM - gg^T)\bar{y} = \bar{y}^TQDQ^T\bar{y}$, whereby $D_n < 0$. Next suppose that every $y \in \mathcal{F}$ satisfies $\|My\| = g^Ty$, and let $\bar{y} \in \mathbf{int}\ \mathcal{F}$. Since $\bar{y} \in \mathbf{int}\ \mathcal{F}$, we have $\|M(\bar{y} + \beta d)\| = g^T(\bar{y} + \beta d)$ for all $d \in B(0,1)$ and all sufficiently small positive β. Squaring the previous equation, then rearranging and cancelling terms yields $2(d^TM^TM\bar{y} - \bar{y}^Tgg^Td) + \beta(d^TM^TMd - d^Tgg^Td) = 0$, which is true only if $g^Td = 0 \Rightarrow Md = 0$. This in turn implies that $\mathrm{rank}(M) = 1$, violating Assumption 1. Therefore there exists y satisfying $\|My\| < g^Ty$. ∎

One characterization of $\mathcal{F}^*$ is as follows:

$$\mathcal{F}^* = \mathbf{cl}\ \left\{M^T\lambda + g\alpha : \|\lambda\| \le \alpha\right\}. \tag{2}$$

This result admits an elementary proof by a separating hyperplane argument and has been part of the folklore of convex analysis for several decades. For a standard proof, see, for example, Theorem 3.1 of Berman [3] applied to the second-order cone. The lack of closure of $\mathcal{T} := \{M^T\lambda + g\alpha : \|\lambda\| \le \alpha\}$ can arise easily. Let $M = \left[\begin{smallmatrix} -1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ and $g = \left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]$. In this case, $\mathcal{T} = \{(-\lambda_1 + \alpha, \lambda_2) \mid \|(\lambda_1, \lambda_2)\| \le \alpha\}$. It is easy to verify that $(0,1) \notin \mathcal{T}$ but $(\varepsilon, 1) \in \mathcal{T}$ for every $\varepsilon > 0$ (set $\lambda_1 = \frac{1}{2\varepsilon} - \frac{\varepsilon}{2}$, $\lambda_2 = 1$, and $\alpha = \frac{1}{2\varepsilon} + \frac{\varepsilon}{2}$), which shows that $\mathcal{T}$ is not closed. For an analysis of cases when $\mathcal{T}$ is guaranteed to be closed, see Pataki [5].

Proof of Theorem 1. Since $\mathbf{int}\ \mathcal{F} \neq \emptyset$, Proposition 1 implies that $D_n < 0$, and so, for the sake of this proof, we rescale (M, g) by $1/\sqrt{|D_n|}$ in order to conveniently satisfy $D_n = -1$. (i) Define $\mathcal{H} := \{y : y^TQDQ^Ty \le 0,\ y^TQ_n \ge 0\}$ and $\mathcal{L} := \{y : y^TQDQ^Ty \le 0,\ y^Tg \ge 0\}$. We need to prove that $\mathcal{H} = \mathcal{F}$. It is straightforward to check that $\mathcal{F} = \mathcal{L}$ and, indeed, $\mathbf{int}\ \mathcal{F} = \mathbf{int}\ \mathcal{L} = \{y : y^TQDQ^Ty < 0,\ y^Tg > 0\}$ from Proposition 1, and this also readily establishes that $Q_n \in \mathbf{int}\ \mathcal{F}$. For any y we can write

$$y^TQ_n = y^T\left(-QDQ^T\right)Q_n = y^T\left(gg^T - M^TM\right)Q_n = \left(My, g^Ty\right)^T\left(-MQ_n, g^TQ_n\right), \tag{3}$$

where the final term's two parenthetic vectors lie in $\mathbb{R}^{m+1}$. Notice that $(-MQ_n, g^TQ_n) \in \mathbf{int}\ \mathcal{Q}^{m+1}$, since $Q_n \in \mathbf{int}\ \mathcal{F}$. If $y \in \mathcal{F}$, then both vectors in the last term of (3) are in $\mathcal{Q}^{m+1}$; hence $y^TQ_n \ge 0$ follows from the self-duality of $\mathcal{Q}^{m+1}$. Therefore $y \in \mathcal{H}$, showing that $\mathcal{F} \subset \mathcal{H}$. Next suppose that $y \in \mathcal{H}$. Then $y \in \mathcal{F}$ unless $g^Ty < 0$, in which case $-y \in \mathcal{F}$. Using (3) we have

$$0 \le y^TQ_n = \left(My, g^Ty\right)^T\left(-MQ_n, g^TQ_n\right) = -\left(-My, -g^Ty\right)^T\left(-MQ_n, g^TQ_n\right) \le 0,$$

again because the two vectors in the last term lie in the self-dual cone $\mathcal{Q}^{m+1}$. This implies that equality holds throughout, and hence $(-My, -g^Ty) = (0,0)$ since $(-MQ_n, g^TQ_n) \in \mathbf{int}\ \mathcal{Q}^{m+1}$, yielding the contradiction that $g^Ty = 0$. This establishes that $\mathcal{F} \subset \mathcal{H}$, completing the proof of (i).

(ii) Having established (i), suppose that $D_i = 0$ for some $i \in \{1, \ldots, n-1\}$. Then $(\theta Q_i)^TQDQ^T(\theta Q_i) = 0$ and $Q_n^T(\theta Q_i) = 0$, whereby $\theta Q_i \in \mathcal{F}$ for all θ, violating the hypothesis that $\mathcal{F}$ is regular. Therefore $D_i > 0$ for all $i \in \{1, \ldots, n-1\}$, and hence D^{-1} exists. Define $\mathcal{J} := \{z : z^TQD^{-1}Q^Tz \le 0,\ z^TQ_n \ge 0\}$. Suppose that $z \in \mathcal{J}$

and $y \in \mathcal{F}$, in which case

$$y^T z = y^T Q Q^T z = \sum_{i=1}^{n-1} D_i^{\frac{1}{2}} \left(Q_i^T y\right) D_i^{-\frac{1}{2}} \left(Q_i^T z\right) + y^T Q_n z^T Q_n$$
$$\geq -\sqrt{\sum_{i=1}^{n-1} D_i \left(Q_i^T y\right)^2} \sqrt{\sum_{i=1}^{n-1} D_i^{-1} \left(Q_i^T z\right)^2} + y^T Q_n z^T Q_n \geq 0,$$

where the first inequality is an application of the Cauchy–Schwarz inequality and the second inequality follows, since $z \in \mathcal{J}$ and $y \in \mathcal{F}$ using part (i). Thus $z \in \mathcal{F}^*$, which shows that $\mathcal{J} \subset \mathcal{F}^*$. Next let $\bar{Q}$ denote the matrix of the first $n-1$ columns of Q, and let $\bar{D}$ denote the diagonal matrix composed of the $n-1$ diagonal components $D_1, \ldots, D_{n-1}$. Then from part (i) we have $\mathcal{F} = \{y : \sqrt{y^T \bar{Q} \bar{D} \bar{Q}^T y} \leq Q_n^T y\} = \{y : \|\bar{D}^{\frac{1}{2}} \bar{Q}^T y\| \leq Q_n^T y\}$, and using (2) we know that $\mathcal{F}^* = \mathbf{cl}\, \mathcal{T}$, where $\mathcal{T} = \{\bar{Q}\bar{D}^{\frac{1}{2}}\lambda + Q_n \alpha : \|\lambda\| \leq \alpha\}$. Let $z \in \mathcal{T}$, where $z = \bar{Q}\bar{D}^{\frac{1}{2}}\lambda + Q_n\alpha$ and $\|\lambda\| \leq \alpha$. Then

$$z^T Q D^{-1} Q^T z = \left(\bar{Q}\bar{D}^{\frac{1}{2}}\lambda + Q_n\alpha\right)^T Q D^{-1} Q^T \left(\bar{Q}\bar{D}^{\frac{1}{2}}\lambda + Q_n\alpha\right) = \lambda^T\lambda - \alpha^2 \leq 0,$$

and furthermore $Q_n^T z = \alpha \geq 0$, whereby $z \in \mathcal{J}$. Thus $\mathcal{T} \subset \mathcal{J}$. It then follows that $\mathcal{F}^* = \mathbf{cl}\, \mathcal{T} \subset \mathbf{cl}\, \mathcal{J} = \mathcal{J}$, which completes the proof of (ii).

To prove (iii), notice that $Q_n^T z = -\alpha D_n Q_n^T y \geq 0$ and

$$z^T Q D^{-1} Q^T z = \alpha^2 y^T Q D Q^T Q D^{-1} Q^T Q D Q^T y = \alpha^2 y^T Q D Q^T y \leq (=)\ 0,$$

since $y \in \mathcal{F}$ ($y \in \partial\mathcal{F}$) implies that $y^T Q D Q^T y \leq (=)\ 0$, and hence $z \in \mathcal{F}^*$ ($z \in \partial\mathcal{F}^*$) from part (ii). Furthermore $y^T z = -\alpha y^T Q D Q^T y = 0$ when $y \in \partial\mathcal{F}$, completing the proof of (iii). The proof of (iv) follows similar logic. ∎

Before proving Corollary 1 we first prove the following.

PROPOSITION 2. *Suppose that* $\mathbf{int}\ \mathcal{F} \neq \emptyset$ *and* $\mathrm{rank}(M) = n$. *Then* $g^T(M^TM)^{-1}g > 1$ *and* $\bar{y} := (M^TM)^{-1}g \in \mathbf{int}\ \mathcal{F}$.

Proof. Let $\alpha := g^T(M^TM)^{-1}g > 0$, since $g \neq 0$ from Assumption 1. From Proposition 1 we know there exists $\hat{y}$ satisfying $\|M\hat{y}\| < g^T\hat{y}$ and rescale $\hat{y}$ if necessary so that $g^T\hat{y} = \alpha$. Notice that $\bar{y}$ optimizes the function $f(y) = y^TM^TMy - 2g^Ty$, whose optimal objective function value is $-\alpha$. Therefore

$$-\alpha \leq \hat{y}^TM^TM\hat{y} - 2g^T\hat{y} < \alpha^2 - 2\alpha,$$

which implies that $\alpha^2 > \alpha > 0$, and hence $\alpha > 1$. Next observe that $\|M\bar{y}\| = \sqrt{\bar{y}^TM^TM\bar{y}} = \sqrt{\alpha} < \alpha = g^T\bar{y}$, whereby $\bar{y} \in \mathbf{int}\ \mathcal{F}$. ∎

Proof of Corollary 1. (i) is a restatement of the definition of $\mathcal{F}$, (iii) is a restatement of part (iii) of Theorem 1, and (iv) is a restatement of part (iv) of Theorem 1 using the Sherman–Morrison formula

$$QD^{-1}Q^T = \left(M^TM - gg^T\right)^{-1} = \left(M^TM\right)^{-1} - \frac{(M^TM)^{-1}gg^T\left(M^TM\right)^{-1}}{g^T\left(M^TM\right)^{-1}g - 1},$$

together with the fact from Proposition 2 that $g^T(M^TM)^{-1}g > 1$.

It remains to prove (ii). Let $\mathcal{K} := \{z \in \mathbb{R}^n : z^TQD^{-1}Q^Tz \leq 0\}$. Then from Theorem 1 we have $\mathcal{K} = \mathcal{F}^* \cup -\mathcal{F}^*$. Let $\bar{y} = (M^TM)^{-1}g$, and note that $\bar{y} \in \mathbf{int}\ \mathcal{F}$

from Proposition 2. Define $\mathcal{H} := \{z \in \mathbb{R}^n : \bar{y}^T z \geq 0\}$, and note that $\mathcal{H} \cap \mathcal{F}^* = \mathcal{F}^*$ and $\mathcal{H} \cap -\mathcal{F}^* = \{0\}$. Therefore $\mathcal{F}^* = \mathcal{K} \cap \mathcal{H} = \{z \in \mathbb{R}^n : z^T Q D^{-1} Q^T z \leq 0,\ g^T (M^T M)^{-1} z \geq 0\}$. Using the Sherman–Morrison formula we obtain

$$\mathcal{F}^* = \left\{ z^T \left((M^T M)^{-1} - \frac{\left(M^T M\right)^{-1} g g^T \left(M^T M\right)^{-1}}{g^T \left(M^T M\right)^{-1} g - 1} \right) z \leq 0,\ g^T \left(M^T M\right)^{-1} z \geq 0 \right\},$$

which after rearranging yields the expression in (ii). ∎

Remark 2 (the case when $\mathcal{F}$ is not regular). Let Z and N partition the set of indices according to zero and nonzero values of D_i. If $D_n = 0$, then one can show that $\mathcal{F}$ is a half-subspace in the subspace spanned by the Q_i for $i \in Z$. If $D_n > 0$, then $\mathcal{F} = \{0\}$. If $D_n < 0$, then $\mathcal{F}$ has an interior, and we can interpret $D_i^{-1} = \infty$ for $i \in Z$. Then Theorem 1 remains valid if we interpret "$z^T Q D^{-1} Q^T z \leq 0$" in (ii) as "$\sum_{i \in N} D_i (Q^T z)_i^2 \leq 0,\ (Q^T z)_i^2 = 0$ for $i \in Z$," and "$y := -\alpha Q D^{-1} Q^T z$" in (iv) as "$Q_i^T y := -\alpha D_i^{-1} Q_i^T z$ for $i \in N$ and $Q_i^T y$ is set *arbitrarily* for $i \in N$."

3. An algorithm for approximately solving (1).

3.1. Basic properties of (1) and the polar problem pair. Returning to (1) where x is the given vector, consider the following conditions in (y, z, θ):

$$\begin{array}{l} y - \theta z = x, \\ y \in \mathcal{F}, \\ z \in \mathcal{F}^*, \\ \|z\| \leq 1, \\ \theta \geq 0,\ \theta \|z\| = \theta. \end{array} \tag{4}$$

Examining (4), we see that x is decomposed into $x = y - \theta z$, where $y \in \mathcal{F}$ and $-\theta z \in -\mathcal{F}^*$ and (y, z) is feasible for the problems (1). Let G denote the duality gap for (1), namely, $G = \|y - x\| + x^T z$. We also consider the following pair of conic problems that are "polar" to (1):

$$\begin{array}{lll} \mathcal{P}^\circ:\ f^* := \min_v & \|v - x\| & \mathcal{D}^\circ:\ f^* := \max_w \quad -x^T w \\ \text{s.t.} & v \in -\mathcal{F}^*, & \text{s.t.} \quad \|w\| \leq 1 \\ & & \qquad w \in -\mathcal{F}, \end{array} \tag{5}$$

together with the following conditions in (v, w, ρ):

$$\begin{array}{l} v - \rho w = x, \\ v \in -\mathcal{F}^*, \\ w \in -\mathcal{F}, \\ \|w\| \leq 1, \\ \rho \geq 0,\ \rho \|w\| = \rho; \end{array} \tag{6}$$

here x is decomposed into $x = v - \rho w$, where now (v, w) is feasible for the problems (5), $-\rho w \in \mathcal{F}$, and $v \in -\mathcal{F}^*$. Let G° denote the duality gap for (5), namely, $G^\circ = \|v - x\| + x^T w$.

It is a straightforward exercise to show that conditions (4) together with the complementarity condition $y^T z = 0$ constitute necessary and sufficient optimality conditions for (1), and similarly, (6) together with $v^T w = 0$ are necessary and sufficient for optimality for (5). Furthermore, the solutions of (4) and (6) transform to one another:

$$\begin{array}{rcl} (y, z, \theta) & \rightarrow & (v, w, \rho) = (-\theta z, -y/\|y\|, \|y\|), \\ (v, w, \rho) & \rightarrow & (y, z, \theta) = (-\rho w, -v/\|v\|, \|v\|), \end{array}$$

with necessary modifications for the cases when $y = 0$ (set $w = 0$) and/or $v = 0$ (set $z = 0$).

PROPOSITION 3. *Suppose* (y, z, θ) *satisfy* (4) *and* (v, w, ρ) *satisfy* (6). *Then* (y, z) *and* (v, w) *are feasible for their respective problems with respective duality gaps:*

(i) $G = y^T z$;
(ii) $G^\circ = v^T w$.

Furthermore,

(iii) *if* (y, z) *is optimal for* (1), *then* $t^* = \theta$;
(iv) *if* (v, w) *is optimal for* (5), *then* $f^* = \rho$;
(v) $(t^*)^2 + (f^*)^2 = \|x\|^2$.

Proof. To prove (i), observe that $y^T z = z^T x + \theta\|z\|^2 = z^T x + \theta\|z\| = z^T x + \|y - x\| = G$, and a similar argument establishes (ii). To prove (iii), observe that $t^* = \|x - y\| = \|\theta z\| = \theta$ with similar arguments for (iv). To prove (v), notice that (y, z, θ) satisfy (4), and $y^T z = 0$ if and only if (y, z) is optimal for (1), in which case it is easy to verify that $(v, w, \rho) \leftarrow (-\theta z, -y/\|y\|, \|y\|)$ satisfy (6) and (v, w) is optimal for (5). Therefore $\|x\|^2 = (y - \theta z)^T(y - \theta z) = y^T y + \theta^2 = \rho^2 + \theta^2 = (f^*)^2 + (t^*)^2$. ∎

PROPOSITION 4. *If* $Q_n^T x \le 0$, *then* $t^* \ge \tau_{\mathcal{F}^*}\|x\|$.

Proof. We assume for the proof that $\|x\| = 1$, since t^*, f^* scale positively with $\|x\|$. If $f^* = 0$, the result follows trivially since $\tau_{\mathcal{F}^*} \le 1$, and $t^* = 1$ from Proposition 3. If $f^* > 0$, define $c = -\frac{t^*}{f^*}Q_n$, and note that $\|c\| = \frac{t^*}{f^*}$. By definition of the width, $B(c, \frac{t^*}{f^*}\tau_{\mathcal{F}^*}) \subset -\mathcal{F}^*$. Note that $\|x - c\| = \sqrt{x^T x + 2\frac{t^*}{f^*}Q_n^T x + \frac{t^{*2}}{f^{*2}}Q_n^T Q_n} \le \sqrt{1 + \frac{t^{*2}}{f^{*2}}} = \frac{1}{f^*}$. Therefore $\frac{1}{f^*\|x-c\|} \ge 1$.

Next observe that $c + \frac{\tau_{\mathcal{F}^*}\|c\|(x-c)}{\|x-c\|} \in -\mathcal{F}^*$, which is equivalent to $c + \frac{\tau_{\mathcal{F}^*}t^*(x-c)}{f^*\|x-c\|} \in -\mathcal{F}^*$. By the previous inequality, we have $c + \tau_{\mathcal{F}^*}t^*(x - c) \in -\mathcal{F}^*$. Thus we have

$$f^* \le \|c + \tau_{\mathcal{F}^*}t^*(x - c) - x\| = (1 - \tau_{\mathcal{F}^*}t^*)\|x - c\| \le (1 - \tau_{\mathcal{F}^*}t^*)\frac{1}{f^*}.$$

Therefore, $1 - t^{*2} = f^{*2} \le 1 - \tau_{\mathcal{F}^*}t^*$, which implies that $\tau_{\mathcal{F}^*} \le t^*$. ∎

PROPOSITION 5. *Given* x *satisfying* $\|x\| = 1$ *and* $Q_n^T x \le 0$, *suppose that* (v, w, ρ) *satisfies* (6), *with duality gap* $G^\circ \le \sigma\tau_{\mathcal{F}^*}/2$ *for* (5), *where* $\sigma \le 1$. *Consider the assignment* $(y, z, \theta) \leftarrow (-\rho w, -v/\|v\|, \|v\|)$ *(with the necessary modification that* $y = 0$ *if* $v = 0$*). Then* (y, z, θ) *satisfies* (4), *with duality gap* $G \le \sigma$ *for* (1).

Proof. Note that $y^T z = \frac{(w^T v)\rho}{\|v\|} \le \frac{\sigma\tau_{\mathcal{F}^*}\rho}{2\|v\|}$, and we have the following relations: (i) $w^T v \le \sigma\tau_{\mathcal{F}^*}/2 \le 1/2$, (ii) $\|v\| = \theta = \|y - x\| \ge t^* \ge \tau_{\mathcal{F}^*}$ from Proposition 4, and (iii) $\rho = \|v - x\| = v^T w - w^T x \le 1/2 + f^* \le 3/2$ from Proposition 3. Therefore $y^T z \le \frac{\tau_{\mathcal{F}^*}\sigma}{2}\frac{3}{2}\frac{1}{\tau_{\mathcal{F}^*}} \le \sigma$. ∎

3.2. The six cases. We assume here that the given x has unit norm, i.e., $\|x\| = 1$, and that we seek feasible solutions to (1) with a duality gap at most σ, where $\sigma \le 1$. Armed with Propositions 3, 4, and 5, we now show how to compute a feasible solution (y, z) of (1) with duality gap $G \le \sigma$. Our method is best understood with the help of Figure 1. We know from section 3.1 and the conditions (4) and/or (6) that we need to decompose x into the sum of a vector in $\mathcal{F}$ plus a vector in $-\mathcal{F}^*$ and that the central axes of $\mathcal{F}$ and $-\mathcal{F}$ are the rays corresponding to Q_n and $-Q_n$, respectively. Define the "dividing hyperplane" $L_{\mathcal{F}} := \{y : Q_n^T y = 0\}$ perpendicular to the central axes of $\mathcal{F}$ and $-\mathcal{F}$, and define $L_{\mathcal{F}}^+ := \{y \in \mathbb{R}^n : Q_n^T y \ge 0\}$ and $L_{\mathcal{F}}^- := -L_{\mathcal{F}}^+$. We divide $L_{\mathcal{F}}^+$ into three regions: region 1 corresponds to points in $\mathcal{F}$, region 2 corresponds to points in $L_{\mathcal{F}}^+$ "near" the dividing hyperplane (where our nearness criterion will be defined

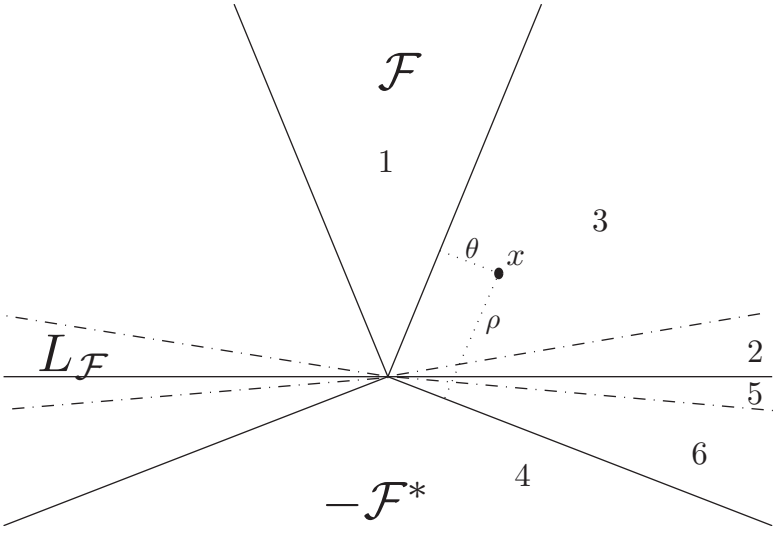


FIG. 1. *The geometry of the sets $\mathcal{F}$, $-\mathcal{F}^*$, and $L_{\mathcal{F}}$ and the six cases. The central axes of $\mathcal{F}$ and $-\mathcal{F}^*$ are the rays generated by $\pm Q_n$, respectively, which are orthogonal to the hyperplane $L_{\mathcal{F}}$. The regions corresponding to the six cases are shown as well.*

shortly), and region 3 corresponds to points in $L_{\mathcal{F}}^+ \setminus \mathcal{F}$ that are "far" from $L_{\mathcal{F}}$. We divide $L_{\mathcal{F}}^-$ similarly, into regions 4, 5, and 6. For each of the three regions in $L_{\mathcal{F}}^+$, we will work with the problem pair (1) and show how to compute a feasible solution (y, z) of (1) with duality gap $G \leq \sigma$. For each of the three regions in $L_{\mathcal{F}}^-$, we will instead work with the problem pair (5) and show how to compute a feasible solution (w, s) of (5) with duality gap $G^\circ \leq \sigma\tau_{\mathcal{F}^*}/2$, whereby from Proposition 5 we obtain a feasible solution (y, z) of (1) with duality gap $G \leq \sigma$. We will consider six cases, one for each of the regions described above and in Figure 1.

We first describe how we choose whether x is in region 2 or 3. For $x \in L_{\mathcal{F}}^+ \setminus \mathcal{F}$, define

$$\varepsilon_{\mathcal{P}} = \varepsilon_{\mathcal{P}}(x) := \frac{Q_n^T x \sqrt{|D_n|}}{\sqrt{\sum_{i=1}^{n-1} D_i \left(Q_i^T x\right)^2}}, \tag{7}$$

and notice that $x \in L_{\mathcal{F}}^+$ implies that $\varepsilon_{\mathcal{P}} \geq 0$, $x \notin \mathcal{F}$ implies that $\varepsilon_{\mathcal{P}} < 1$, and smaller values of $\varepsilon_{\mathcal{P}}$ correspond to $Q_n^T x$ closer to zero and hence x closer to $L_{\mathcal{F}}$. We specify a tolerance $\bar{\varepsilon}_{\mathcal{P}}$ and determine whether x is in region 2 or 3 depending on whether $\varepsilon_{\mathcal{P}} \leq \bar{\varepsilon}$ or $\varepsilon_{\mathcal{P}} > \bar{\varepsilon}$, respectively, where we set $\bar{\varepsilon} = \bar{\varepsilon}_{\mathcal{P}} := \sigma\tau_{\mathcal{F}}$.

Case 1: $Q_n^T x \geq 0$ *and* $x^T Q D Q^T x \leq 0$. From Theorem 1 we know that $x \in \mathcal{F}$. Then it is elementary to show that $(y, z, \theta) \leftarrow (x, 0, 0)$ satisfy (4), with $y^T z = 0$, whereby from Proposition 3 the duality gap is $G = 0$.

Case 2: $Q_n^T x \geq 0$ *and* $x^T Q D Q^T x > 0$, $\varepsilon_{\mathcal{P}} \leq \bar{\varepsilon}_{\mathcal{P}} := \sigma\tau_{\mathcal{F}}$. Let $\hat{y}$ solve the following system of equations:

$$\begin{aligned} [I + 1/|D_n| D] Q^T \hat{y} &= Q^T x - e_n Q_n^T x, \\ Q_n^T \hat{y} &= 0, \end{aligned} \tag{8}$$

where $e_n = (0, \ldots, 0, 1) \in \mathbb{R}^n$. Notice that the last row of the first equation system

has all zero entries. Therefore this system is not overdetermined, and one can write the closed-form solution $(Q^T\hat{y})_i = (Q^T x)_i/(1 + 1/|D_n|D_i)$ for $i = 1, \ldots, n-1$ and $(Q^T\hat{y})_n = 0$, in the transformed variables $\hat{s} := Q^T\hat{y}$. Having computed $\hat{y}$, next compute $\alpha := \sqrt{\hat{y}^T QDQ^T\hat{y}}/\sqrt{|D_n|}$, and then make the following assignments to variables:

$$\begin{aligned}
&\bar{y} \leftarrow \hat{y} + \alpha Q_n,\\
&\theta \leftarrow \sqrt{\bar{y}^T QD^2Q^T\bar{y}}/|D_n|,\\
&z \leftarrow -QDQ^T\bar{y}/(|D_n|\theta),\\
&y \leftarrow \bar{y} + Q_n^T x Q_n.
\end{aligned}$$

PROPOSITION 6. *Suppose that* $\|x\| = 1$, $\sigma \le 1$, *and* $\varepsilon_{\mathcal{P}} \le \bar{\varepsilon} < 1$ *and that* (y, z, θ) *are computed according to Case* 2 *above. Then* (y, z, θ) *is feasible for* (4) *with duality gap* $G \le \bar{\varepsilon}/\tau_{\mathcal{F}}$ *for* (1).

Applying Proposition 6 using $\bar{\varepsilon} = \bar{\varepsilon}_{\mathcal{P}} := \sigma\tau_{\mathcal{F}}$ ensures that the resulting duality gap satisfies $G \le \bar{\varepsilon}/\tau_{\mathcal{F}} = \sigma$. Note that the complexity of the computations in Case 2 is $O(mn^2)$ (assuming that square roots are sufficiently accurately computed in $O(1)$ operations).

Proof of Proposition 6. It is easy to establish that $(Q_1^T x, \ldots, Q_{n-1}^T x) \ne 0$, and hence $\alpha > 0$. This in turn implies that $Q_n^T\bar{y} = \alpha > 0$, and hence $\theta > 0$, so z is well defined. It is straightforward to verify that

$$\bar{y}^T QDQ^T\bar{y} = (\hat{y} + \alpha Q_n)^T QDQ^T(\hat{y} + \alpha Q_n) = \hat{y}^T QDQ^T\hat{y} - \alpha^2|D_n| = 0,$$

which shows via Theorem 1 that $\bar{y} \in \mathcal{F}$, and therefore $z \in \mathcal{F}^*$ and $z^T\bar{y} = 0$. It is also straightforward to verify that $\|z\| = 1$. Finally, we have from (8) that

$$\begin{aligned}
[I + 1/|D_n|D]Q^T\bar{y} &= [I + 1/|D_n|D]\left(Q^T\hat{y} + \alpha e_n\right)\\
&= [I + 1/|D_n|D]\left(Q^T\hat{y}\right) = Q^T\left(x - Q_n Q_n^T x\right)
\end{aligned}$$

(where the second equality above follows since the last row and column of the matrix are zero); hence $\bar{y} + 1/|D_n|QDQ^T\bar{y} = x - Q_n Q_n^T x$. Substituting the values of y, z, θ into this expression yields $y - \theta z = x$, which then shows that (y, z, θ) satisfy (4). Therefore from Proposition 3 (y, z) is feasible for (1) with duality gap

$$\begin{aligned}
G = z^T y = z^T\bar{y} + z^T Q_n Q_n^T x \le Q_n^T x &= \frac{\varepsilon_{\mathcal{P}}\sqrt{\sum_{i=1}^{n-1} D_i(Q_i^T x)^2}}{\sqrt{|D_n|}}\\
&\le \frac{\bar{\varepsilon}\sqrt{D_1}}{\sqrt{|D_n|}} \le \frac{\bar{\varepsilon}\sqrt{D_1 + |D_n|}}{\sqrt{|D_n|}} = \bar{\varepsilon}/\tau_{\mathcal{F}}. \qquad \square
\end{aligned}$$

Case 3*:* $Q_n^T x \ge 0$ *and* $x^T QDQ^T x > 0$, $\varepsilon_{\mathcal{P}} > \bar{\varepsilon}_{\mathcal{P}} := \sigma\tau_{\mathcal{F}}$. Here x is on the same side of the dividing hyperplane $L_{\mathcal{F}}$ as $\mathcal{F}$ but is neither in $\mathcal{F}$ nor close enough to $L_{\mathcal{F}}$ in the nearness measure. Consider the following univariate function in γ:

$$f(\gamma) := x^T Q[I + \gamma D]^{-1} D[I + \gamma D]^{-1} Q^T x = \sum_{i=1}^{n} \frac{D_i\left(x^T Q_i\right)^2}{(1 + D_i\gamma)^2}, \tag{9}$$

shown canonically in Figure 2.

Notice that $f(0) = x^T QDQ^T x > 0$, and since $D_n < 0$, we have $f(\gamma) \to -\infty$ as $\gamma \to 1/|D_n|$. Furthermore, $f'(\gamma) = -2\sum_{i=1}^{n} D_i^2(x^T Q_i)^2(1 + \gamma D_i)^{-3} < 0$ for $\gamma \in$

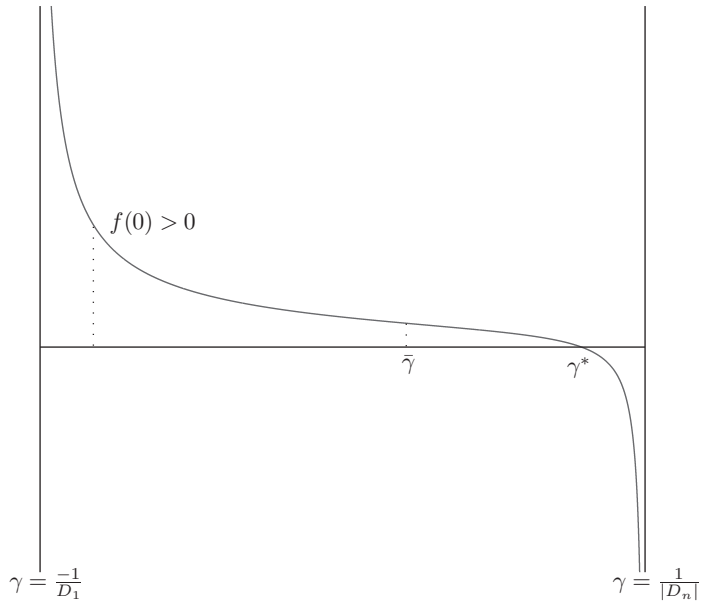


FIG. 2. *The function f on the interval $(-1/D_1, 1/|D_n|)$. Among many desirable properties, f is strictly decreasing and analytic and has a unique root $\gamma^* \in (0, 1/|D_n|)$. Moreover, f is convex over $(-1/D_1, \bar{\gamma})$ and concave over $(\bar{\gamma}, 1/|D_n|)$, where $\bar{\gamma}$ is the unique point satisfying $f''(\bar{\gamma}) = 0$. Note that one can have $\gamma^* \leq \bar{\gamma}$ or $\gamma^* \geq \bar{\gamma}$.*

$[0, 1/|D_n|)$. Therefore $f(\gamma)$ is strictly decreasing in the domain $[0, 1/|D_n|)$, whereby from the mean value theorem there is a unique value $\gamma^* \in (0, 1/|D_n|)$ for which $f(\gamma^*) = 0$. We show in section 5 how to combine a binary search and Newton's method to very efficiently compute $\gamma \in (0, 1/|D_n|)$ satisfying $f(\gamma) \leq 0$ and $f(\gamma) \approx 0$ (and $\gamma \approx \gamma^*$). Presuming that this can be done very efficiently, consider the following variable assignment:

$$\begin{array}{l} y \leftarrow Q\left[I + \gamma D\right]^{-1} Q^T x, \\ \theta \leftarrow \gamma\sqrt{y^T Q D^2 Q^T y}, \\ z \leftarrow -\gamma Q D Q^T y/\theta. \end{array} \tag{10}$$

We now show that (y, θ, z) satisfy (4). First note that $Q_n^T y = Q_n^T x/(1 - \gamma|D_n|) > 0$, and furthermore this shows that $\theta > 0$, and so z is well defined. By the hypothesis that $f(\gamma) \leq 0$ we have

$$y^T Q D Q^T y = x^T Q[I + \gamma D]^{-1} D[I + \gamma D]^{-1} Q^T x = f(\gamma) \leq 0,$$

which implies that $y \in \mathcal{F}$, and hence $z \in \mathcal{F}^*$ from Theorem 1. It is also straightforward to verify that $\|z\| = 1$. Finally, rearranging the formula for y yields $x = y + \gamma Q D Q^T y = y - \theta z$, which shows that (4) is satisfied. From Proposition 3, (y, z) is feasible for (1), and using the above assignments the duality gap works out to be

$$G = y^T z = -f(\gamma)/\sqrt{x^T Q D^2 [I + \gamma D]^{-2} Q^T x},$$

whereby G will be small if $f(\gamma) \approx 0$. To make this more precise requires a detailed analysis of a binary search and Newton's method, which is postponed to section 5 where we will prove the following.

PROPOSITION 7. *Suppose that* $\|x\| = 1$, $1 > \varepsilon_{\mathcal{P}} > \bar{\varepsilon}$, *and* $g > 0$ *is a given gap tolerance. If* $Q_n^T x > 0$ *and* $x^T Q D Q^T x > 0$, *then a solution* (y, z, θ) *of* (4) *with duality gap* $G \leq g$ *for* (1) *is computable in* $O(n \ln\ln(1/\tau_{\mathcal{F}} + 1/\bar{\varepsilon} + 1/g))$ *operations.*

Substituting $\bar{\varepsilon} = \bar{\varepsilon}_{\mathcal{P}} := \sigma\tau_{\mathcal{F}}$ and $g = \sigma$, it follows that the complexity of computing a feasible of solution of (y, z) of (1) with duality gap at most σ is $O(n \ln\ln(1/\tau_{\mathcal{F}} + 1/\sigma))$=$O(n \ln\ln(1/\min\{\tau_{\mathcal{F}}, \tau_{\mathcal{F}^*}\} + 1/\sigma))$ operations.

Case 4: $Q_n^T x \leq 0$ *and* $x^T Q D^{-1} Q^T x \leq 0$. From Theorem 1 we know that $x \in -\mathcal{F}^*$. Then it is elementary to show that $(y, z, \theta) \leftarrow (0, -x/\|x\|, \|x\|)$ satisfy (4), with $y^T z = 0$, whereby from Proposition 3 the duality gap is $G = 0$.

Before describing how we treat Cases 5 and 6 (corresponding to regions 5 and 6), we need to describe how we choose whether x is in region 5 or 6. We use a parallel concept to that used to distinguish regions 2 and 3, except that $\mathcal{F}$ is replaced by $-\mathcal{F}^*$; see Figure 1. For $x \in L_{\mathcal{F}}^- \setminus -\mathcal{F}^*$, define the following quantity analogous to (7):

$$\varepsilon_{\mathcal{P}^*} = \varepsilon_{\mathcal{P}^*}(x) := \frac{-Q_n^T x \sqrt{1/|D_n|}}{\sqrt{\sum_{i=1}^{n-1} (1/D_i) \left(Q_i^T x\right)^2}}, \tag{11}$$

and notice that $x \in L_{\mathcal{F}}^-$ implies $\varepsilon_{\mathcal{P}^*} \geq 0$, $x \notin -\mathcal{F}^*$ implies $\varepsilon_{\mathcal{P}^*} < 1$, and smaller values of $\varepsilon_{\mathcal{P}^*}$ correspond to $Q_n^T x$ closer to zero and hence x closer to $L_{\mathcal{F}}$. We specify a tolerance $\bar{\varepsilon}_{\mathcal{P}^*}$ and determine whether x is in region 5 or 6 depending on whether $\varepsilon_{\mathcal{P}^*} \leq \bar{\varepsilon}$ or $\varepsilon_{\mathcal{P}^*} > \bar{\varepsilon}$, respectively, where we set $\bar{\varepsilon} = \bar{\varepsilon}_{\mathcal{P}^*} := \sigma\tau_{\mathcal{F}^*}^2/2$.

Case 5: $Q_n^T x \leq 0$ *and* $x^T Q D^{-1} Q^T x > 0$, *and* $\varepsilon_{\mathcal{P}^*} \leq \bar{\varepsilon}_{\mathcal{P}^*} := \sigma\tau_{\mathcal{F}^*}^2/2$. This case is an exact analogue of Case 2, with $\mathcal{F}$ replaced by $-\mathcal{F}^*$ and the pair (1) replaced by (5). Therefore the methodology of Case 2 can be used to compute (v, w, ρ) satisfying (6), and hence (v, w) is feasible for (5). Applying Proposition 6 to the context of the polar pair (5) with $\bar{\varepsilon} = \bar{\varepsilon}_{\mathcal{P}^*}$, it follows that the duality gap for (5) will be $G^\circ = v^T w$ and will satisfy $G^\circ \leq \bar{\varepsilon}/\tau_{\mathcal{F}^*} = \sigma\tau_{\mathcal{F}^*}^2/(2\tau_{\mathcal{F}^*}) \leq \sigma\tau_{\mathcal{F}^*}/2$. Converting (v, w, ρ) to (y, z, θ) using Proposition 5, we obtain (y, z) feasible for (1) with duality gap $G \leq \sigma$. Here the complexity of the computations is of the same order as Case 2.

Case 6: $Q_n^T x \leq 0$ *and* $x^T Q D^{-1} Q^T x > 0$, *and* $\varepsilon_{\mathcal{P}^*} > \bar{\varepsilon}_{\mathcal{P}^*} := \sigma\tau_{\mathcal{F}^*}^2/2$. In concert with the previous case, this case is an exact analogue of Case 3, with $\mathcal{F}$ replaced by $-\mathcal{F}^*$ and the pair (1) replaced by (5). Therefore the methodology of Case 3 can be used to compute (v, w, ρ) satisfying (6), and hence (v, w) is feasible for (5). Applying Proposition 7 to the context of the polar pair (5) with $\bar{\varepsilon} = \bar{\varepsilon}_{\mathcal{P}^*}$ and $g = \sigma\tau_{\mathcal{F}^*}/2$, it follows that a solution (v, w, ρ) of (6) with duality gap $G^\circ \leq g = \sigma\tau_{\mathcal{F}^*}/2$ for (5) is computable in $O(n \ln\ln(1/\tau_{\mathcal{F}^*} + 1/\bar{\varepsilon} + 1/g)) = O(n \ln\ln(1/\min\{\tau_{\mathcal{F}}, \tau_{\mathcal{F}^*}\} + 1/\sigma))$ operations. Converting (v, w, ρ) to (y, z, θ) using Proposition 5, we obtain (y, z) feasible for (1) with duality gap $G \leq \sigma$.

Proof of Theorem 2. The spectral decomposition of $M^T M - gg^T = QDQ^T$ is assumed to take $O(mn^2)$ operations. The computations in Cases 1 and 4 are trivial after checking the conditions of the cases, which is $O(mn^2)$ operations, and similarly for Cases 2 and 5. Regarding Cases 3 and 6, the discussion in the description of these cases establishes the desired operation bound. ∎

Remark 3 (the case when $\mathcal{F}$ is not regular, again). As in Remark 2, let Z and N partition the set of indices according to zero and nonzero values of D_i. Consider the case when $D_n < 0$ (the cases when $D_n > 0$ and $D_n = 0$ were discussed in Remark 2). We interpret $D_i^{-1} = \infty$ for $i \in Z$. Consider the orthonormal transformation $Q^T x$ and $Q^T y$, $Q^T z$ of the given vector x and the variables y, z. Then for $i \in Z$ simply set $Q_i^T y = Q_i^T x$ and $Q_i^T z = 0$ and work in the lower-dimensional problem in the subspace spanned by Q_i, $i \in N$.

TABLE 1
Average computational results from 100 *randomly generated sparse problems.*

Dimension n	Iterations: Proposed method, Theoretical bound	Iterations: Proposed method, Actual	Iterations: SDPT3	Running time (seconds): Proposed method, EIG	Running time (seconds): Proposed method, Total	Running time (seconds): SDPT3	Range of widths: Minimum	Range of widths: Maximum
10	6.6	4.7	11.7	0.0005	0.0013	0.1089	1e-7	0.534015
20	7.2	4.8	13.9	0.0003	0.0011	0.1489	0.046400	0.510048
50	7.7	4.5	14.2	0.0011	0.0014	0.2828	0.087697	0.441248
100	8.0	4.3	18.8	0.0065	0.0075	1.1620	0.095081	0.414343
200	8.0	4.0	23.8	0.0556	0.0575	5.7393	0.176357	0.430163
500	8.0	3.8	18.8	1.0492	1.0571	33.9511	0.179974	0.317520

4. Comparison with interior-point methods. The primal-dual pair of problems (1) can be reformulated as second-order cone programs; one formulation of the primal problem is $\max_{y,t}\{-t : (My, g^T y) \in \mathcal{Q}^{m+1},\ (y - x, t) \in \mathcal{Q}^{n+1}\}$, for example. It then follows from interior-point complexity theory that approximate solutions to (1) with duality gap at most δ can be computed in $O(\ln(1/\delta))$ interior-point iterations. This iteration bound follows from Theorem 2.4.1 of Renegar [6], noting that an interior starting feasible solution with good symmetry is easy to precompute. Unlike the complexity bound for the proposed method in Theorem 2, the interior-point method bound has a stronger dependence on the duality gap δ (global linear convergence), but, unlike the proposed method, there is no dependence on the widths $\tau_{\mathcal{F}}, \tau_{\mathcal{F}^*}$. Therefore from a complexity viewpoint one cannot assert that one algorithm dominates the other. (There is a theory of local quadratic convergence for interior-point methods (see, for example, [1]) that could possibly be used to prove a weaker interior-point method dependence on δ for this class of problems.) In terms of computational practice, it is relevant to compare the two methods on randomly generated problems. We generated 100 relatively sparse random problem instances ((M, g) has density, respectively, 10% and 30% on average) for each of dimensions $(m, n) = (2n, n)$ for $n = 10$, 20, 50, 100, 200, and 500 and solved them using both our method and the conic convex interior-point method software SDPT3s [9]. All computation was performed in MatLab on a recent model laptop computer. We used MatLab's EIG command to compute the eigendecomposition for $M^T M - gg^T$ for our proposed method. Table 1 shows our computational results. Columns 2 and 3 of the table show the average theoretical iteration bound and the average actual iterations of our method. Columns 5, 6, and 7 report running time information. Note as expected that the eigendecomposition is the dominant computation in our method. Our method substantially outperforms SDPT3, as one would expect, since SDPT3 is not optimized for the problem class (1).

A fairer comparison can be done by presuming that the problem instances are preprocessed and transformed by the eigendecomposition, replacing $Q^T y$ with y, whereby the second-order cone formulation takes the diagonal form:

$$\max_{y,t} \left\{ -t : \left(\sqrt{D_1} y_1, \ldots, \sqrt{D_{n-1}} y_{n-1}, \sqrt{|D_n|} y_n \right) \in \mathcal{Q}^n,\ (y - x, t) \in \mathcal{Q}^{n+1} \right\},$$

and SDPT3 naturally exploits the sparse structure of this problem. We generated 100 random problem instances each, for a range of n from $n = 20$ to $n = 5000$, where each instance was generated by randomly choosing D but ensuring that $\tau_{\mathcal{F}} = \tau_{\mathcal{F}^*} = 10^{-7}$. These instances need no eigendecomposition for our method; also SDPT3 can take good advantage of the problem's natural sparsity as well. Table 2 shows our

TABLE 2

Average computational results from 100 *randomly generated diagonal problems with* $\tau_{\mathcal{F}} = \tau_{\mathcal{F}^*} = 10^{-7}$.

Dimension n	Iterations: Proposed method: Theoretical bound	Iterations: Proposed method: Actual	Iterations: SDPT3	Running time (seconds): Proposed method	Running time (seconds): SDPT3
10	8.0	5.0	23.9	0.0004	0.2320
20	8.0	5.0	26.9	0.0005	0.2778
50	8.0	5.0	28.1	0.0005	0.4869
100	8.0	5.0	18.3	0.0018	1.1063
200	8.0	5.0	17.6	0.0033	1.5806
500	8.0	4.9	20.2	0.0154	0.5290
1000	8.0	4.9	16.9	0.0567	1.0957
2000	8.0	5.0	16.4	0.2153	1.8111
5000	8.0	5.2	19.9	1.2656	9.3998

computational results. Our method still substantially outperforms SDPT3 but not as dramatically when n is very large. However, the running time numbers in Table 2 are the running time until the stopping criteria are met for each method. The stopping criteria for SDPT3 includes stopping when the duality gap is sufficiently small or when insufficient progress is made in satisfying primal/dual feasibility/optimality. For the diagonal problems generated, this latter stopping criteria is unfortunately encountered quite often: the relative error of the final solution from SDPT3 was at least 0.01 for 65% of the diagonal problem instances and was at least 0.001 for 81% of the instances. In fact, SDPT3 stopped with a relative error of at most 10^{-6} in only 2% of the instances. In contrast, our proposed method terminated with a relative error of 10^{-12} in all instances.

5. Proof of Proposition 7. This section is devoted to the proof of Proposition 7. Our algorithmic approach is motivated by Ye [10], and it consists of a combination of a binary search and Newton's method to approximately solve $f(\gamma) = 0$ for the function f given in (9). An alternate approach would be to use interpolation methods as presented and analyzed in Melman [4], for which global quadratic convergence is proved but there is no complexity analysis of associated constants. While Proposition 7 indicates that a solution (y, z, θ) of (4) with duality gap $G \leq g$ for (1) can be computed extremely efficiently, unfortunately our proof is not nearly as efficient as we or the reader might wish. We assume throughout this section that the hypotheses of Proposition 7 hold. We start with a review of Smale's main result for Newton's method in [8].

5.1. Newton's method and Smale's results. Let g be an analytic function, and consider the Newton iterate from a given point $\hat{\gamma}$:

$$\gamma^+ = \hat{\gamma} - \frac{g(\hat{\gamma})}{g'(\hat{\gamma})},$$

and let $\{\gamma_k\}_{k\geq 0}$ denote the sequence of points generated starting from $\hat{\gamma} = \gamma_0$.

DEFINITION 1. *A point* γ_0 *is said to be an approximate zero of* g *if*

$$|\gamma_k - \gamma_{k-1}| \leq (1/2)^{2^{k-1}-1}|\gamma_1 - \gamma_0| \quad \textit{for } k \geq 1.$$

For an approximate zero γ_0, let $\gamma^* = \lim_{k\to\infty} \gamma_k$. Then γ^* is a zero of g, and Newton's method starting from γ_0 converges quadratically to γ^* from the very first iteration. The main result in [8] can be restated as follows.

THEOREM 3 (Smale [8]). *Let g be an analytic function. If $\hat{\gamma}$ satisfies*

$$\sup_{k>1}\left|\frac{g^{(k)}(\hat{\gamma})}{k!g'(\hat{\gamma})}\right|^{1/(k-1)} \leq \frac{1}{8}\left|\frac{g'(\hat{\gamma})}{g(\hat{\gamma})}\right|, \tag{12}$$

then $\hat{\gamma}$ is an approximate zero of g. Furthermore, if $\hat{\gamma}$ is an approximate zero of g, then $|\gamma_k - \gamma^*| \leq 2(1/2)^{2^{k-1}}|\gamma_1 - \gamma_0|$ for all $k \geq 1$.

5.2. Properties of $f(\gamma)$. We employ the change of variables $s = Q^T x$, whereby from the hypotheses of Proposition 7 we have $s_n > 0$, $s^T Ds > 0$, and $\varepsilon_{\mathcal{P}} = s_n\sqrt{|D_n|}/\sqrt{\sum_{j=1}^{n-1} D_i s_i^2} > \bar{\varepsilon}$. We consider computing a zero of our function of interest:

$$f(\gamma) = s^T\left(I+\gamma D\right)^{-2} Ds = \sum_{i=1}^{n}\frac{D_i s_i^2}{(1+\gamma D_i)^2}. \tag{13}$$

LEMMA 1. *Under the hypotheses of Proposition 7, f has the following properties:*

(i) *$f(0) > 0$, $\lim_{\gamma\to 1/|D_n|} f(\gamma) = -\infty$, and f has a unique root $\gamma^* \in (0, 1/|D_n|)$.*

(ii) *f is analytic on $(-1/D_1, 1/|D_n|)$, and for $k \geq 1$ the kth derivative of f is*

$$\begin{aligned} f^{(k)}(\gamma) &= (-1)^k(k+1)!s^T\left(I+\gamma D\right)^{-(k+2)} D^{k+1}s \\ &= (-1)^k(k+1)!\sum_{i=1}^{n}\frac{D_i^{k+1}s_i^2}{(1+\gamma D_i)^{k+2}}. \end{aligned}$$

(iii) $\displaystyle\sup_{k>1}\left|\frac{f^{(k)}(\gamma)}{k!f'(\gamma)}\right|^{1/(k-1)} \leq \frac{3}{2}\max\left\{\frac{D_1}{1+\gamma D_1},\ \frac{|D_n|}{1-\gamma|D_n|}\right\}.$

(iv) $\displaystyle\frac{1-\varepsilon_{\mathcal{P}}}{|D_n|+\varepsilon_{\mathcal{P}}D_1} \leq \gamma^* \leq \frac{1-\varepsilon_{\mathcal{P}}}{|D_n|}$, *where $\varepsilon_{\mathcal{P}}$ is given by (7).*

(v) *There exists a unique value $\bar{\gamma} \in (-1/D_1, 1/|D_n|)$ such that f is convex on $(-1/D_1, \bar{\gamma}]$ and concave on $[\bar{\gamma}, 1/|D_n|)$.*

Proof. (i) follows from the mean value theorem and the observation that f is decreasing on $(0, 1/|D_1|)$, and (ii) follows using a standard derivation. To prove (iii) observe that

$$\begin{aligned} \left|\frac{f^{(k)}(\gamma)}{k!f'(\gamma)}\right|^{1/(k-1)} &= \left|\frac{(k+1)!}{2k!}\right|^{1/(k-1)}\left|\frac{s^T\left(I+\gamma D\right)^{-(k+2)}D^{k+1}s}{s^T\left(I+\gamma D\right)^{-3}D^2 s}\right|^{1/(k-1)} \\ &\leq \frac{3}{2}\left|\frac{s^T\left(I+\gamma D\right)^{-3/2}D\left[\left(I+\gamma D\right)^{-1}D\right]^{k-1}D\left(I+\gamma D\right)^{-3/2}s}{s^T\left(I+\gamma D\right)^{-3/2}D^2\left(I+\gamma D\right)^{-3/2}s}\right|^{1/(k-1)} \\ &\leq \frac{3}{2}\max_{v\neq 0}\left|\frac{v^T P^{k-1}v}{v^T v}\right|^{1/(k-1)} \\ &= \frac{3}{2}\max_{i=1,\ldots,n}\left\{\frac{|D_i|}{1+\gamma D_i}\right\}, \end{aligned}$$

where $P = \left(I+\gamma D\right)^{-1}D$. Therefore

$$\left|\frac{f^{(k)}(\gamma)}{k!f'(\gamma)}\right|^{1/(k-1)} \leq \frac{3}{2}\max_{i=1,\ldots,n}\left\{\frac{|D_i|}{1+\gamma D_i}\right\} \leq \frac{3}{2}\max\left\{\frac{D_1}{1+\gamma D_1},\ \frac{|D_n|}{1-\gamma|D_n|}\right\},$$

which proves (iii). To prove the first inequality of (iv), note that

$$f(\gamma) = \sum_{i=1}^{n} \frac{D_i s_i^2}{(1+\gamma D_i)^2} \geq \frac{1}{(1+\gamma D_1)^2} \sum_{i=1}^{n-1} D_i s_i^2 - \frac{|D_n| s_n^2}{(1+\gamma D_n)^2}.$$

The right-hand side of the expression above equals zero only at $\tilde{\gamma} := \frac{1-\varepsilon_{\mathcal{P}}}{|D_n|+\varepsilon_{\mathcal{P}} D_1}$. This implies that $f(\tilde{\gamma}) \geq 0$, whereby $\tilde{\gamma} \leq \gamma^*$, since f is strictly decreasing. For the second inequality note that $\varepsilon_{\mathcal{P}} \in (0,1)$ since $s_n > 0$ and $s^T D s > 0$. We have $f(\gamma) < \sum_{i=1}^{n-1} s_i^2 D_i \ - |D_n| s_n^2/(1+\gamma D_n)^2$, and substituting $\gamma = \frac{1-\varepsilon_{\mathcal{P}}}{|D_n|}$ into this strict inequality yields $f(\frac{1-\varepsilon_{\mathcal{P}}}{|D_n|}) < 0$, which then implies that $\gamma^* < \frac{1-\varepsilon_{\mathcal{P}}}{|D_n|}$. To prove (v), examine the derivatives of f in (ii), and notice that $f^{(k)}(\gamma) < 0$ for any odd value of k, whereby f'' is strictly decreasing. Let $\bar{\gamma}$ be the unique point in $(-1/D_1, 1/|D_n|)$ such that $f''(\bar{\gamma}) = 0$. Since f'' is strictly decreasing, f is convex on $(-1/D_1, \bar{\gamma})$ and concave on $(\bar{\gamma}, 1/|D_n|)$. □

Figure 2 illustrates the geometry underlying some of the analytical properties of f described by Lemma 1.

Remark 4. In the interval $(\frac{-1}{D_1}, \frac{1}{2|D_n|} - \frac{1}{2D_1}]$ the maximum in (iii) of Lemma 1 is $\frac{D_1}{1+\gamma D_1}$, and in the interval $[\frac{1}{2|D_n|} - \frac{1}{2D_1}, \frac{1}{|D_n|})$ the maximum is $\frac{|D_n|}{1+\gamma D_n}$.

5.3. Locating an approximate zero of f by binary search. From Lemma 1 we know that $\gamma^* \in (0, \bar{U}]$, where $\bar{U} := (1-\bar{\varepsilon})/|D_n|$. We will cover this interval with subintervals and use a binary search to locate an approximate zero of f, motivated by the method of Ye [10]. Noticing from Remark 4 that the maximum in (iii) of Lemma 1 depends on the "midpoint" $M := \frac{1}{2|D_n|} - \frac{1}{2D_1}$, we will consider two types of subintervals: the *left intervals* will cover $[0, \max\{0, M\}]$, and the *right intervals* will cover $[\max\{0, M\}, \bar{U}]$. (Of course, in the case when $M \leq 0$, there is no need to create the left intervals.)

The left intervals will be of the form $[L^{i-1}, L^i]$, where $L^i := \frac{1}{D_1}((\frac{13}{12})^i - 1)$ for $i = 0, 1, \ldots$. If $M \leq 0$, we do not consider creating these intervals. The right intervals will have the form $[R^i, R^{i-1}]$, where $R^i := \frac{1}{|D_n|} - (\frac{1}{|D_n|} - \bar{U})(\frac{13}{12})^i$ for $i = 0, 1, \ldots$.

Let $[a, b]$ denote one of these intervals (either $[L^{i-1}, L^i]$ or $[R^i, R^{i-1}]$ for some i). Note that if $f(a) \geq 0$ and $f(b) \leq 0$, then $\gamma^* \in [a, b]$. Supposing that this is the case, it follows from Lemma 1 that f is either convex on $[a, \gamma^*]$ or concave on $[\gamma^*, b]$ (or both), and consider starting Newton's method from $\hat{\gamma} = a$ in the first case or $\hat{\gamma} = b$ in the second case. Then the Newton step

$$\gamma^+ = \hat{\gamma} - \frac{f(\hat{\gamma})}{f'(\hat{\gamma})}$$

satisfies

$$\left| \frac{f(\hat{\gamma})}{f'(\hat{\gamma})} \right| = |\gamma^+ - \hat{\gamma}| \leq |\gamma^* - \hat{\gamma}| \leq b - a, \tag{14}$$

where the first inequality follows from either the convexity of f on $[a, \gamma^*]$ or the concavity of f on $[\gamma^*, b]$. In particular, we have

$$|f(\hat{\gamma})| \leq |f'(\hat{\gamma})||\gamma^* - \hat{\gamma}|, \tag{15}$$

which relates the value of the function at an approximate solution and the error in our approximation.

LEMMA 2. *Under the hypotheses of Proposition 7 the intervals described herein have the following properties:*

(i) *The total number of left intervals and right intervals needed to cover* $[0, \bar{U}]$ *is* $K_L := \lceil \frac{\ln(1/2)+2\ln(1/\tau_{\mathcal{F}})}{\ln(13/12)} \rceil^+$ *and* $K_R := \lceil \frac{\ln(1/\bar{\varepsilon})}{\ln(13/12)} \rceil$, *respectively.*

(ii) *Let* $[a, b]$ *denote one of these intervals, and suppose that* $f(a) \geq 0$ *and* $f(b) \leq 0$. *Then either* a *or* b *is an approximate zero of* f, *and* $\gamma^* \in [a, b]$.

(iii) $R^{i-1} - R^i \leq \frac{1}{12|D_n|}$ *for* $i = 1, \ldots, K_R$ *and* $L^i - L^{i-1} \leq \frac{1}{12|D_n|}$ *for* $i = 1, \ldots, K_L$.

Proof. We first prove (i) for the right intervals. We have $R^0 = \bar{U}$ and

$$R^{K_R} = \frac{1}{|D_n|} - \frac{\bar{\varepsilon}}{|D_n|}\left(\frac{13}{12}\right)^{K_R} \leq \frac{1}{|D_n|} - \frac{\bar{\varepsilon}}{|D_n|}\frac{1}{\bar{\varepsilon}}\min\left\{1, \frac{|D_n|}{2D_1} + \frac{1}{2}\right\}$$
$$= \max\left\{0, \frac{1}{2|D_n|} - \frac{1}{2D_1}\right\} = \max\{0, M\},$$

and thus the right intervals cover $[\max\{0, M\}, \bar{U}]$. Note that, using the above reasoning, one easily shows that, because $K_R \leq 1 + \ln(1/\bar{\varepsilon})/\ln(13/12)$, one also has

$$\left(\frac{13}{12}\right)^{K_R} \leq \frac{13}{12\bar{\varepsilon}}. \tag{16}$$

For the left intervals, first consider the case when $M \geq 0$. Then $|D_n| \leq D_1$ and $\tau_{\mathcal{F}} \leq \frac{1}{\sqrt{2}}$, whereby there is no need to take the nonnegative part in the definition of K_L. We have $L^0 = 0$ and

$$L^{K_L} = \frac{1}{D_1}\left(\left(\frac{13}{12}\right)^{K_L} - 1\right) \geq \frac{1}{D_1}\left(\frac{1}{2\tau_{\mathcal{F}}^2} - 1\right) = \frac{1}{D_1}\left(\frac{D_1 + |D_n|}{2|D_n|} - 1\right) = M,$$

and thus the left intervals cover $[0, M] = [0, \max\{0, M\}]$. Note that, using the above reasoning, one easily shows that, because $K_L \leq 1 + \frac{\ln(1/2)+2\ln(1/\tau_{\mathcal{F}})}{\ln(13/12)}$, one also has

$$\left(\frac{13}{12}\right)^{K_L} \leq \frac{13}{24\tau_{\mathcal{F}}^2}. \tag{17}$$

When $M \leq 0$ there is nothing to prove.

To prove (ii), we consider the two cases of $[a, b]$ being either a left or right interval. If $[a, b]$ is a left interval, then $M \geq 0$ and $b = a(13/12) + \frac{1}{12D_1}$. In this case, for one of $\hat{\gamma} = a$ or $\hat{\gamma} = b$, we have for all $k > 1$:

$$\frac{1}{8}\left|\frac{f'(\hat{\gamma})}{f(\hat{\gamma})}\right| \geq \frac{1/8}{b-a} = \frac{1/8}{(1/12)(a + 1/D_1)} \geq \frac{3}{2}\frac{D_1}{1 + \hat{\gamma}D_1} \geq \left|\frac{f^{(k)}(\hat{\gamma})}{k!f'(\hat{\gamma})}\right|^{1/(k-1)},$$

where the first inequality uses (14), the second inequality uses $a \leq \hat{\gamma}$, and the third inequality uses Remark 4 and the fact that $\hat{\gamma} \leq M$ in conjunction with Lemma 1. Therefore $\hat{\gamma}$ is an approximate zero of f. If $[a, b]$ is a right interval, then $a = b(13/12) - \frac{1}{12|D_n|}$ and $M \leq a \leq b$. In this case, for one of $\hat{\gamma} = a$ or $\hat{\gamma} = b$, we have for all $k > 1$:

$$\frac{1}{8}\left|\frac{f'(\hat{\gamma})}{f(\hat{\gamma})}\right| \geq \frac{1/8}{b-a} = \frac{1/8}{b - b(13/12) + \frac{1}{12|D_n|}} = \frac{1/8}{\frac{1}{12}\left(\frac{1}{|D_n|} - b\right)} = \frac{3}{2}\,\frac{|D_n|}{1 - b|D_n|}$$
$$\geq \frac{3}{2}\,\frac{|D_n|}{1 - \hat{\gamma}|D_n|} \geq \left|\frac{f^{(k)}(\hat{\gamma})}{k!f'(\hat{\gamma})}\right|^{1/(k-1)},$$

where the first inequality uses (14), the second inequality uses $M \le a \le \hat{\gamma} \le b$, and the third inequality uses Remark 4 and the fact that $\hat{\gamma} \ge M$ in conjunction with Lemma 1. Therefore $\hat{\gamma}$ is an approximate zero of f.

To prove (iii), for the right intervals

$$R^{i-1} - R^i = \frac{\bar{\varepsilon}}{13|D_n|}\left(\frac{13}{12}\right)^i \le \frac{\bar{\varepsilon}}{13|D_n|}\left(\frac{13}{12}\right)^{K_R} \le \frac{13}{12}\frac{1}{13|D_n|} = \frac{1}{12|D_n|},$$

by the definition of K_R, and the second inequality derives from (16).

For the left intervals, we can assume $M \ge 0$ (otherwise they are not constructed), in which case $D_1 \ge |D_n|$. In this case, we have

$$L^i - L^{i-1} = \frac{1}{13D_1}\left(\frac{13}{12}\right)^i \le \frac{1}{13D_1}\left(\frac{13}{12}\right)^{K_L} \le \frac{1}{13D_1}\frac{13}{24\tau_{\mathcal{F}}^2} = \frac{1}{24}\left(\frac{1}{D_1} + \frac{1}{|D_n|}\right) \le \frac{1}{12|D_n|},$$

by the definition of K_L, and the second inequality derives from (17). ∎

Based on these properties, consider the following method for locating an approximate zero of f. Perform a binary search on the end points of the intervals, testing the end points to locate an interval $[a, b]$ for which $f(a) \ge 0$ and $f(b) \le 0$. Then either a or b is an approximate zero of f. Then initiate Newton's method from *both* a and b either in parallel or iterate-sequentially. Notice that, in order to perform a binary search on the left and right intervals, there is no need to compute and evaluate f for all of the end points. In fact, the operation complexity of a binary search will be $O(n \ln K_L)$ and $O(n \ln K_R)$, respectively, since each function evaluation of f requires $O(n)$ operations.

5.4. Computing a solution of (1) with duality gap at most σ. Under the hypotheses of Proposition 7, suppose that $[a, b]$ is one of the constructed intervals, $f(a) \ge 0$, and $f(b) \le 0$. Then, from Lemmas 1 and 2, $\gamma^* \in [a, b]$ and either f is convex on $[a, \gamma^*]$ or concave on $[\gamma^*, b]$ (or both). We first analyze the latter case, i.e., when f is concave on $[\gamma^*, b]$, whereby b is an approximate zero of f, and we analyze the iterates of Newton's method for k iterations starting at $\gamma_0 = b$. Let $\gamma := \gamma_k$ be the final iterate. It follows from the concavity of f on $[\gamma^*, b]$ that $\gamma \ge \gamma^*$ and consequently $f(\gamma) \le 0$. Then the analysis in Case 3 shows that the assignment (10) yields a feasible solution of (1) with duality gap $G = -f(\gamma)/\sqrt{s^T D^2 [I + \gamma D]^{-2} s}$. The following result bounds the value of this duality gap.

LEMMA 3. *Let $g \in (0, 1]$ be the desired duality gap for* (1)*, and let*

$$k = 1 + \left\lceil \frac{\ln\ln\left(\left(\frac{1}{3g}\right)\left(\frac{1}{\tau_{\mathcal{F}}^2} + \frac{1}{\bar{\varepsilon}^2}\right)\right) - \ln\ln 2}{\ln 2} \right\rceil.$$

Under the hypotheses of Proposition 7 *and the setup above where b is an approximate zero of f, let $\gamma_0 := b$ and $\gamma_1, \ldots, \gamma_k$ be the Newton iterates, and define $\gamma := \gamma_k$. Then the assignment* (10) *will be feasible for* (1) *with duality gap at most g.*

Proof. We have $|f(\gamma)| \le |f'(\gamma)|\,|\gamma^* - \gamma|$ from the concavity of f on $[\gamma^*, b]$. Also, we have

$$|f'(\gamma)| = 2\sum_{i=1}^{n} \frac{D_i^2 s_i^2}{(1 + \gamma D_i)^3} \le 2\sum_{i=1}^{n-1} \frac{D_i^2 s_i^2}{(1 + \gamma D_i)^2} + 2\frac{D_n^2 s_n^2}{(1 + \gamma D_n)^2}\frac{1}{(1 + \gamma D_n)}.$$

Substitute $\frac{1}{1+\gamma D_n} = 1 + \frac{-\gamma D_n}{1+\gamma D_n}$ to obtain

$$|f'(\gamma)| \le 2\sum_{i=1}^{n} \frac{D_i^2 s_i^2}{(1+\gamma D_i)^2} - 2\frac{\gamma D_n^3 s_n^2}{(1+\gamma D_n)^3}.$$

Let $G = y^T z$ denote the duality gap. Then

$$\begin{aligned}
G &= \frac{-f(\gamma)}{\sqrt{s^T D^2 [I+\gamma D]^{-2} s}} \le \frac{|f'(\gamma)|\ |\gamma^* - \gamma|}{\sqrt{s^T D^2 [I+\gamma D]^{-2} s}} \\
&\le \frac{2\sum_{i=1}^{n} \frac{D_i^2 s_i^2}{(1+\gamma D_i)^2} + 2\frac{\gamma |D_n|^3 s_n^2}{(1+\gamma D_n)^3}}{\sqrt{s^T D^2 [I+\gamma D]^{-2} s}} |\gamma^* - \gamma| \\
&= \left(2\sqrt{s^T D^2 [I+\gamma D]^{-2} s} + 2\frac{\gamma |D_n|^3 s_n^2}{(1+\gamma D_n)^3 \sqrt{s^T D^2 [I+\gamma D]^{-2} s}} \right) |\gamma^* - \gamma| \\
&\le \left(2D_1 + 2\frac{|D_n|}{1+\gamma D_n} + 2\frac{\gamma D_n^2 s_n}{(1+\gamma D_n)^2} \right) |\gamma^* - \gamma|,
\end{aligned}$$

where we used $\sqrt{s^T D^2 [I+\gamma D]^{-2} s} \ge |D_n| s_n / (1+\gamma D_n)$ in the last inequality. Next note that $\gamma \le \bar{U} = \frac{1-\bar{\varepsilon}}{|D_n|}$, which implies that $\frac{1}{\bar{\varepsilon}} \ge \frac{1}{1+\gamma D_n}$. Therefore, recalling that γ is the kth iterate, we have

$$\begin{aligned}
G &\le 2|\gamma^* - \gamma| \left(D_1 + \frac{|D_n|}{\bar{\varepsilon}} + \frac{(1-\bar{\varepsilon}) D_n^2}{|D_n| \bar{\varepsilon}^2} \right) \\
&\le 2|\gamma^* - \gamma| |D_n| \left(\frac{1}{\tau_{\mathcal{F}}^2} + \frac{1}{\bar{\varepsilon}^2} \right) \\
&\le 4|\gamma_1 - \gamma_0| |D_n| \left(\frac{1}{\tau_{\mathcal{F}}^2} + \frac{1}{\bar{\varepsilon}^2} \right) \left(\frac{1}{2} \right)^{2^{k-1}} \\
&\le 4\frac{1}{12|D_n|} |D_n| \left(\frac{1}{\tau_{\mathcal{F}}^2} + \frac{1}{\bar{\varepsilon}^2} \right) \left(\frac{1}{2} \right)^{2^{k-1}} \\
&= \frac{1}{3} \left(\frac{1}{\tau_{\mathcal{F}}^2} + \frac{1}{\bar{\varepsilon}^2} \right) \left(\frac{1}{2} \right)^{2^{k-1}},
\end{aligned}$$

where we used Theorem 3 for the third inequality and Lemma 2 for the fourth inequality. Substituting the value of k above yields $G \le g$. ∎

Last of all, we analyze the case when f is convex on $[a, \gamma^*]$, whereby a is an approximate zero of f, and we analyze the iterates of Newton's method for k iterations starting at $\gamma_0 = a$. Let γ_k be the final iterate. It follows from the convexity of f on $[a, \gamma^*]$ that $\gamma_k \le \gamma^*$ and consequently $f(\gamma_k) \ge 0$, in which case the assignment (10) is not necessarily feasible for (1). However, invoking Theorem 3, we know that $\gamma_k + 2(1/2)^{2^{k-1}} |\gamma_1 - \gamma_0| \ge \gamma^*$, we also know that $\bar{U} \ge \gamma^*$, and we can set $\gamma := \min\{\gamma_k + 2(1/2)^{2^{k-1}} |\gamma_1 - \gamma_0|, \bar{U}\}$. Then the analysis in Case 3 shows that the assignment (10) yields a feasible solution of (1), with duality gap $G = -f(\gamma)/\sqrt{s^T D^2 [I+\gamma D]^{-2} s}$. The following result bounds the value of this duality gap.

LEMMA 4. *Let $g \in (0,1]$ be the desired duality gap for* (1)*, and let*

$$k = 1 + \left\lceil \frac{\ln\ln\left(\left(\frac{16}{3g} \right) \left(\frac{1}{\tau_{\mathcal{F}}^2} + \frac{1}{\bar{\varepsilon}^2} \right) \right) - \ln\ln 2}{\ln 2} \right\rceil.$$

Under the hypotheses of Proposition 7 and the setup above where a is an approximate zero of f, let $\gamma_0 := a$ and $\gamma_1, \ldots, \gamma_k$ be the Newton iterates, and define $\gamma := \min\{\gamma_k + 2(1/2)^{2^{k-1}}|\gamma_1 - \gamma_0|, \bar{U}\}$. Then the assignment (10) will be feasible for (1) with duality gap at most g.

Proof. Define $\delta := \gamma - \gamma_k$, and it follows that $\delta \geq 0$ and $\gamma_k + \delta \leq \bar{U}$. Furthermore,

$$\begin{aligned}\delta &\leq 2(1/2)^{2^{k-1}}|\gamma_1 - \gamma_0| \\ &\leq \frac{2}{\left(\frac{16}{3g}\right)[1/\tau_{\mathcal{F}}^2 + 1/\bar{\varepsilon}^2]\,12|D_n|} \\ &\leq \frac{\min\{\bar{\varepsilon}^2, \tau_{\mathcal{F}}^2\}}{|D_n|} \leq \frac{\min\{\bar{\varepsilon}, \tau_{\mathcal{F}}^2/(1-\tau_{\mathcal{F}}^2)\}}{|D_n|} = \min\{\bar{\varepsilon}/|D_n|, 1/D_1\}.\end{aligned} \tag{18}$$

Therefore $\delta \leq \bar{\varepsilon}/|D_n|$, whereby $1 + \gamma_k D_n + 2\delta D_n = 1 - (\gamma_k + \delta)|D_n| - \delta|D_n| \geq 1 + \bar{\varepsilon} - 1 - \bar{\varepsilon} = 0$, where we also used $\gamma_k + \delta \leq \bar{U} = (1 - \bar{\varepsilon})/|D_n|$. Therefore

$$1 + \gamma_k D_n \leq 2(1 + (\gamma_k + \delta)D_n) \leq 2(1 + tD_n) \quad \text{for all} \quad t \in [\gamma_k, \gamma_k + \delta]. \tag{19}$$

We also have from (18) that $\delta \leq 1/D_1 \leq 1/D_i \leq 1/D_i + \gamma_k$ for $i = 1, \ldots, n-1$; hence

$$1 + \gamma_k D_i + \delta D_i \leq 2(1 + \gamma_k D_i), \ i = 1, \ldots, n-1. \tag{20}$$

The duality gap of the assignment (10) is

$$G = y^T z = \frac{-f(\gamma)}{\sqrt{s^T D^2 [I + \gamma D]^{-2} s}} = \frac{-f(\gamma_k + \delta)}{\sqrt{s^T D^2 [I + (\gamma_k + \delta) D]^{-2} s}}.$$

We now proceed to bound the numerator and denominator of the rightmost expression. For the numerator we have

$$-f(\gamma_k + \delta) = |f(\gamma_k + \delta)| = \left| f(\gamma_k) + \int_{\gamma_k}^{\gamma_k + \delta} f'(t)dt \right|.$$

However, observe that $f(\gamma_k) \geq 0$, $f(\gamma_k + \delta) \leq 0$, and $f'(t) \leq 0$ for all $t \in [0, 1/|D_n|)$, whereby

$$|f(\gamma_k + \delta)| \leq \int_{\gamma_k}^{\gamma_k + \delta} |f'(t)|dt.$$

Using (19) for $t \in [\gamma_k, \gamma_k + \delta]$, we have

$$\begin{aligned}|f'(t)| &= 2\sum_{i=1}^{n-1} \frac{D_i^2 s_i^2}{(1 + tD_i)^3} + 2\frac{D_n^2 s_n^2}{(1 + tD_n)^3} \\ &\leq 2\sum_{i=1}^{n-1} \frac{D_i^2 s_i^2}{(1 + \gamma_k D_i)^3} + 16\frac{D_n^2 s_n^2}{(1 + \gamma_k D_n)^3} \leq 8|f'(\gamma_k)|,\end{aligned}$$

and it follows that $-f(\gamma_k + \delta) \leq 8\delta|f'(\gamma_k)|$. To bound the denominator, simply notice from (20) and $1 + \gamma_k D_n + \delta D_n \leq 1 + \gamma_k D_n$ that $\sqrt{s^T D^2 [I + (\gamma_k + \delta) D]^{-2} s} \geq (1/2)\sqrt{s^T D^2 [I + \gamma_k D]^{-2} s}$. Therefore

$$G = \frac{-f(\gamma_k + \delta)}{\sqrt{s^T D^2 [I + (\gamma_k + \delta) D]^{-2} s}} \leq 16\frac{\delta|f'(\gamma_k)|}{\sqrt{s^T D^2 [I + \gamma_k D]^{-2} s}}.$$

Next notice from the logic from the proof of Lemma 3 that

$$\frac{|f'(\gamma_k)|}{\sqrt{s^T D^2 [I+\gamma_k D]^{-2} s}} \leq 2|D_n| \left(\frac{1}{\tau_{\mathcal{F}}^2} + \frac{1}{\bar{\varepsilon}^2}\right);$$

therefore

$$G \leq 32\delta|D_n| \left(\frac{1}{\tau_{\mathcal{F}}^2} + \frac{1}{\bar{\varepsilon}^2}\right) \leq 32|D_n| \left(\frac{1}{\tau_{\mathcal{F}}^2} + \frac{1}{\bar{\varepsilon}^2}\right) \frac{2}{\left(\frac{16}{3g}\right)[1/\tau_{\mathcal{F}}^2 + 1/\bar{\varepsilon}^2]\, 12|D_n|} = g,$$

where the last inequality uses the second inequality of (18). ∎

Proof of Proposition 7. Note from the discussion at the end of section 5.3 that the operation complexity of the binary search is $O(n \ln K_L + n \ln K_R) = O(n \ln \ln(1/\tau_{\mathcal{F}} + 1/\bar{\varepsilon}))$ from Lemma 2. The number of Newton steps is $O(\ln \ln(1/\tau_{\mathcal{F}} + 1/\bar{\varepsilon} + 1/g))$ from Lemmas 3 and 4, with each Newton step requiring $O(n)$ operations, yielding the desired complexity bound. ∎

Acknowledgments. We are grateful to two anonymous referees for their suggestions on ways to improve the paper.

REFERENCES

[1] F. Alizadeh, J.-P. A. Haeberly, and M. L. Overton, *Primal-dual interior-point methods for semidefinite programming: Convergence rates, stability, and numerical results*, SIAM J. Optim., 8 (1998), pp. 746–768.

[2] A. Belloni, R. M. Freund, and S. Vempala, *Efficiency of a Re-scaled Perceptron Algorithm for Conic Systems*, Working paper OR 379-06, MIT Operations Research Center, Cambridge, MA, 2006.

[3] A. Berman, *Cones, Matrices, and Mathematical Programming*, Springer-Verlag, New York, 1973.

[4] A. Melman, *A unifying convergence analysis of second-order methods for secular equations*, Math. Comp., 66 (1997), pp. 333–344.

[5] G. Pataki, *On the closedness of the linear image of a closed convex cone*, Math. Oper. Res., 32 (2007), pp. 395–412.

[6] J. Renegar, *A Mathematical View of Interior-Point Methods in Convex Optimization*, SIAM, Philadelphia, 2001.

[7] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[8] S. Smale, *Newton's method estimates from data at one point*, in The Merging of Disciplines: New Directions in Pure, Applied and Computational Mathematics, R. Ewing, K. Gross, and C. Martin, eds., Springer-Verlag, New York, 1986, pp. 185–196.

[9] J. Sturm, *Using SeDuMi* 1.02*, a Matlab toolbox for optimization over symmetric cones*, Optim. Methods Softw., 11 & 12 (1999), pp. 625–653.

[10] Y. Ye, *A new complexity result for minimizing a general quadratic function with a sphere constraint*, in Recent Advances in Global Optimization, C. Floudas and P. Pardalos, eds., Princeton University Press, Princeton, NJ, 1992, pp. 19–31.

SIAM J. OPTIM.
Vol. 19, No. 3, pp. 1093–1106

RELAXED ALTERNATING PROJECTION METHODS*

ANDRZEJ CEGIELSKI† AND AGNIESZKA SUCHOCKA†

Abstract. In this paper we deal with the von Neumann alternating projection method $x_{k+1} = P_A P_B x_k$ and with its generalization of the form $x_{k+1} = P_A(x_k + \lambda_k(P_A P_B x_k - x_k))$, where A, B are closed and convex subsets of a Hilbert space $\mathcal{H}$ and $\operatorname{Fix} P_A P_B \neq \varnothing$. We do not suppose that $A \cap B \neq \varnothing$. We give sufficient conditions for the weak convergence of the sequence (x_k) to $\operatorname{Fix} P_A P_B$ in the general case and in the case A is a closed affine subspace. We present also the results of preliminary numerical experiments.

Key words. alternating projection method, Fejér monotonicity, weak convergence

AMS subject classification. 65K05

DOI. 10.1137/070698750

1. Introduction. Let $\mathcal{H}$ be a real Hilbert space equipped with a scalar product $\langle \cdot, \cdot \rangle$ and with the norm $\| \cdot \|$ induced by $\langle \cdot, \cdot \rangle$. Further, let $A, B \subset \mathcal{H}$ be nonempty, convex, and closed subsets. In the practical considerations one often needs to find an element of the intersection $A \cap B$ or, more generally, to solve the following problem:

$$\text{find } a^* \in A \text{ and } b^* \in B \text{ such that } \|a^* - b^*\| = \inf_{a \in A, b \in B} \|a - b\|. \tag{1}$$

We suppose that this infimum is attained. Of course, $a^* = b^*$ if and only if $A \cap B \neq \varnothing$. Several optimization problems, e.g., the convex feasibility problem, can be reduced to problem (1) (see, e.g., [18, section 2.9] for details). Problems of this kind have many practical applications, e.g., in signal reconstruction (see, e.g., [8] or [18, Chapter 6]), in image reconstruction, or in intensity modulated radiation therapy (see, e.g., [7, 10, 18, 9, 14]), where the convex subsets are described by a large and sparse system of linear equalities or inequalities.

An important method generating sequences which converge weakly to a solution of problem (1) is the von Neumann alternating projection (AP) method (see, e.g., [11, Chapter 9] or [1, section 4]). In this method the metric projections onto A and B are successively applied. Recall that for a closed and convex subset $D \subset \mathcal{H}$ and for any $u \in \mathcal{H}$ there exists the uniquely determined metric projection $P_D u$. Furthermore, a point $y \in D$ is the projection $P_D u$ if and only if

$$\langle u - y, z - y \rangle \leq 0 \text{ for all } z \in D; \tag{2}$$

i.e., inequality (2) characterizes the metric projection $P_D u$ (see, e.g., [12, Lemma 12.1] or [1, section 1]). It is known that $a^* \in A$ and $b^* \in B$ realize the distance between A and B if and only if $a^* = P_A b^*$ and $b^* = P_B a^*$, i.e., $a^* \in \operatorname{Fix} P_A P_B$ or $b^* \in \operatorname{Fix} P_B P_A$ (see, e.g., [1, Lemma 2.2(i)]). Therefore, it is enough to find an element of $\operatorname{Fix} P_A P_B$ in order to find a solution of problem (1). In this paper we construct a generalization of the von Neumann alternating projection method and prove its Fejér monotonicity

*Received by the editors July 30, 2007; accepted for publication (in revised form) June 11, 2008; published electronically October 31, 2008.
http://www.siam.org/journals/siopt/19-3/69875.html
†Faculty of Mathematics, Computer Sciences, and Econometrics, University of Zielona Góra, 65-516 Zielona Góra, ul. Szafrana 4a, Poland (a.cegielski@wmie.uz.zgora.pl, a.suchocka@wmie.uz.zgora.pl).

with respect to the solution set $\operatorname{Fix} P_A P_B$, as well as prove the weak convergence of the method to a solution. Recall that a sequence $(x_k) \subset \mathcal{H}$ is called *Fejér monotone* with respect to a subset $D \subset \mathcal{H}$ if for all $z \in D$ there holds $\|x_{k+1} - z\| \leq \|x_k - z\|$, $k = 1, 2, \ldots$.

Consider a sequence $(x_k) \subset \mathcal{H}$ generated by the following iterative scheme:

$$\begin{aligned} &x_0 \in A \text{ – arbitrary,} \\ &x_{k+1} = P_A(x_k + \lambda_k \sigma_k (P_A P_B x_k - x_k)), \end{aligned} \tag{3}$$

where the *relaxation parameter* $\lambda_k \in [0, 2]$ and the *step size* $\sigma_k \geq 0$. We call method (3) the *relaxed alternating projection (RAP) method*. If $\lambda_k = \sigma_k = 1$, we obtain the *von Neumann AP method*:

$$\begin{aligned} &x_0 \in A \text{ – arbitrary,} \\ &x_{k+1} = P_A P_B x_k \end{aligned} \tag{4}$$

(see, e.g., [1]). Some modifications of the AP method (4) for $A \cap B \neq \varnothing$ and for B being an obtuse cone, different from (3), were proposed in [5, section 3], where the projection P_B in (4) is replaced by the reflection $R_B = 2P_B - I$.

One can show that any sequence (x_k) generated by the AP method (4) converges weakly to an element $x^* \in \operatorname{Fix} P_A P_B$ (see, e.g., [1, Theorem 4.8 and Lemma 2.2]). Note that $\operatorname{Fix} P_A P_B \neq \varnothing$ since we have supposed that the infimum in (1) is attained. If $A \cap B \neq \varnothing$, then any sequence (x_k) generated by the RAP method (3) converges weakly to an element $x^* \in \operatorname{Fix} P_A P_B = A \cap B$ if $\sigma_k = 1$ and $\lambda_k \in [\varepsilon, 2 - \varepsilon]$, where $\varepsilon > 0$ (see, e.g., [2, Corollary 3.22] for a more general result). Gurin, Polyak, and Raik have proposed the following step size in order to accelerate the convergence of the RAP method in the case $A \cap B \neq \varnothing$:

$$\sigma_k = \frac{\|P_B x_k - x_k\|^2}{\langle P_B x_k - x_k, P_A P_B x_k - x_k \rangle} \tag{5}$$

(see [13, Theorem 4]). Recently, the idea of [13] was applied in the case A and B are subspaces of $\mathcal{H}$ (see [4, Theorem 3.23]) and in the case A is a closed affine subspace of $\mathcal{H}$ with $A \cap B \neq \varnothing$ (see [3, Corollary 4.11]). Unfortunately, the weak convergence of the RAP method with the relaxation parameter $\lambda_k \in [\varepsilon, 2 - \varepsilon]$ and the step size $\sigma_k = 1$ or the step size defined by (5) is not guaranteed if $A \cap B = \varnothing$. A new question arises in this context: What should we impose on the relaxation parameters λ_k and on the step sizes σ_k in order to obtain the weak convergence of the RAP method (3) to an element $x^* \in \operatorname{Fix} P_A P_B$, without assumption $A \cap B \neq \varnothing$? The answers to these questions are contained in Theorem 15, which is the main result of the paper.

In sections 2 and 3 we give some sufficient conditions for the quasi nonexpansivity of operators determining RAP methods. Recall that an operator $U : C \to \mathcal{H}$ is *quasi-nonexpansive* if for all $x \in C$ and for all $z \in \operatorname{Fix} U$ there holds the inequality

$$\|Ux - z\| \leq \|x - z\|$$

(see, e.g., [16]). Quasi-nonexpansive operators are also known in the literature under the name *attracting* operators (see, e.g., [2, Definition 2.1]) or *Fejér monotone* operators or mappings (see, e.g., [17, Definition 2.1]. In section 4 we show the weak convergence of RAP methods to a fixed point of the operator $P_A P_B$ for special choices of step sizes σ_k. In section 5 we present the results of preliminary numerical experiments.

2. Quasi nonexpansivity of relaxed alternating projections. Let A, B be nonempty, closed, and convex subsets of $\mathcal{H}$. Define the operator of *alternating projections* $T : A \to A$ by the equality

$$T = P_A P_B.$$

For a constant $\lambda \in [0, 2]$ we call the operator $T_\lambda = (1 - \lambda)I + \lambda T$ the *relaxation* of T and the operator $P_A T_\lambda$ the *projected relaxation* of T. Furthermore, for a relaxation parameter $\lambda \in [0, 2]$ we call the operator $T_{\sigma,\lambda} : A \to A$ defined by

$$T_{\sigma,\lambda} x = P_A(x + \lambda\sigma(x)(Tx - x)) \tag{6}$$

the RAP operator, where the nonnegative *step size function* $\sigma(x)$ depends on x; i.e., σ is a function, $\sigma : A \to \mathbb{R}_+ = [0, +\infty)$. Of course, $T_{\sigma,\lambda} = T$ if $\sigma(x) = 1$ for all $x \in A$ and $\lambda = 1$. The operator T defines the AP method since the iteration (4) can be written in the form $x_{k+1} = Tx_k$. Similarly, for a function $\sigma : A \to \mathbb{R}_+$ and for a sequence of relaxation parameters (λ_k) the operator $T_{\sigma,\lambda}$ defines the RAP method by the equality

$$x_{k+1} = T_{\sigma,\lambda_k} x_k, \tag{7}$$

which is equivalent to (3) with $\sigma_k = \sigma(x_k)$. First we give some properties of the operators T and $T_{\sigma,\lambda}$, which we use later to show the quasi nonexpansivity of $T_{\sigma,\lambda}$ and the weak convergence of a sequence generated by the recurrence (7) for special choices of the step size function $\sigma : A \to \mathbb{R}_+$.

LEMMA 1. *Let $\sigma(x) > 0$ for all $x \in A$, and let $\lambda > 0$. Then* $\operatorname{Fix} T_{\sigma,\lambda} = \operatorname{Fix} T$.

Proof. Denote by $N_D(y) = \{u \in \mathcal{H} : \langle u - y, z - y\rangle \le 0 \text{ for all } z \in D\}$ the *normal cone* to a closed and convex subset $D \subset \mathcal{H}$ at the point $y \in D$. By the equivalence

$$y = P_D u \Leftrightarrow u - y \in N_D(y), \tag{8}$$

where $D \subset \mathcal{H}$ is a closed and convex subset (see, e.g., [15, Chapter 1, Proposition 5.3.3]), and by the obvious fact that $N_D(y)$ is a cone, we have

$$\begin{aligned} x \in \operatorname{Fix} T_{\sigma,\lambda} &\Leftrightarrow P_A(x + \lambda\sigma(x)(Tx - x)) = x \\ &\Leftrightarrow \lambda\sigma(x)(Tx - x) \in N_A(x) \Leftrightarrow Tx - x \in N_A(x) \\ &\Leftrightarrow x = P_A Tx = Tx \Leftrightarrow x \in \operatorname{Fix} T, \end{aligned}$$

which completes the proof. □

It is easily seen that the characterization (2) of the metric projection $P_D u$ is equivalent to the condition

$$\langle z - u, P_D u - u\rangle \ge \|P_D u - u\|^2 \text{ for any } u \in \mathcal{H} \text{ and } z \in D. \tag{9}$$

Denote by $\delta = \mathrm{d}(A, B) = \inf_{x\in A, y\in B} \|x - y\|$ the distance between the subsets A and B. As we have supposed in section 1, δ is attained, and, consequently, $\operatorname{Fix} T \ne \varnothing$.

LEMMA 2. *Let $z \in \operatorname{Fix} T$. Then for any $x \in A$ there holds the inequality*

$$\langle z - x, Tx - x\rangle \ge \|Tx - P_B x\|^2 - \tilde{\delta}\|P_B x - x\| + \langle P_B x - x, Tx - x\rangle, \tag{10}$$

where $\tilde{\delta} \in [\delta, \|Tx - P_B x\|]$ is an upper bound of the distance δ.

Proof. Let $w = P_B z$. Observe that $\|z - w\| = \delta$ (see, e.g., [1, Lemma 2.2(i)]). We have by the characterization (2) of the metric projection and by the Cauchy–Schwarz inequality

$$\begin{aligned}\langle z - P_B x, P_B x - x\rangle &= \langle z - w, P_B x - x\rangle + \langle w - P_B x, P_B x - x\rangle \\ &\geq \langle z - w, P_B x - x\rangle \\ &\geq -\|z - w\| \cdot \|P_B x - x\| = -\delta\|P_B x - x\|.\end{aligned}$$

Therefore, if we apply condition (9) we obtain

$$\begin{aligned}\langle z - x, Tx - x\rangle &= \langle z - P_B x, Tx - x\rangle + \langle P_B x - x, Tx - x\rangle \\ &= \langle z - P_B x, Tx - P_B x\rangle + \langle z - P_B x, P_B x - x\rangle \\ &\quad + \langle P_B x - x, Tx - x\rangle \\ &\geq \|Tx - P_B x\|^2 - \delta\|P_B x - x\| + \langle P_B x - x, Tx - x\rangle.\end{aligned}$$

Now (10) follows from the inequality $\delta \leq \tilde{\delta}$. ∎

Let $x \in A$. Let

$$\bar{\delta} = \bar{\delta}(x) = \|Tx - P_B x\|. \tag{11}$$

Let

$$\tilde{\delta} = \tilde{\delta}(x) \in [\delta, \bar{\delta}(x)] \tag{12}$$

be an upper bound of the distance δ, and let $T_{\sigma,\lambda}$ be defined by (6), where the function $\sigma : A \to \mathbb{R}_+$ is given by

$$\sigma(x) = \frac{\|Tx - P_B x\|^2 - \tilde{\delta}\|P_B x - x\| + \langle P_B x - x, Tx - x\rangle}{\|Tx - x\|^2} \tag{13}$$

for $x \notin \operatorname{Fix} T$ and $\sigma(x) = 1$ for $x \in \operatorname{Fix} T$. For $x \in A$ define

$$\alpha(x) = \begin{cases} \sphericalangle(x - P_B x, Tx - P_B x) & \text{if } x \notin \operatorname{Fix} T \text{ and } P_B x \notin A, \\ 0 & \text{if } x \in \operatorname{Fix} T \text{ or } P_B x \in A, \end{cases} \tag{14}$$

where the symbol $\sphericalangle(a, b)$ denotes the angle between two nonzero vectors $a, b \in \mathcal{H}$, i.e., $\sphericalangle(a, b) = \arccos \frac{\langle a,b\rangle}{\|a\| \cdot \|b\|}$. Note that $\alpha(x)$ is well defined since $x - P_B x$ and $Tx - P_B x$ are obviously nonzero vectors in the first case of (14). Furthermore, we have by the characterization of the metric projection $P_A(P_B x)$

$$\langle P_B x - x, P_B x - Tx\rangle \geq \|P_B x - Tx\|^2 > 0 \tag{15}$$

and, consequently,

$$0 < \frac{\|P_B x - Tx\|}{\|P_B x - x\|} \leq \cos\alpha(x). \tag{16}$$

LEMMA 3. *Let $x \in A$, and let the step size $\sigma(x)$ be defined by* (13). *Then*

$$\sigma(x) \geq \frac{1}{1 + \cos\alpha(x)} \geq \frac{1}{2}. \tag{17}$$

Proof. Inequality (17) is clear if $x \in \operatorname{Fix} T$ or $P_B x \in A$. Suppose now that $x \notin \operatorname{Fix} T$ and $P_B x \notin A$. Let $a = P_B x - x$, $b = Tx - x$, and $c = P_B x - Tx$. Of course, $a, b, c \neq 0$ and $\alpha(x) = \sphericalangle(a, c)$. Observe that $b = a - c$, $\tilde{\delta} \leq \|c\|$, and that the function $y \longmapsto \frac{y+\rho}{y+2\rho}$ is increasing for $y > -2\rho$. Therefore, for $\rho = 1 - \cos\alpha(x)$, we have

$$\begin{aligned}
\sigma(x) &= \frac{\|c\|^2 - \tilde{\delta}\|a\| + \langle a, b\rangle}{\|b\|^2} \\
&\geq \frac{\|c\|^2 - \|a\| \cdot \|c\| + \langle a, b\rangle}{\|b\|^2} \\
&= \frac{\|c\|^2 - \|a\| \cdot \|c\| + \langle a, a - c\rangle}{\|a - c\|^2} \\
&= \frac{(\|a\| - \|c\|)^2 + \|a\| \cdot \|c\| - \langle a, c\rangle}{(\|a\| - \|c\|)^2 + 2(\|a\| \cdot \|c\| - \langle a, c\rangle)} \\
&= \frac{\left(1 - \frac{\|c\|}{\|a\|}\right)\left(\frac{\|a\|}{\|c\|} - 1\right) + 1 - \cos\alpha(x)}{\left(1 - \frac{\|c\|}{\|a\|}\right)\left(\frac{\|a\|}{\|c\|} - 1\right) + 2(1 - \cos\alpha(x))} \\
&\geq \frac{(1 - \cos\alpha(x))\left(\frac{1}{\cos\alpha(x)} - 1\right) + 1 - \cos\alpha(x)}{(1 - \cos\alpha(x))\left(\frac{1}{\cos\alpha(x)} - 1\right) + 2(1 - \cos\alpha(x))} \\
&= \frac{1}{1 + \cos\alpha(x)} \geq \frac{1}{2},
\end{aligned}$$

which completes the proof. ∎

LEMMA 4. *Let $x \in A$ be such that $Tx \notin \operatorname{Fix} T$. Then $\alpha(x) \in (0, \frac{\pi}{2})$ and, consequently, the vectors $x - P_B x$ and $Tx - P_B x$ are linearly independent.*

Proof. Suppose that $\alpha(x) = 0$, i.e.,

$$Tx - P_B x = \gamma(x - P_B x) \tag{18}$$

for some $\gamma > 0$. By the equivalence (8) we have $x - P_B x \in N_B(P_B x)$ and, consequently, $\gamma(x - P_B x) \in N_B(P_B x)$, and again by the equivalence (8),

$$P_B(P_B x + \gamma(x - P_B x)) = P_B x. \tag{19}$$

Now we obtain by (18) and (19)

$$\begin{aligned}
TTx &= T(P_B x + \gamma(x - P_B x)) \\
&= P_A P_B(P_B x + \gamma(x - P_B x)) \\
&= P_A P_B x = Tx,
\end{aligned}$$

a contradiction with the assumption $Tx \notin \operatorname{Fix} T$. Therefore, $\alpha(x) > 0$. Furthermore, $\alpha(x) < \frac{\pi}{2}$ by (16). Consequently, the vectors $x - P_B x$ and $Tx - P_B x$ are linearly independent. ∎

Remark 5. According to Lemma 4, we can stop the RAP algorithm (3) with $Tx_k \in \operatorname{Fix} T$ if we state that $P_B x_k - x_k$ and $Tx_k - P_B x_k$ are linearly dependent.

Let $x \in A$ be such that $Tx \notin \operatorname{Fix} T$. Let $y \in \operatorname{aff}(x, P_B x, Tx)$ be a solution of the system

$$\langle P_B x - y, P_B x - x\rangle = \tilde{\delta}\|P_B x - x\|, \tag{20}$$

$$\langle P_B x - y, Tx - P_B x\rangle = -\|Tx - P_B x\|^2, \tag{21}$$

where $\tilde{\delta} \in [\delta, \|Tx - P_B x\|]$. By Lemma 4 such a solution is defined uniquely.

LEMMA 6. *Let $x \in A$ be such that $Tx \notin \operatorname{Fix} T$. The step size $\sigma(x)$ given by* (13) *is characterized by the equality*

$$\langle x + \sigma(x)(Tx - x) - y, Tx - x\rangle = 0. \tag{22}$$

Proof. For y being a solution of the system (20)–(21) and for any σ we have

$$\begin{aligned}
&\langle x + \sigma(Tx - x) - y, Tx - x\rangle \\
&= \langle x - y, Tx - x\rangle + \sigma\|Tx - x\|^2 \\
&= \langle x - P_B x, Tx - x\rangle + \langle P_B x - y, Tx - x\rangle + \sigma\|Tx - x\|^2 \\
&= \langle x - P_B x, Tx - x\rangle + \langle P_B x - y, Tx - P_B x\rangle \\
&\quad + \langle P_B x - y, P_B x - x\rangle + \sigma\|Tx - x\|^2 \\
&= \langle x - P_B x, Tx - x\rangle - \|Tx - P_B x\|^2 + \tilde{\delta}\|P_B x - x\| + \sigma\|Tx - x\|^2.
\end{aligned}$$

Therefore, equalities (22) and (13) are equivalent. ∎

LEMMA 7. *Let $z \in \operatorname{Fix} T$, $x \in A$, and let $\sigma(x)$ be defined by* (13). *There holds the inequality*

$$\langle z - x, Tx - x\rangle \geq \sigma(x)\|Tx - x\|^2.$$

Proof. The lemma follows directly from Lemma 2 and from equality (13). ∎

THEOREM 8. *Let $x \in A$, and let $\sigma(x)$ be given by* (13). *Then for any $z \in \operatorname{Fix} T$ and for any $\lambda \geq 0$ there holds the inequality*

$$\|T_{\sigma,\lambda}x - z\|^2 \leq \|x - z\|^2 - \lambda(2 - \lambda)\sigma^2(x)\|Tx - x\|^2. \tag{23}$$

Consequently, the operator $T_{\sigma,\lambda}$ defined by (6) *is quasi-nonexpansive for $\lambda \in [0, 2]$.*

Proof. Let $z \in \operatorname{Fix} T$, $x \in A$, and $\lambda \geq 0$. Of course, $z = P_A z$. We have by the nonexpansivity of the metric projection P_A and by Lemma 7

$$\begin{aligned}
\|T_{\sigma,\lambda}x - z\|^2 &= \|P_A(x + \lambda\sigma(x)(Tx - x)) - z\|^2 \\
&= \|P_A(x + \lambda\sigma(x)(Tx - x)) - P_A z\|^2 \\
&\leq \|x + \lambda\sigma(x)(Tx - x) - z\|^2 \\
&= \|x - z\|^2 + \lambda^2\sigma^2(x)\|Tx - x\|^2 - 2\lambda\sigma(x)\langle z - x, Tx - x\rangle \\
&\leq \|x - z\|^2 + \lambda^2\sigma^2(x)\|Tx - x\|^2 - 2\lambda\sigma^2(x)\|Tx - x\|^2 \\
&= \|x - z\|^2 - \lambda(2 - \lambda)\sigma^2(x)\|Tx - x\|^2,
\end{aligned}$$

and we see that $T_{\sigma,\lambda}$ is quasi-nonexpansive if $\lambda \in [0, 2]$. ∎

Remark 9. Let $A \cap B \neq \varnothing$, and let $x \in A \setminus B$. We have $\delta = 0$, and the step size given by (13) with $\tilde{\delta} = 0$ has the form

$$\sigma(x) = \frac{\|Tx - P_B x\|^2 + \langle P_B x - x, Tx - x\rangle}{\|Tx - x\|^2}. \tag{24}$$

Gurin, Polyak, and Raik have proposed the relaxation parameter $\lambda = 1$ and the following step size $\sigma(x)$ in the relaxed alternating projection method:

$$\sigma(x) = \frac{\|P_B x - x\|^2}{\langle P_B x - x, Tx - x\rangle} \tag{25}$$

(see [13, equality (15)]). Observe that the step size $\sigma(x)$ defined by (25) is the unique solution of the equality

$$\langle x + \sigma(x)(Tx - x) - P_B x, P_B x - x\rangle = 0.$$

Consider two cases of the RAP method with the step size (25):

(i) If A, B are subspaces of H, then

$$\sigma(x) = \frac{\langle x, x - Tx\rangle}{\|Tx - x\|^2}.$$

In this case the RAP method is equivalent to an acceleration method of Bauschke et al. (see [4, equality (3.1.2) and Theorem 3.23]).

(ii) If A is a closed affine subspace, then $\langle P_B x - x, Tx - x\rangle = \|Tx - x\|^2$ and

$$\sigma(x) = \frac{\|P_B x - x\|^2}{\|Tx - x\|^2}.$$

In this case the RAP method is equivalent to the extrapolated alternating projection method (see [3, equality (4.35)]).

LEMMA 10. *Let $A \cap B \neq \varnothing$, and let $x \in A \setminus B$. Then we have*

$$\frac{\|Tx - P_B x\|^2 + \langle P_B x - x, Tx - x\rangle}{\|Tx - x\|^2} \geq \frac{\|P_B x - x\|^2}{\langle P_B x - x, Tx - x\rangle}; \tag{26}$$

i.e., the step size $\sigma(x)$ defined by (24) *is not shorter than the one proposed by Gurin, Polyak, and Raik* (*equality* (25)). *Furthermore, both step sizes are equal if A is a closed affine subspace.*

Proof. Observe that $\delta = 0$ since $A \cap B \neq \varnothing$ and that $x \notin \operatorname{Fix} T$ since $\operatorname{Fix} T = A \cap B$ for $A \cap B \neq \varnothing$. It follows from the characterization of the metric projection $P_A(P_B x)$ that

$$\langle x - Tx, P_B x - Tx\rangle \leq 0. \tag{27}$$

If we apply inequality (27), the Cauchy–Schwarz inequality, the nonexpansivity of the metric projection P_A, and the fact $x \neq Tx$ we easily obtain

$$0 < \|Tx - x\|^2 \leq \langle P_B x - x, Tx - x\rangle \leq \|P_B x - x\|^2. \tag{28}$$

A simple computation shows that

$$\begin{aligned}&\frac{\|Tx - P_B x\|^2 + \langle P_B x - x, Tx - x\rangle}{\|Tx - x\|^2}\\ &= \frac{\|Tx - x\|^2 + \|P_B x - x\|^2 - \langle P_B x - x, Tx - x\rangle}{\|Tx - x\|^2}.\end{aligned}$$

If we apply the last equality, we easily see that (26) is equivalent to the inequality

$$\left(\|P_B x - x\|^2 - \langle P_B x - x, Tx - x\rangle\right)\left(\langle P_B - x, Tx - x\rangle - \|Tx - x\|^2\right) \geq 0,$$

which is true by (28). Suppose now that A is a closed affine subspace. The equality in (26) follows easily from the fact that $\langle Tx - P_B x, Tx - x\rangle = 0$ for A being an affine subspace. ∎

3. Quasi nonexpansivity of the RAP operator for closed and affine A. In this section we suppose that $A \subset \mathcal{H}$ is a closed affine subspace. In this case $x + \sigma(Tx - x) \in A$ for any $x \in A$ and $\sigma \in \mathbb{R}$, where $T = P_A P_B$. Consequently, the RAP operator $T_{\sigma,\lambda} : A \to A$ defined by (6) has the form

$$T_{\sigma,\lambda}(x) = x + \lambda\sigma(x)(Tx - x) \tag{29}$$

and one iteration of the RAP method has the form

$$x_{k+1} = x_k + \lambda_k \sigma_k (Tx_k - x_k). \tag{30}$$

It is known that for A being an affine subspace the operator $T = P_A P_B$ restricted to A is firmly nonexpansive [6, Proposition 3(i)] and that the RAP method converges to an element of $\operatorname{Fix} T$ for $\sigma_k = 1$ and for $\lambda_k \in [\varepsilon, 2 - \varepsilon]$, where $\varepsilon > 0$ [6, Theorem 1] (see, e.g., [12, Chapter 12] for the definition and the properties of firmly nonexpansive operators). We generalize these results. We start with the following lemma.

LEMMA 11. *Let $A \subset \mathcal{H}$ be a closed affine subspace and $B \subset \mathcal{H}$ be a closed and convex subset. For all $x, y \in A$ there holds the inequality*

$$\langle Tx - Ty, x - y\rangle \geq \|Tx - Ty\|^2 + (\|Tx - P_B x\| - \|Ty - P_B y\|)^2.$$

Proof. Since the metric projection P_A is a firmly nonexpansive operator (see, e.g., [2, Fact 1.5]), we have for any $u, v \in \mathcal{H}$

$$\langle Tu - Tv, P_B u - P_B v\rangle \geq \|Tu - Tv\|^2. \tag{31}$$

Further, for any $u, v \in A$ we have, by the affinity of A,

$$\langle Tu - P_B u, u - P_B u\rangle = \|Tu - P_B u\|^2 \tag{32}$$

and

$$\langle P_B v - Tv, u - Tu\rangle = 0. \tag{33}$$

The characterization of the metric projection $P_B v$ yields

$$\langle P_B u - P_B v, v - P_B v\rangle \leq 0 \tag{34}$$

for any $u, v \in \mathcal{H}$. Now let $x, y \in A$. It follows from (31)–(34) and from the Cauchy–Schwarz inequality that

$$\begin{aligned}
&\langle Tx - Ty, x - y\rangle \\
&= \langle Tx - Ty, P_B x - P_B y\rangle + \langle Tx - Ty, (x - P_B x) - (y - P_B y)\rangle \\
&\geq \|Tx - Ty\|^2 \\
&\quad + \langle (Tx - P_B x) + (P_B x - P_B y) + (P_B y - Ty), (x - P_B x) - (y - P_B y)\rangle \\
&= \|Tx - Ty\|^2 + \langle Tx - P_B x, x - P_B x\rangle - \langle Tx - P_B x, y - P_B y\rangle \\
&\quad + \langle P_B x - P_B y, x - P_B x\rangle + \langle P_B y - P_B x, y - P_B y\rangle \\
&\quad + \langle P_B y - Ty, x - P_B x\rangle + \langle Ty - P_B y, y - P_B y\rangle \\
&\geq \|Tx - Ty\|^2 + \|Tx - P_B x\|^2 - (\langle Tx - P_B x, y - Ty\rangle \\
&\quad - \langle P_B x - Tx, Ty - P_B y\rangle) + \langle P_B y - Ty, x - Tx\rangle \\
&\quad + \langle P_B y - Ty, Tx - P_B x\rangle + \|Ty - P_B y\|^2 \\
&\geq \|Tx - Ty\|^2 + \|Tx - P_B x\|^2 - 2\|P_B x - Tx\| \cdot \|Ty - P_B y\| \\
&\quad + \|Ty - P_B y\|^2 \\
&= \|Tx - Ty\|^2 + (\|Tx - P_B x\| - \|Ty - P_B y\|)^2,
\end{aligned}$$

which completes the proof. ∎

COROLLARY 12 (Combettes [6]). *Let $A \subset \mathcal{H}$ be a closed affine subspace and $B \subset \mathcal{H}$ be a closed and convex subset. Then the operator $T : A \to A$, $T = P_A P_B$, is firmly nonexpansive.*

Let the function $\sigma : A \to \mathbb{R}_+$ be defined by

$$\sigma(x) = 1 + \frac{\left(\|Tx - P_B x\| - \tilde{\delta}\right)^2}{\|Tx - x\|^2}, \tag{35}$$

for $x \notin \operatorname{Fix} T$ and $\sigma(x) = 1$ for $x \in \operatorname{Fix} T$, where $\tilde{\delta}$ is given by (12).

LEMMA 13. *Let $z \in \operatorname{Fix} T$, $x \in A$, and let $\sigma(x)$ be defined by* (35). *There holds the inequality*

$$\langle z - x, Tx - x\rangle \geq \sigma(x)\|Tx - x\|^2.$$

Proof. The lemma is obvious for $x \in \operatorname{Fix} T$. Now let $x \notin \operatorname{Fix} T$. Since $\delta = \|Tz - P_B z\|$ we have by Lemma 11 that

$$\begin{aligned}
\langle z - x, Tx - x\rangle &= \|Tx - x\|^2 + \langle z - Tx, Tx - x\rangle \\
&= \|Tx - x\|^2 + \langle Tz - Tx, z - x\rangle - \|Tz - Tx\|^2 \\
&\geq \|Tx - x\|^2 + (\|Tx - P_B x\| - \|Tz - P_B z\|)^2 \\
&= \left(1 + \frac{(\|Tx - P_B x\| - \delta)^2}{\|Tx - x\|^2}\right)\|Tx - x\|^2 \\
&\geq \left(1 + \frac{\left(\|Tx - P_B x\| - \tilde{\delta}\right)^2}{\|Tx - x\|^2}\right)\|Tx - x\|^2,
\end{aligned}$$

and the lemma follows now from equality (35). ∎

COROLLARY 14. *Let $A \subset \mathcal{H}$ be a closed affine subspace and $B \subset \mathcal{H}$ be a closed and convex subset. Further, let $T_{\sigma,\lambda} : A \to A$ be defined by* (29), *where σ is defined by* (35). *Then for any $x \in A$, $z \in \operatorname{Fix} T$, and $\lambda \geq 0$ there holds the inequality*

$$\|T_{\sigma,\lambda}x - z\|^2 \leq \|x - z\|^2 - \lambda(2 - \lambda)\sigma^2(x)\|Tx - x\|^2, \tag{36}$$

and, consequently, $T_{\sigma,\lambda}$ is quasi-nonexpansive for $\lambda \in [0, 2]$.

Proof. Let $x \in A$, $z \in \operatorname{Fix} T$, and $\lambda \geq 0$. We have by Lemma 13 that

$$\begin{aligned}
\|T_{\sigma,\lambda}x - z\|^2 &= \|x + \lambda\sigma(x)(Tx - x) - z\|^2 \\
&= \|x - z\|^2 + \lambda^2\sigma^2(x)\|Tx - x\|^2 - 2\lambda\sigma(x)\langle z - x, Tx - x\rangle \\
&\leq \|x - z\|^2 - \lambda(2 - \lambda)\sigma^2(x)\|Tx - x\|^2.
\end{aligned}$$

Now we see that for $\lambda \in [0, 2]$ the operator $T_{\sigma,\lambda}$ is quasi-nonexpansive. ∎

4. Convergence of the RAP method. We consider in this section two cases of the RAP method (3) with $\lambda_k \in [\varepsilon, 2 - \varepsilon]$ for $\varepsilon > 0$:

(i) $A, B \subset \mathcal{H}$ are closed convex subsets, and the step size σ_k is given by

$$\sigma_k = \frac{\|Tx_k - P_B x_k\|^2 - \tilde{\delta}_k\|P_B x_k - x_k\| + \langle P_B x_k - x_k, Tx_k - x_k\rangle}{\|Tx_k - x_k\|^2}. \tag{37}$$

(ii) $A \subset \mathcal{H}$ is a closed affine subspace, $B \subset \mathcal{H}$ is a closed convex subset, and the step size σ_k is given by

$$\sigma_k = 1 + \frac{\left(\|Tx_k - P_B x_k\| - \tilde{\delta}_k\right)^2}{\|Tx_k - x_k\|^2}. \tag{38}$$

In both cases $\tilde{\delta}_k = \tilde{\delta}(x_k) \in [\delta, \bar{\delta}_k]$, where $\bar{\delta}_k = \bar{\delta}(x_k) = \|Tx_k - P_B x_k\|$.

THEOREM 15. *In both cases* (i) *and* (ii) *the sequence* (x_k) *converges weakly to an element* $x^* \in \operatorname{Fix} T$.

Proof. By Lemma 1 we have $\operatorname{Fix} T_{\sigma,\lambda} = \operatorname{Fix} T$. If we set $x = x_k$ in inequality (23) or in inequality (36) if A is a closed and affine subspace, we obtain in both cases for any $z \in \operatorname{Fix} T$

$$\|x_{k+1} - z\|^2 \le \|x_k - z\|^2 - \lambda_k(2 - \lambda_k)\sigma_k^2 \|Tx_k - x_k\|^2.$$

Therefore, $(\|x_k - z\|)$ converges as a nonincreasing sequence. Consequently,

$$\|Tx_k - x_k\| \to 0 \tag{39}$$

since for σ_k given by (37) we have $\sigma_k \ge \frac{1}{2}$ (see Lemma 3), for a closed affine subspace A and for σ_k given by (38) we have $\sigma_k \ge 1$, and $\lambda_k(2 - \lambda_k) \ge \varepsilon^2 > 0$. Let (x_{n_k}) be any weakly convergent subsequence of (x_k), and let $x \in A$ be the weak limit of x_{n_k}. Note that such a subsequence exists since (x_k) is bounded. Since T is nonexpansive we have by (39) that $x \in \operatorname{Fix} T$ (see, e.g., [2, Fact 1.2]). We have proved that all weak cluster points of (x_k) lie in $\operatorname{Fix} T$. Furthermore, $\operatorname{Fix} T$ is closed and convex (see, e.g., [1, Lemma 2.2(ii)]). Since the sequence (x_k) is Fejér monotone with respect to $\operatorname{Fix} T$, it converges weakly to some point $x^* \in \operatorname{Fix} T$; see [2, Theorem 2.16(ii)]. ∎

5. The results of preliminary numerical experiments. In this section we present the results of preliminary numerical tests for problem (1), where $\mathcal{H} = \mathbb{R}^n$.

5.1. Problems. We consider the following test problems:

P1. $A = B(z_1, 1)$ and $B = B(z_2, 1)$ are two balls in $\mathbb{R}^n$ with centers $z_1, z_2 \in \mathbb{R}^n$ and radius one. We consider this problem for various distances $d = \|z_1 - z_2\|$. Of course, $\delta = \mathrm{d}(A, B) = \max\{0, d - 2\}$, and, consequently, $A \cap B \ne \varnothing$ if and only if $d \le 2$. Without loss of generality we suppose that $n = 2$ and $z_1 = (0, d)$ for $d \in \mathbb{R}_+$ and $z_2 = (0, 0)$. We set $x_0 = (1, d) \in \mathbb{R}^2$ as the starting point. The exact solution of (1) can be easily evaluated analytically for the test problem P1, $x^* = (0, d-1)$ for $d \ge 2$ and $x^* = (\sqrt{4 - d^2}/2, d/2)$ for $d < 2$.

P2. A is a hyperplane and $B = B(z, 1)$ is a ball in $\mathbb{R}^n$. We consider this problem for various distances $d = \inf_{y \in A} \|z - y\|$. Of course, $\delta = \mathrm{d}(A, B) = \max\{0, d - 1\}$, and, consequently, $A \cap B \ne \varnothing$ if and only if $d \le 1$. Without loss of generality we suppose that $n = 2$ and $A = \{(\xi_1, \xi_2) \in \mathbb{R}^2 : \xi_2 = d\}$ for $d \in \mathbb{R}_+$ and $B = B((0, 0), 1)$. We set $x_0 = (3, d) \in \mathbb{R}^2$ as the starting point. The exact solution of (1) can be easily evaluated analytically for the test problem P1, $x^* = (0, d)$ for $d \ge 1$ and $x^* = (\sqrt{1 - d^2}, d)$ for $d < 1$.

Note that in all problems the starting point $x_0 \in \mathbb{R}^n$ belongs to A.

5.2. Tests. Now we present the results of numerical tests for the following methods:

- AP. The von Neumann alternating projection method (4), applied to problem (1),

TABLE 1
*Comparison of the AP, GPR, and RAP*1(i) *methods for problem* P1 *in case* $A \cap B \neq \emptyset$.

Method →		AP	GPR	RAP1(i)		
$\lambda \rightarrow$				1	$\frac{3}{2}$	2
d	ε	k	k	k	k	k
1.9	10^{-1}	2	1	1	1	1
	10^{-2}	6	2	2	1	1
	10^{-3}	11	3	3	1	1
1.99	10^{-1}	7	2	2	1	2
	10^{-2}	43	3	3	1	2
	10^{-3}	97	4	4	1	2
2	10^{-1}	25	3	3	1	2
	10^{-2}	2499	6	6	2	2
	10^{-3}	249999	10	9	4	3

- RAP1. The relaxed alternating projection method (3), where the step size σ_k is defined by (37), applied to problem (1),
- RAP2. The relaxed alternating projection method (3), where the step size σ_k is defined by (38), applied to problem (1) with affine A,
- GPR. The method proposed by Gurin, Polyak, and Raik, i.e., the relaxed alternating projection method (3), where the step size σ_k is defined by (5), applied to problem (1) with $A \cap B \neq \varnothing$.

In the presented tests we employ various values of constant relaxation parameter $\lambda_k = \lambda \in (0, 2)$. For the RAP1 and RAP2 methods we consider two cases:

(i) The value δ is known, and we set $\tilde{\delta}_k = \delta$ in (37) and in (38).

(ii) We set $\tilde{\delta}_k = \bar{\delta}_k = \|Tx_k - P_B x_k\|$ in (37) and in (38).

For both test problems P1 and P2 we know the exact solution x^* of (1) and we apply the condition $\|x_k - x^*\| \leq \varepsilon$ or $x_k \in A \cap B$ as the stopping criterion. Let k denote the number of iterations after which the corresponding algorithm terminates. All tested methods were programmed in MATLAB 6.1.

In Table 1 we present the numerical results of the AP, GPR, and RAP1(i) methods for problem P1 with $A \cap B \neq \varnothing$ for various distances d between the centers of two balls and for various optimality tolerances ε. The results of RAP1(i) are presented for three values of relaxation parameter λ (note that the AP and GPR methods are originally constructed only for $\lambda = 1$). The results for $\lambda = 1$ are repeated in Figure 1. We see that for all optimality tolerances the behavior of the RAP1(i) and GPR methods is similar and is considerably better than for the AP method. Observe that RAP1(i) behaves a little bit better if $\lambda > 1$.

In Table 2 we compare the AP, RAP1(i), and RAP1(ii) methods for problem P1 for various optimality tolerances ε. We consider here both cases: $A \cap B \neq \varnothing$ ($d \leq 2$) as well as $A \cap B = \varnothing$ ($d > 2$). The results for $d > 2$ are repeated in Figure 2. Note that we cannot apply GPR if $A \cap B = \varnothing$. Observe that RAP1(i) behaves essentially better than RAP1(ii) and the AP method. The most considerable differences are in case $d = 2$. In this case RAP1(i) behaves very well, while RAP1(ii) as well as the AP method converge very slowly because of zigzagging (the angle between the vectors $P_B x_k - x_k$ and $P_A P_B x_k - P_B x_k$ is close to π for x_k closed to the solution x^*).

In Table 3 we present the results of numerical tests for problem P2. We compare RAP2(i) ($\tilde{\delta}_k = \delta$) and RAP2(ii) ($\tilde{\delta}_k = \bar{\delta}_k$). In the second case $\sigma_k = 1$ for both methods. Furthermore, for $\lambda = 1$ RAP2(ii) reduces to the AP method. We set $d = 1$ (the hyperplane A is tangent to B in the solution x^*). For such d the termination

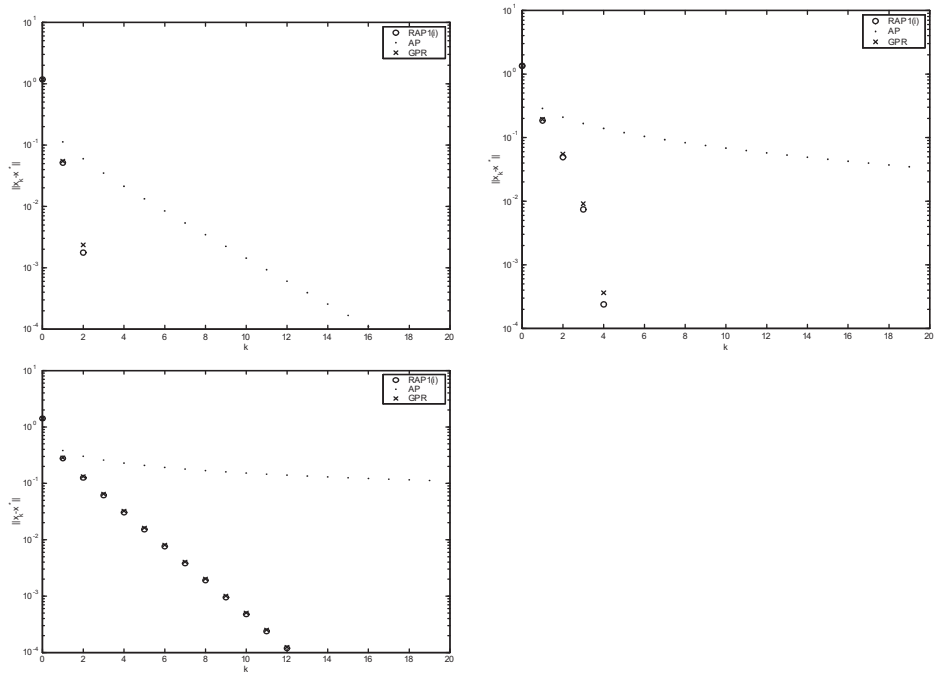


FIG. 1. *Comparison of the AP, GPR, and RAP*1(i) *methods for problem* P1 *in case* $A \cap B \neq \varnothing$.

TABLE 2
*Comparison of the AP and RAP*1 *methods for problem* P1.

Method →		AP	RAP1					
$\lambda \to$			1		$\frac{3}{2}$		2	
$\tilde{\delta}_k \to$			δ	$\bar{\delta}_k$	δ	$\bar{\delta}_k$	δ	$\bar{\delta}_k$
d	ε	k	k	k	k	k	k	k
1.9	10^{-1}	2	1	3	1	1	1	1
	10^{-2}	6	2	12	1	1	1	1
	10^{-3}	11	3	20	1	1	1	1
1.99	10^{-1}	7	2	13	1	1	2	1
	10^{-2}	43	3	86	1	1	2	1
	10^{-3}	97	4	194	1	1	2	1
2	10^{-1}	25	3	49	1	1	2	1
	10^{-2}	2499	6	5000	2	3297	2	2474
	10^{-3}	249999	9	500000	4	333296	3	249972
2.01	10^{-1}	17	3	34	1	1	2	2
	10^{-2}	116	6	232	2	130	2	100
	10^{-3}	231	9	464	4	284	3	215
2.1	10^{-1}	6	3	12	1	1	2	3
	10^{-2}	18	6	37	3	14	3	14
	10^{-3}	30	9	62	4	30	4	26
3	10^{-1}	2	2	4	2	1	2	2
	10^{-2}	3	5	9	2	4	4	4
	10^{-3}	5	7	14	3	6	6	5

of RAP2(ii) requires essentially more iterations than that of RAP2(i). Note that in this case A and B are almost "parallel" near the solution and the angle between the vectors $P_B x_k - x_k$ and $P_A P_B x_k - P_B x_k$ is close to π. We observe a small influence of

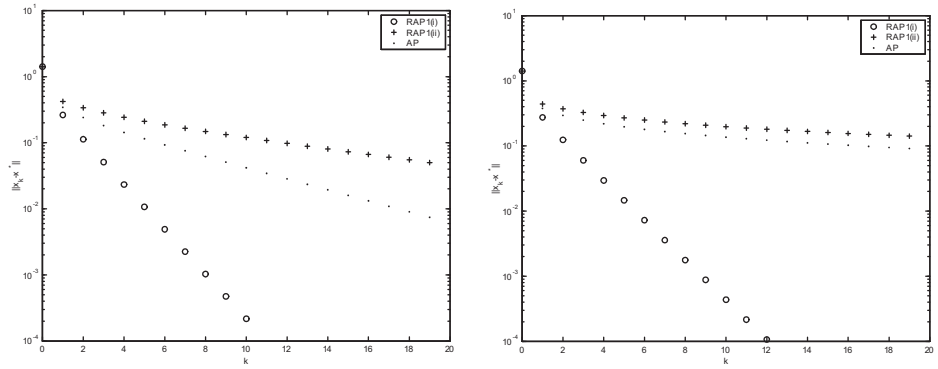


FIG. 2. *Comparison of the AP and RAP*1 *methods for problem P*1 *in case* $A \cap B = \varnothing$.

TABLE 3
*Numerical results of RAP*2 *for problem P*2.

Method →	RAP2(i)	RAP2(ii)	RAP2(i)	RAP2(ii)	RAP2(i)	RAP2(ii)
$\lambda \rightarrow$	1		$\frac{3}{2}$		2	
$\bar{\delta}_k \rightarrow$	δ	$\bar{\delta}_k$	δ	$\bar{\delta}_k$	δ	$\bar{\delta}_k$
ε	k	k	k	k	k	k
10^{-1}	5	100	3	1	4	48
10^{-2}	8	10000	5	6555	5	4996
10^{-3}	12	$> 5 \cdot 10^5$	7	$> 5 \cdot 10^5$	5	499995
10^{-4}	15	$> 5 \cdot 10^5$	8	$> 5 \cdot 10^5$	5	$> 5 \cdot 10^5$
10^{-5}	18	$> 5 \cdot 10^5$	10	$> 5 \cdot 10^5$	5	$> 5 \cdot 10^5$
10^{-6}	22	$> 5 \cdot 10^5$	12	$> 5 \cdot 10^5$	5	$> 5 \cdot 10^5$

parameter λ on the convergence. Furthermore, the behavior of RAP2(i) is essentially better (we use here the known distance δ between A and B) than that of RAP2(ii).

Preliminary numerical experiments show that both relaxed alternating projection methods behave essentially better than the original alternating projection method if the distance $\delta = d(A, B)$ is known. The most significant difference in the behavior of both methods with respect to the AP method can be observed if $A \cap B$ consists of one point or the distance δ is close to zero and the subsets A and B are almost "parallel" close to the solution.

Acknowledgment. The authors wish to thank the anonymous referee for his helpful remarks.

REFERENCES

[1] H. H. BAUSCHKE AND J. M. BORWEIN, *Dykstra's alternating projection algorithm for two sets*, J. Approx. Theory, 79 (1994), pp. 418–443.
[2] H. H. BAUSCHKE AND J. M. BORWEIN, *On projection algorithms for solving convex feasibility problems*, SIAM Rev., 38 (1996), pp. 367–426.
[3] H. H. BAUSCHKE, P. L. COMBETTES, AND S. G. KRUK, *Extrapolation algorithm for affine-convex feasibility problems*, Numer. Algorithms, 41 (2006), pp. 239–274.
[4] H. H. BAUSCHKE, F. DEUTSCH, H. HUNDAL, AND S.-H. PARK, *Accelerating the convergence of the method of alternating projection*, Trans. Amer. Math. Soc., 355 (2003), pp. 3433–3461.
[5] H. H. BAUSCHKE AND S. G. KRUK, *The method of reflection-projection for convex feasibility problems with an obtuse cone*, J. Optim. Theory Appl., 120 (2004), pp. 503–531.

[6] P. L. Combettes, *Inconsistent signal feasibility problems: Least-square solutions in a product space*, IEEE Trans. Signal Process., 42 (1994), pp. 2955–2966.
[7] P. L. Combettes, *The convex feasibility problem in image recovery*, in Advances in Imaging and Electron Physics, Vol. 95, P. Hawkes, ed., Academic Press, New York, 1996, pp. 155–270.
[8] P. L. Combettes and B. Bondon, *Hard-constrained inconsistent signal feasibility problems*, IEEE Trans. Signal Process., 47 (1999), pp. 2460–2468.
[9] Y. Censor, D. Gordon, and R. Gordon, *Component averaging: An efficient iterative parallel algorithm for large and sparse unstructured problems*, Parallel Comput., 27 (2001), pp. 777–808.
[10] Y. Censor and S. A. Zenios, *Parallel Optimization, Theory, Algorithms and Applications*, Oxford University Press, New York, 1997.
[11] F. Deutsch, *Best Approximation in Inner Product Spaces*, Springer-Verlag, New York, 2001.
[12] K. Goebel and W. A. Kirk, *Topics in Metric Fixed Point Theory*, Cambridge University Press, Cambridge, 1990.
[13] L. G. Gurin, B. T. Polyak, and E. V. Raik, *The method of projection for finding the common point in convex sets*, Zh. Vychisl. Mat. Mat. Fiz., 7 (1967), pp. 1211–1228 (in Russian); U.S.S.R. Comput. Math. Phys., 7 (1967), pp. 1–24 (in English).
[14] H. W. Hamacher and K.-H. Küfer, *Inverse radiation therapy planning – a multiple objective optimization approach*, Discrete Appl. Math., 118 (2002), pp. 145–161.
[15] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* I, Springer-Verlag, Berlin, 1993.
[16] S. A. Hirstoaga, *Iterative selection methods for common fixed point problems*, J. Math. Anal. Appl., 324 (2006), pp. 1020–1035.
[17] D. Schott, *Basic properties of Fejér monotone mappings*, Rostock. Math. Kolloq., 50 (1997), pp. 71–84.
[18] H. Stark and Y. Yang, *Vector Space Projections. A Numerical Approach to Signal and Image Processing, Neural Nets and Optics*, John Wiley, New York, 1998.

SIAM J. OPTIM.
Vol. 19, No. 3, pp. 1107–1130

FIXED-POINT CONTINUATION FOR ℓ_1-MINIMIZATION: METHODOLOGY AND CONVERGENCE*

ELAINE T. HALE†, WOTAO YIN†, AND YIN ZHANG†

Abstract. We present a framework for solving the large-scale ℓ_1-regularized convex minimization problem:

$$\min \|x\|_1 + \mu f(x).$$

Our approach is based on two powerful algorithmic ideas: operator-splitting and continuation. Operator-splitting results in a fixed-point algorithm for any given scalar μ; continuation refers to approximately following the path traced by the optimal value of x as μ increases. In this paper, we study the structure of optimal solution sets, prove finite convergence for important quantities, and establish q-linear convergence rates for the fixed-point algorithm applied to problems with $f(x)$ convex, but not necessarily strictly convex. The continuation framework, motivated by our convergence results, is demonstrated to facilitate the construction of practical algorithms.

Key words. ℓ_1 regularization, fixed-point algorithm, q-linear convergence, continuation, compressed sensing

AMS subject classifications. 65K05, 90C06, 90C25, 90C90

DOI. 10.1137/070698920

1. Introduction.

Under suitable conditions, minimizing the ℓ_1-norm is equivalent to minimizing the so-called "ℓ_0-norm," that is, the number of nonzeros in a vector. The former is always more computationally tractable than the latter. Thus, minimizing or limiting the magnitude of $\|x\|_1$ has long been recognized as a practical avenue for obtaining sparse solutions x. Some early work is in the area of geophysics, where sparse spike train signals are often of interest, and data may include large sparse errors [10, 39, 55, 57]. The signal processing and statistics communities use the ℓ_1-norm to describe a signal with just a few waveforms or a response quantity with just a few explanatory variables [9, 24, 44, 58]. More references on ℓ_1-regularization for signal processing and statistics can be found in [46].

In this work, we present an algorithmic framework and related convergence analysis for solving general problems of the form

$$\min_{x\in\mathbb{R}^n} \|x\|_1 + \mu f(x), \tag{1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable and convex, but not necessarily strictly convex, and $\mu > 0$. Interesting special cases of this problem include

$$\min_{x\in\mathbb{R}^n} \|x\|_1 + \frac{\mu}{2}\|Ax - b\|_2^2, \tag{1.2}$$

*Received by the editors July 31, 2007; accepted for publication (in revised form) June 17, 2008; published electronically October 31, 2008.

http://www.siam.org/journals/siopt/19-3/69892.html

†Department of Computational and Applied Mathematics, Rice University, 6100 Main Street, MS-134, Houston, TX 77005 (ehale@rice.edu, wotao.yin@rice.edu, yzhang@rice.edu). The work of E. Hale was supported by an NSF VIGRE grant (DMS-0240058). The work of W. Yin was supported in part by NSF CAREER award DMS-0748839 and ONR grant N00014-08-1-1101. The work of Y. Zhang was supported in part by NSF grants DMS-0405831 and DMS-0811188 and ONR grant N00014-08-1-1101.

and its generalization

$$\min_{x\in\mathbb{R}^n} \|x\|_1 + \frac{\mu}{2}\|Ax-b\|_M^2, \tag{1.3}$$

where $M \in \mathbb{R}^{m\times m}$ is a positive definite matrix, $\|x\|_M := \sqrt{x^\top Mx}$ is the associated M-norm, $A \in \mathbb{R}^{m\times n}$ is dense, $m \le n$ or even $m \ll n$, and n is large.

As a general principle, a sparse solution, $x \in \mathbb{R}^n$, of an underdetermined linear system of equations, $Ax = b$, may be obtained by minimizing the ℓ_1-norm of x. If the "observation" b is contaminated with noise, then an appropriate norm of the residual $Ax-b$ should be minimized or constrained. Such considerations yield several related optimization problems. For instance, if there is Gaussian noise distributed as $N(0,\sigma^2 I)$ in b, then the ℓ_1-regularized least squares problem (1.2) would be appropriate, as would the least absolute shrinkage and selection operator (LASSO) problem (6.3) [58].

Such problems are of fundamental importance to *compressed sensing*. Compressed sensing is the name assigned to the idea of encoding a large sparse signal using a relatively small number of linear measurements, and minimizing the ℓ_1-norm (or its variants) in order to decode the signal. Recent results reported by Candes et al. [4, 5, 6], Donoho et al. [16, 20, 61], and others ([54, 60], for example) stimulated the current burst of research in this area. Applications of compressed sensing include compressive imaging [56, 65, 66], medical imaging [42], multisensor and distributed compressed sensing [1], analog-to-information conversion [36, 37, 38, 59], and missing data recovery [67]. Compressed sensing is attractive for these and other potential applications because it reduces the number of measurements required to obtain a given amount of information. The tradeoff is the addition of a nontrivial decoding process that consists of solving problems like (1.2), where the data matrix A is usually either random or has its rows taken from an orthogonal matrix such as a discrete cosine transform (DCT) matrix. Such data matrices are invariably dense and large in applications of interest. Thus we are motivated to study algorithms that do not require any linear system solves or matrix factorizations, and are able to take advantage of available fast transforms like FFT and DCT.

1.1. Our approach and main results. The objective function in (1.1) is the sum of two convex functions. While the ℓ_1-norm term is not smooth, it is easily transformed into a linear function plus some linear constraints, such that standard interior-point methods utilizing a direct linear solver can be applied to, say, problem (1.2). However, such a standard approach is too costly for large-scale problems with dense data.

Our approach is based on operator splitting. It is well known in convex analysis that minimizing a convex function $\phi(x)$ is equivalent to finding a zero of the subdifferential $\partial\phi(x)$, i.e., finding x such that $\mathbf{0} \in \partial\phi(x) := T(x)$, where T is a maximal monotone operator [53]. In many cases, one can split ϕ into the sum of two convex functions, $\phi = \phi_1 + \phi_2$, which implies the decomposition of T into the sum of two maximal monotone operators T_1 and T_2, i.e., $T = T_1 + T_2$. For $\tau > 0$, if T_2 is single-valued and $(I + \tau T_1)$ is invertible, then

$$\begin{aligned} \mathbf{0} \in T(x) &\iff \mathbf{0} \in (x + \tau T_1(x)) - (x - \tau T_2(x)) \\ &\iff (I - \tau T_2)x \in (I + \tau T_1)x \\ &\iff x = (I + \tau T_1)^{-1}(I - \tau T_2)x. \end{aligned} \tag{1.4}$$

Equation (1.4) leads to the *forward-backward splitting* algorithm for finding a zero of T:

$$x^{k+1} := (I + \tau T_1)^{-1}(I - \tau T_2)x^k, \tag{1.5}$$

which is a fixed-point algorithm. For the minimization problem (1.1), $T_2 = \mu\nabla f$ and $(I + \tau T_1)^{-1}$ is component-wise shrinkage (or soft-thresholding), which is related to the ℓ_1-norm term in (1.1) and is described fully in sections 3 and 4 below.

The forward-backward splitting method was first proposed by Lions and Mercier [40] and Passty [51] at about the same time in 1979. Over the years, this scheme and its modifications have been extensively studied by various authors, including, to name a few, Mercier [45], Gabay [30], Glowinsky and Le Tallec [31], Eckstein [22], Chen and Rockafellar [8], Haubruge, Nguyen, and Strodiot [33], Noor [48], and Tseng [62]. The idea of splitting operators can be traced back to the mid-1950's in the works of Peaceman and Rachford [52] and Douglas and Rachford [21] for solving second-order elliptic and parabolic partial differential equations.

General convergence theory exists for forward-backward splitting methods [8, 30, 45]. Unfortunately, it requires rather strong conditions on T_2, or on T as a whole. In short, when reduced to our setting with $\phi_1 = \|x\|_1$ and $\phi_2 = \mu f(x)$, the classical convergence theory requires either f or the whole objective in (1.1) to be strictly convex (though such strong assumptions may be weakened with modifications to the basic algorithm [62]). An anonymous referee brought our attention to the recent paper [12] by Combettes and Wajs, which applies forward-backward splitting methods to various concrete forms of minimizing a sum of two convex functions, including problem (1.1). Combettes and Wajs [12] show that the fixed-point iterations (1.5), and some extensions of it, converge to a global minimum without a strict convexity assumption.

In the present work, we aim to address the following two questions, "Can stronger convergence results be obtained for algorithm (1.5) applied to problem (1.1)?" and "Can algorithm (1.5) be computationally competitive when applied to problem (1.2) or (1.3)?" Our answers to both questions are affirmative.

On the theoretical side, we have obtained finite convergence for some interesting quantities (cf. Theorems 4.5 and 4.7) and q-linear[1] rates of convergence (cf. Proposition 4.9 and Theorems 4.10 and 4.11) without assuming strict convexity, nor uniqueness of solution. Furthermore, we show that these q-linear rates of convergence are not determined by the conditioning of the Hessian of f, as is normally the case for gradient-type methods, but by that of a "reduced" Hessian whose condition number can be much smaller than that of the full Hessian when the solution x is sparse.

On the computational side, we devised a continuation strategy that significantly reduces the number of iterations required for a given value of μ. Our extensive numerical results, which will be presented in a separate paper [32] due to space limitation, indicate that our algorithm is especially well suited for large-scale instances of problem (1.2) when x^* is sufficiently sparse and A is a partial transform matrix such as a partial DCT matrix. In comparison with several recently developed algorithms, our algorithm appears to be the most robust, and in many cases also the fastest.

1.2. Related work. Recently, solving problems (1.1) or (1.2), especially (1.2), has been actively studied by many authors, largely because of its newly found applications in signal and image processing. These problems can be solved by the forward-backward operator-splitting method given by (1.5) if one substitutes T_1 and T_2 by the subdifferential of $\|x\|_1$ and the gradient of $\mu f(x)$, respectively, resulting in the formula (3.1) of section 3. Becau se the operator-splitting approach was priorly not widely known in signal and image processing areas, in most recent works the

[1] $\{x^k\}$ converges to x^* q-linearly, where q stands for "quotient," if $\limsup_k \|x^{k+1} - x^*\|/\|x^k - x^*\| < 1$.

fixed-point iteration (3.1), namely (1.5) specialized to (1.1), was derived by different authors, often independently, based on different motivations and approaches.

Here we mention a number of recent works that proposed, derived, or analyzed the fixed-point iteration scheme (1.5) or its variants when applied to either problem (1.2) or (1.1). These contributions include [29] by Figueiredo and Nowak, [15] by De Mol and Defrise, [13] by Daubechies, Defrise, and De Mol, [2] by Aubert, Bect, Blanc-Feraud, and Chambolle, [25, 26, 27] by Elad et al., and [**?**] by Darbon and Osher (though this list is unavoidably nonexhaustive). Some of these works were done independently around the same time, such as the two earlier papers [29] and [15]. Although the derivations and analyses in these papers were conducted through different means other than forward-backward operator splitting, the theoretical and numerical results in these papers contribute to a better understanding on the behavior of the fixed-point iterations (1.5) when applied to problem (1.2) or (1.1). For example, the authors of [2] and those of [13] analyzed the model (1.2) and independently proved global convergence without a strict convexity assumption, while classic convergence results for forward-backward operator splitting require stronger assumptions, and the authors of [26] proposed extensions and enhancements to the basic fixed-point iterations to improve its practical performance. As is already mentioned, Combettes and Wajs [12] recently established global convergence of the fixed-point iterations (1.5) when applied to (1.1) without a strict convexity assumption. A more recent paper by Combettes and Pesquet [11] studies closely related proximal soft-thresholding algorithms.

Alternative algorithms for the unconstrained ℓ_1-problem (1.3) include an iterative linear solver in an interior-point framework [35] by Kim et al., a gradient projection and Barzilai–Borwein method applied to an equivalent box-constrained QP [28] by Figueiredo, Nowak, and Wright, a direct and accelerated projected gradient method [14] by Daubechies, Fornasier, and Loris, an accelerated multistep gradient method [47] with an error convergence rate $O(1/k^2)$ by Nesterov, and a "two-step" shrinkage-based algorithm [3] by Bioucas–Dias and Figueiredo. In [64], Van den Berg and Friedlander apply an iterative method for solving the LASSO problem (6.3).

1.3. Notation and organization. For simplicity, we let $\|\cdot\| := \|\cdot\|_2$, the Euclidean norm, unless otherwise specified. The *support* of $x \in \mathbb{R}^n$ is $\mathrm{supp}(x) := \{i : x_i \neq 0\}$. Let

$$g(x) := \nabla f(x)$$

be the gradient of $f(x)$; in particular, $g(x) = A^\top M(Ax - b)$ for f defined by (1.3), that is

$$f = \frac{1}{2}\|Ax - b\|_M^2. \tag{1.6}$$

For a set E, we use $|E|$ to denote its cardinality. For any symmetric matrix $B \in \mathbb{R}^{n\times n}$, we denote its eigenvalues as $\lambda_i(B)$, $i = 1, \ldots, n$, and its maximum and minimum eigenvalues as, respectively, $\lambda_{\max}(B)$ and $\lambda_{\min}(B)$.

The signum function of $t \in \mathbb{R}$ is

$$\mathrm{sgn}(t) := \begin{cases} +1 & t > 0, \\ 0 & t = 0, \\ -1 & t < 0, \end{cases}$$

while the signum multifunction (i.e., set-valued function) of $t \in \mathbb{R}$ is

$$\mathrm{SGN}(t) := \partial|t| = \begin{cases} \{+1\} & t > 0, \\ [-1,1] & t = 0, \\ \{-1\} & t < 0, \end{cases}$$

which is also the subdifferential of $|t|$. For $x \in \mathbb{R}^n$, we define $\mathrm{sgn}(x) \in \mathbb{R}^n$ and $\mathrm{SGN}(x) \subset \mathbb{R}^n$ component-wise as $(\mathrm{sgn}(x))_i := \mathrm{sgn}(x_i)$ and $(\mathrm{SGN}(x))_i := \mathrm{SGN}(x_i)$, $i = 1, 2, \ldots, n$, respectively. Clearly,

$$\mathrm{sgn}(x) = \mathrm{sgn}(x') \iff \mathrm{SGN}(x) = \mathrm{SGN}(x'),\ \forall\ x, x'.$$

For $x, y \in \mathbb{R}^n$, let $x \odot y \in \mathbb{R}^n$ denote the component-wise product of x and y, i.e., $(x \odot y)_i = x_i y_i$. Furthermore, vector operators such as $|x|$ and $\max\{x, y\}$ are defined to operate component-wise as well, analogous to the definitions of sgn and SGN.

For any index set $I \subseteq \{1, 2, \ldots, n\}$ (later, we will use index sets E and L), x_I is defined as the subvector of x of length $|I|$ consisting only of components x_i, $i \in I$. Similarly, if g is a vector-valued function, then $g_I(x)$ denotes the subvector of $g(x)$ consisting of $g_i(x)$, $i \in I$.

This paper is organized as follows. In section 2, we recall the classic optimality (or in general, stationarity) conditions for problem (1.1), and then characterize the optimal solution sets of problems (1.1) and (1.3). In section 3, we present a fixed-point optimality condition for (1.1). This optimality condition motivates a fixed-point algorithm and introduces the shrinkage operator, the properties of which conclude section 3. In section 4, we present our results on the convergence and rates of convergence of the fixed-point algorithm; the proofs of the main results are given in section 5. We motivate and propose a continuation method in section 6, support this proposal with a few numerical results, and briefly discuss some possible extensions. Finally, we conclude the paper in section 7.

2. Optimality and optimal solution sets. Recall that $f(x)$ in (1.1) is convex, and let X^* be the set of optimal solutions of (1.1). It is well known from convex analysis (see, for example, [53]) that an optimality condition for (1.1) is

$$x \in X^* \iff \mathbf{0} \in \mathrm{SGN}(x) + \mu g(x), \tag{2.1}$$

where $\mathbf{0}$ is the zero vector in $\mathbb{R}^n$, or equivalently,

$$x \in X^* \iff \mu g_i(x) \begin{cases} = -1, & x_i > 0, \\ \in [-1,1], & x_i = 0, \\ = 1, & x_i < 0. \end{cases} \tag{2.2}$$

It follows readily from (2.2) that $\mathbf{0}$ is an optimal solution of (1.1) if and only if $\mu\|g(\mathbf{0})\|_\infty \leq 1$; therefore, it is easy to check whether $\mathbf{0}$ is a solution of (1.1).

We note that the solution set X^* may have more than one element. The following theorem establishes some properties of X^* that are of interest in their own right, but will also be useful in later developments.

THEOREM 2.1. *Let $f \in C^2$ be convex and X^* be the set of optimal solutions of* (1.1)*, which is nonempty.*

1. *If $x^1 \in X^*$ and $x^2 \in X^*$, then $g(x^1) = g(x^2)$.*

2. *$x \in X^*$ if and only if $g(x) \equiv g^*$, where for $i = 1, 2, \ldots, n$,*

$$\mu g_i^* \begin{cases} = -1, & \max\{x_i : x \in X^*\} > 0, \\ = +1, & \min\{x_i : x \in X^*\} < 0, \\ \in [-1, 1], & \text{otherwise.} \end{cases} \tag{2.3}$$

3. *X^* is contained in a single orthant of $\mathbb{R}^n$; more precisely*

$$X^* \subset O := \{x \in \mathbb{R}^n : -\mathrm{sgn}^+(g_i^*)x_i \geq 0, \forall i\}, \tag{2.4}$$

where $\mathrm{sgn}^+(\cdot)$ is equal to $\mathrm{sgn}(\cdot)$ except that $\mathrm{sgn}^+(0) := 1$, i.e.,

$$\mathrm{sgn}^+(t) := \begin{cases} +1 & t \geq 0, \\ -1 & t < 0. \end{cases}$$

(In addition, we let $\mathrm{sgn}^+(x)$ be defined component-wise for any $x \in \mathbb{R}^n$.)

Furthermore, if $f(x)$ is the quadratic defined as in (1.6), then

4. *If $x^1 \in X^*$ and $x^2 \in X^*$, then $Ax^1 = Ax^2$.*
5. *$\|x\|_1$ and $\|Ax - b\|_M$ are constant for all $x \in X^*$.*
6. *X^* is a bounded polyhedron, i.e., a polytope.*

Proof. We prove the statements one by one.

1. This part will be proven later as Corollary 4.2 under Assumption 1, which is slightly weaker than what is assumed for this theorem. That proof is independent of this theorem and the results that follow from it.
2. (2.3) follows directly from part 1 and (2.2) applied to all $x \in X^*$.
3. From (2.1) and (2.3), if there exists an $x \in X^*$ with a strictly positive (negative) x_i, then $\mu g_i^* = -1$ ($\mu g_i^* = 1$), so all other $x \in X^*$ must satisfy $x_i \geq 0$ ($x_i \leq 0$). Consequently, X^* lies in the orthant O.
4. From part 1 and for the quadratic $f(x)$ so specified, $g(x^1) - g(x^2) = A^\top MA(x^1 - x^2) = \mathbf{0}$, which immediately implies that $A(x^1 - x^2) = \mathbf{0}$, given that M is symmetric positive definite.
5. From part 4, Ax is constant over $x \in X^*$, and so is $\|Ax - b\|_M$. Since (1.3) has a unique optimal objective value, $\|x\|_1$ must also be constant.
6. Defining $p = -\mathrm{sgn}^+(g^*)$, from the definition of O we have

$$p^\top x = \|x\|_1, \ \forall x \in O.$$

Consider the linear program

$$\min_x \left\{ p^\top x : Ax = c, \ x \in O \right\}, \tag{2.5}$$

where $c = Ax$ for any $x \in X^*$. It is easy to verify that an optimal solution $\bar{x}$ of (2.5) satisfies both $\|\bar{x}\|_1 = \|x\|_1$ and $\|A\bar{x} - b\|_M = \|Ax - b\|_M$ for any $x \in X^*$ and vice versa. So (2.5) is equivalent to (1.3), as long as c and O (or equivalently, g^*) are known. Consequently, X^*, as the solution set of the linear program (2.5), must be a polyhedron and must be bounded since $\|x\|_1$ is constant for all $x \in X^*$.

This completes the proof. ∎

3. A fixed-point algorithm.

3.1. Optimality as a fixed-point equation. We start with another optimality condition for problem (1.1): for any scalar $\tau > 0$, $x^* \in X^*$ if and only if

$$x^* = \operatorname{sgn}(x^* - \tau g(x^*)) \odot \max\left\{|x^* - \tau g(x^*)| - \frac{\tau}{\mu}, \mathbf{0}\right\}. \tag{3.1}$$

The derivation of (3.1) can be found, for example, in [12].

It is straightforward to verify that optimality condition (3.1) can be replaced by

$$x^* = \operatorname{sgn}(x^* - d(x^*) \odot g(x^*)) \odot \max\left\{|x^* - d(x^*) \odot g(x^*)| - \frac{d(x^*)}{\mu}, \mathbf{0}\right\}, \tag{3.2}$$

where the positive scalar τ in (3.1) is replaced by any mapping d from $\mathbb{R}^n$ to $\mathbb{R}^n$ such that $(d(x))_i = d_i(x_i) > 0$. The algorithm based on (3.1), and its analysis as well, can be readily extended to those based on (3.2) (see [26] for a study of such an extension).

The right-hand side of the fixed-point equation (3.1) is a composition of two mappings from $\mathbb{R}^n$ to $\mathbb{R}^n$ defined as

$$h(\cdot) := I(\cdot) - \tau g(\cdot), \tag{3.3}$$

$$s_\nu(\cdot) := \operatorname{sgn}(\cdot) \odot \max\{|\cdot| - \nu, 0\}, \text{ where } \nu > 0. \tag{3.4}$$

Intuitively, $h(\cdot)$ resembles a gradient descent step for $f(x)$ with the stepsize $\tau > 0$, and $s_\nu(\cdot)$ reduces the magnitude of each nonzero component of the input vector by an amount less than or equal to ν, thus reducing the ℓ_1-norm. Later we will also use s_ν as a mapping from $\mathbb{R}$ to $\mathbb{R}$ in composition with $h_i(\cdot) = (h(\cdot))_i$ from $\mathbb{R}^n$ to $\mathbb{R}$.

Equation (3.1) leads to *the fixed-point iterations*

$$x^{k+1} = s_\nu \circ h\left(x^k\right) \text{ with } \tau > 0,\ \nu = \tau/\mu, \tag{3.5}$$

which can be derived by operator-splitting or other means as has been done by the authors mentioned in subsection 1.2. As many others before us, we originally derived the fixed-point scheme (3.5) from a totally different approach, and later found that it can be interpreted as the forward-backward splitting algorithm (1.5) with

$$T_1(x) = \partial\|x\|_1/\mu, \text{ and } T_2(x) = g(x),$$

since simple calculations show that

$$s_\nu = (I + \tau T_1)^{-1}, \text{ and } h = I - \tau T_2.$$

However, some special properties of the operator s_ν, given below, will allow us to obtain strong convergence results that do not directly follow from the existing theory for forward-backward splitting algorithms applied to more general operators.

The main algorithm of the paper, Algorithm 1, is based on (3.5) and will be presented in subsection 6.2 along with choices for τ and μ.

3.2. Properties of the shrinkage operator. It is easy to verify that $s_\nu(y)$ is the unique solution of

$$\min_{x \in \mathbb{R}^n} \nu\|x\|_1 + \frac{1}{2}\|x - y\|^2$$

for any $y \in \mathbb{R}^n$. Wavelet analysts refer to $s_\nu(\cdot)$ as the soft-thresholding [19] or wavelet shrinkage [7] operator. For convenience, we will refer to $s_\nu(\cdot)$ as the *shrinkage operator*.

The two lemmas below establish some useful properties of the shrinkage operator. Both make immediate use of the component-wise separability of s_ν, that is, for all indices i

$$(s_\nu(y))_i = s_\nu(y_i).$$

The alternative representation of s_ν in Lemma 3.1 will be used to prove Lemma 3.2. Lemma 3.2 proves a number of component-wise properties of s_ν, including nonexpansiveness. These results will be used to prove convergence results for (3.5) that are not implied by convergence results in [12].

With a slight abuse of notation, we let $\mathcal{P}(x)$ denote the projection of a vector $x \in \mathbb{R}^k$ onto the k-cube $[-\nu, \nu]^k$ for any positive integer k, since the projection onto any k-cube is done component-wise. Now Lemma 3.1 below can be trivially verified by enumerating all possible cases.

LEMMA 3.1. *The shrinkage operator can be written as*

$$s_\nu(y) = y - \mathcal{P}(y), \quad \forall\ y \in \mathbb{R}^n, \tag{3.6}$$

and the equation holds component-wise; i.e., $(s_\nu(y))_i \equiv s_\nu(y_i) = y_i - \mathcal{P}(y_i)$. *Moreover, both* $s_\nu(y)$ *and* $\mathcal{P}(y)$ *are component-wise monotone.*

LEMMA 3.2. *The operator* $s_\nu(\cdot)$ *is component-wise nonexpansive and for any* $y^1, y^2 \in \mathbb{R}^n$,

$$\left|s_\nu\left(y_i^1\right) - s_\nu\left(y_i^2\right)\right| = \left|y_i^1 - y_i^2\right| - \left|\mathcal{P}\left(y_i^1\right) - \mathcal{P}\left(y_i^2\right)\right|, \ \forall i. \tag{3.7}$$

Consequently, s_ν *is nonexpansive in any* ℓ_p *(quasi-)norm with* $p \geq 0$, *and if* h *is nonexpansive in a given norm, then* $s_\nu \circ h$ *is as well. Moreover, consider the case when*

$$\left|s_\nu\left(y_i^1\right) - s_\nu\left(y_i^2\right)\right| = \left|y_i^1 - y_i^2\right|, \tag{3.8}$$

which we refer to as the no-shrinkage condition. We have, for each index i:

1. (3.8) $\implies \mathcal{P}(y_i^1) = \mathcal{P}(y_i^2)$, $\operatorname{sgn}(y_i^1) = \operatorname{sgn}(y_i^2)$, $s_\nu(y_i^1) - s_\nu(y_i^2) = y_i^1 - y_i^2$.
2. (3.8) *and* $y_i^1 \neq y_i^2 \implies |y_i^1| \geq \nu$, $|y_i^2| \geq \nu$ *and* $|y_i^1| \neq |y_i^2|$.
3. (3.8) *and* $|y_i^2| < \nu \implies |y_i^1| < \nu$, $y_i^1 = y_i^2$, $s_\nu(y_i^1) = s_\nu(y_i^2) = 0$.
4. (3.8) *and* $|y_i^2| \geq \nu \implies |y_i^1| \geq \nu$.
5. $|y_i^2| \geq \nu$ *and* $\operatorname{sgn}(y_i^1) \neq \operatorname{sgn}(y_i^2) \implies |s_\nu(y_i^1) - s_\nu(y_i^2)| \leq |y_i^1 - y_i^2| - \nu$.
6. $s_\nu(y_i^1) \neq 0 = s_\nu(y_i^2) \implies |y_i^1| > \nu$, $|y_i^2| \leq \nu$, $|s_\nu(y_i^1) - s_\nu(y_i^2)| \leq |y_i^1 - y_i^2| - (\nu - |y_i^2|)$.

Proof. For ease of notation, we drop the subscript i and let $p^1 = y_i^1$ and $p^2 = y_i^2$. Without loss of generality we assume that $p^1 \geq p^2$. Hence, from the monotonicity of s_ν and $\mathcal{P}$ we have $s_\nu(p^1) \geq s_\nu(p^2)$ and $\mathcal{P}(p^1) \geq \mathcal{P}(p^2)$. Therefore,

$$\begin{aligned}\left|s_\nu\left(p^1\right) - s_\nu\left(p^2\right)\right| &= s_\nu\left(p^1\right) - s_\nu\left(p^2\right) = p^1 - p^2 - \left(\mathcal{P}\left(p^1\right) - \mathcal{P}\left(p^2\right)\right) \\ &= \left|p^1 - p^2\right| - \left|\mathcal{P}\left(p^1\right) - \mathcal{P}\left(p^2\right)\right|,\end{aligned}$$

which proves (3.7).

Next, we move to proving parts 5 and 6, omitting the proofs for parts 1 through 4 since, given (3.6) and (3.7), they all can be similarly and easily verified.

For part 5, it suffices to show that $|\mathcal{P}(p^1)-\mathcal{P}(p^2)| \geq \nu$. Without loss of generality, we assume that $p^1 > 0$ and $p^2 < 0$. Hence, $\mathcal{P}(p^1) \geq 0$, $\mathcal{P}(p^2) = -\nu$, and $|\mathcal{P}(p^1) - \mathcal{P}(p^2)| = \mathcal{P}(p^1) - \mathcal{P}(p^2) \geq \nu$.

Finally, we verify part 6. First, it is easy to see that $|p^1| > \nu$ and $|p^2| \leq \nu$. Hence, $|\mathcal{P}(p^1)| = \nu$ and $\mathcal{P}(p^2) = p^2$, and $|\mathcal{P}(p^1) - \mathcal{P}(p^2)| \geq |\mathcal{P}(p^1)| - |\mathcal{P}(p^2)| = \nu - |p^2|$. ∎

4. Convergence analysis. In this section, we study the convergence of the fixed-point iterations (3.5) applied to the general ℓ_1-regularized minimization problem (1.1) and the quadratic case (1.3). Assumption 1 below, which states that f is a convex function with bounded Hessian in a neighborhood of an optimal solution of (1.1), is sufficient for our global convergence result and will be applied throughout. Further assumptions (primarily on the rank of a particular minor of the Hessian of f) will be made to obtain linear convergence rate results in section 4.2.

Assumption 1. Problem (1.1) has an optimal solution set $X^* \neq \emptyset$, and there exists a bounded convex set $\Omega \supset X^*$ such that $f \in C^2(\Omega)$, $H(x) := \nabla^2 f(x) \succeq 0$ for $x \in \Omega$ and

$$\hat{\lambda}_{\max} := \max_{x \in \Omega} \lambda_{\max}(H(x)) < \infty. \tag{4.1}$$

For simplicity, we will use a constant parameter τ in the fixed-point iterations (3.5): $x^{k+1} = s_\nu(x^k - \tau g(x^k))$, where

$$\nu = \tau/\mu. \tag{4.2}$$

In particular, we will always choose

$$\tau \in \left(0, 2/\hat{\lambda}_{\max}\right), \tag{4.3}$$

which guarantees that $h(\cdot) = I(\cdot) - \tau g(\cdot)$ is nonexpansive in Ω, and contractive in the range space of H in the quadratic case. Our analysis can be extended to the case of variable τ, but this would require more complicated notation and a reduction of clarity.

4.1. Global and finite convergence. From the mean-value theorem, we recall that for any $x, x' \in \Omega$

$$g(x) - g(x') = \left(\int_0^1 H(x' + t(x - x'))\,\mathrm{d}t\right)(x - x') := \bar{H}(x, x')(x - x'). \tag{4.4}$$

This fact is used to verify the nonexpansiveness of h and the result that noncontraction between any two points under h implies that the gradient of f is equal at those points.

LEMMA 4.1. *Under Assumption* 1 *and the choice of* τ *specified in* (4.3), $h(\cdot) = I(\cdot) - \tau g(\cdot)$ *is nonexpansive in* Ω, *i.e., for any* $x, x' \in \Omega$,

$$\|h(x) - h(x')\| \leq \|x - x'\|. \tag{4.5}$$

Moreover, $g(x) = g(x')$ *whenever equality holds in* (4.5).

Proof. Let $\bar{H} := \bar{H}(x, x')$. We first note that

$$h(x) - h(x') = x - x' - \tau(g(x) - g(x')) = \left(I - \tau\bar{H}\right)(x - x').$$

Hence, in view of (4.3),

$$\begin{aligned}\|h(x)-h(x')\| &= \left\|\left(I-\tau\bar{H}\right)(x-x')\right\| \\ &\le \max\left\{\left|1-\tau\lambda_{\max}\left(\bar{H}\right)\right|,\left|1-\tau\lambda_{\min}\left(\bar{H}\right)\right|\right\}\|x-x'\| \\ &\le \max\{|1-\tau\hat{\lambda}_{\max}|,1\}\|x-x'\| \\ &\le \|x-x'\|.\end{aligned}$$

To prove the second statement, let $s := x - x'$ and $p := \bar{H}^{1/2}s$. Then

$$\begin{aligned}\|h(x)-h(x')\| = \|x-x'\| &\iff \left\|s-\tau\bar{H}s\right\| = \|s\| \\ &\iff -2\tau\, s^T\bar{H}s + \tau^2 s^T\bar{H}^2 s = 0 \\ &\iff \tau p^T\bar{H}p = 2p^Tp \\ &\implies \tau\,\frac{p^T\bar{H}p}{p^Tp} = 2 \text{ if } p\neq 0,\end{aligned}$$

which contradicts (4.3) since $\frac{p^T\bar{H}p}{p^Tp} \le \hat{\lambda}_{\max} < \frac{2}{\tau}$. Hence, $p = 0$ so that

$$g(x) - g(x') = \bar{H}^{1/2}p = 0$$

whenever h is noncontractive. □

Since any two fixed points, say x and x', of the nonexpansive mapping $s_\nu \circ h$ must satisfy the equality

$$\|x-x'\| = \|s_\nu\circ h(x) - s_\nu\circ h(x')\| = \|h(x)-h(x')\|, \tag{4.6}$$

Lemma 4.1 shows that the gradient of f evaluated at any two fixed points must be equal. Hence, we have the following corollary and the first statement of Theorem 2.1.

COROLLARY 4.2 (Constant optimal gradient). *Under Assumption* 1, *there is a vector* $g^* \in \mathbb{R}^n$ *such that*

$$g(x^*) \equiv g^*, \ \ \forall x^* \in X^*. \tag{4.7}$$

We will use the following partition of all indices $\{1,\dots,n\}$ into L and E to obtain finite convergence for components in L and linear convergence for components in E.

DEFINITION 4.3. *Let* $X^* \neq \emptyset$ *be the solution set of* (1.1) *and* g^* *be the vector specified in Corollary* 4.2. *Define*

$$L := \{i : \mu|g_i^*| < 1\} \ \ \textit{and} \ \ E := \{i : \mu|g_i^*| = 1\}. \tag{4.8}$$

It is clear from the optimality condition (2.2) that $L \cup E = \{1,2,\dots,n\}$,

$$\operatorname{supp}(x^*) \subseteq E, \text{ and } x_i^* = 0, \ \ \forall i \in L, \ \forall x^* \in X^*. \tag{4.9}$$

There are examples in which $\operatorname{supp}(x^*) \neq E$, so the two vectors $|x^*|$ and $\mathbf{1} - \mu|g^*|$ are always complementary, but may not be strictly complementary.

The positive scalar ω defined below will also play a key role in the finite convergence property of the fixed-point iterations:

DEFINITION 4.4. *Let* g^* *be the vector specified in Corollary* 4.2. *Define*

$$\omega := \min\{\nu(1-\mu|g_i^*|) : i \in L\}. \tag{4.10}$$

From the definition, clearly $\omega > 0$. In addition, since (4.9) implies that for all $x^* \in X^*$ and all $i \in L$

$$\nu(1 - \mu|g_i^*|) = \nu - \tau|g_i^*| = \nu - |x_i^* - \tau g_i(x^*)| = \nu - |h_i(x^*)|,$$

and consequently, for any $x^* \in X^*$,

$$\min\{\nu - |h_i(x^*)| : i \in L\} = \omega > 0. \tag{4.11}$$

We now claim that Assumption 1 is sufficient for obtaining convergence of the fixed-point iterations (3.5) and finite convergence for components in L and signs in E. We reiterate that under similar conditions, convergence has been established in [12].

THEOREM 4.5 (the general case). *Under Assumption* 1, *the sequence* $\{x^k\}$ *generated by the fixed-point iterations* (3.5) *applied to problem* (1.1) *from any starting point* $x^0 \in \Omega$ *converges to some* $x^* \in X^* \cap \Omega$. *In addition, for all but finitely many iterations, we have*

$$x_i^k = x_i^* = 0, \ \ \forall i \in L, \tag{4.12}$$

$$\operatorname{sgn}\left(h_i\left(x^k\right)\right) = \operatorname{sgn}\left(h_i\left(x^*\right)\right) = -\mu g_i^*, \ \ \forall i \in E, \tag{4.13}$$

where the numbers of iterations not satisfying (4.12) *and* (4.13) *do not exceed* $\|x^0 - x^*\|^2/\omega^2$ *and* $\|x^0 - x^*\|^2/\nu^2$, *respectively, for* ω *defined in* (4.10) *and* $\nu = \tau/\mu$.

The proof of Theorem 4.5 is rather lengthy and is therefore relegated to the next section. A majority of the proof concerns the finite convergence properties which are new. For the sake of completeness, we also include a proof for global convergence which is known.

In light of this theorem, every starting point $x^0 \in \Omega$ determines a converging sequence $\{x^k\}$ whose limit is a solution of (1.1). Generally, the solutions of (1.1) may be nonunique, as it is not difficult to construct simple examples for which different starting points lead to different solutions.

We recall that x_E and g_E^* are defined as the subvectors of x and g^* with components x_i and g_i^*, $i \in E$, respectively. Without loss of generality, we assume $E = \{1, 2, \ldots, |E|\}$, and let $(x_E; \mathbf{0})$ denote the vector in $\mathbb{R}^n$ obtained from x by setting the components x_i $\forall i \in L$ to zero. The following corollary enables one to apply any convergence results for the gradient projection method to the fixed-point iterations (3.5).

COROLLARY 4.6. *Under Assumption* 1 *and starting from some* $x^0 \in \Omega$, *after a finite number of iterations the fixed-point iterations* (3.5) *reduce to gradient projection iterations for minimizing* $\phi(x_E)$ *over a constraint set* O_E, *where*

$$\phi(x_E) := -(g_E^*)^\top x_E + f((x_E; \mathbf{0})), \ \ \textit{and} \tag{4.14}$$

$$O_E = \{x_E \in \mathbb{R}^{|E|} : -\operatorname{sgn}(g_E^*) \odot x_E \geq 0\}. \tag{4.15}$$

Specifically, we have $x^{k+1} = (x_E^{k+1}; \mathbf{0})$ *in which*

$$x_E^{k+1} := P_{O_E}\left(x_E^k - \tau \nabla \phi\left(x_E^k\right)\right), \tag{4.16}$$

where P_{O_E} *is the orthogonal projection onto* O_E, *and* $\nabla\phi(x_E) = -g_E^* + g_E((x_E; \mathbf{0}))$.

Proof. From Theorem 4.5, there exists $K > 0$ such that for $k \geq K$ (4.12)–(4.13) hold. Let $k > K$. Since $x_i^k = 0$ for $i \in L$, it suffices to consider $i \in E$. For

$i \in E$, we have $x_i^k \geq 0$ if $\operatorname{sgn}(h_i(x^{k-1})) = 1$ (equivalently, $g_i^* < 0$) and $x_i^k \leq 0$ if $\operatorname{sgn}(h_i(x^{k-1})) = -1$ ($g_i^* > 0$). Therefore, for any i, $-g_i^* x_i^k \geq 0$ for all $k > K$. Hence, $x^k \in O$ according to the definition (2.4) of O and $x_E^k \in O_E$.

For $i \in E$, we calculate the quantity

$$\begin{aligned} y_i^{k+1} &:= x_i^k - \tau \left(\nabla \phi \left(x^k\right)\right)_i \\ &= x_i^k - \tau \left(-g_i^* + g_i \left(x^k\right)\right) \\ &= h_i \left(x^k\right) + \nu \mu g_i^* \\ &= \operatorname{sgn}\left(h_i \left(x^k\right)\right) \left(\left|h_i \left(x^k\right)\right| - \nu\right), \end{aligned}$$

where (4.13) was used to obtain the last expression. Clearly, the fixed-point iterations (3.5) restricted to the components $i \in E$ are

$$\left(x_E^{k+1}\right)_i = s_\nu \circ h_i \left(x^k\right) = \begin{cases} y_i^{k+1}, & -g_i^* y_i^{k+1} \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Equivalently,

$$\left(x_E^{k+1}\right)_i = \left(P_{O_E} \left(x_E^k - \tau \nabla \phi \left(x_E^k\right)\right)\right)_i,$$

which completes the proof. □

Finally, a stronger global convergence result for convex quadratic functions, namely, f as in (1.6), follows directly from the general convergence result. We note that Assumption 1 is no longer necessary if the convex quadratic is bounded below. Due to the importance of the quadratic case, we state a separate theorem.

THEOREM 4.7 (the quadratic case). *Let f be a convex quadratic function that is bounded below, H be its Hessian, and τ satisfy*

$$0 < \tau < 2/\lambda_{\max}(H). \tag{4.17}$$

Then the sequence $\{x^k\}$, generated by the fixed-point iterations (3.5) *from any starting point x^0, converges to some $x^* \in X^*$. In addition,* (4.12)–(4.13) *hold for all but finitely many iterations.*

4.2. Linear rate of convergence. Let $\{x^k\}$ be generated by the fixed-point iterations (3.5) starting from any $x^0 \in \Omega$. We know that the sequence converges to some point in X^*. Throughout this subsection, we let $x^0 \in \Omega$,

$$x^* := \lim_{k \to \infty} x^k,$$

and study the rate of convergence of $\{x^k\}$ to x^* under different assumptions. Note that $\Omega = \mathbb{R}^n$ if f is convex quadratic, and recall that a sequence $\{\|x^k - x^*\|\}$ converges to zero q-linearly if its q_1-factor is less than one, i.e., if

$$q_1 := \limsup_{k \to \infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} < 1,$$

while it is r-linearly convergent if it is bounded by a q-linearly convergent sequence.

As we will show, under appropriate assumptions q-linear convergence holds for any $\tau \in (0, 2/\hat{\lambda}_{\max})$. However, the q_1-factor may vary with different choices of τ. In particular, we consider choices of the form

$$\tau(\lambda) := \frac{\gamma(\lambda)}{\gamma(\lambda) + 1} \frac{2}{\hat{\lambda}_{\max}}, \quad \gamma(\lambda) := \frac{\hat{\lambda}_{\max}}{\lambda}, \tag{4.18}$$

where $\hat{\lambda}_{\max}$ is defined in (4.1) and $\lambda > 0$ will take different values under different assumptions. It is easy to see that $\tau(\lambda) \in (0, 2/\hat{\lambda}_{\max})$ since $\gamma(\lambda) > 0$.

Some of our assumptions will involve the matrix H_{EE}:

DEFINITION 4.8. *Let $H(x)$ denote the Hessian of f evaluated at $x \in \Omega$, and*

$$H_{EE}(x) := [H_{i,j}(x)]_{i,j \in E}$$

denote the square submatrix of H corresponding to the index set E defined in (4.8).

Based on Corollary 4.6, we first apply existing convergence results for the gradient projection method applied to (4.16).

PROPOSITION 4.9. *Let Assumption* 1 *hold. If* (i) *$H_{EE}(x^*)$ has full rank, or* (ii) *f is defined as in* (1.6)*, then $\{\|x^k\|_1 + \mu f(x^k)\}$ converges to $\|x^*\| + \mu f(x^*)$ q-linearly and $\{x^k\}$ converges to x^* r-linearly.*

Under the first condition, the above result follows from [50] and [41], while under the second condition it follows from [41]. However, by directly analyzing the original fixed-point iterations, we can strengthen the convergence rate of $\{x^k\}$ from r-linear to q-linear. Theorem 4.10 does this under the assumption that $H_{EE}(x^*)$ is full rank; Theorem 4.11 instead assumes that $\operatorname{supp}(x^*) = E$. We first define

$$\bar{H}^k \equiv \bar{H}\left(x^k, x^*\right) := \int_0^1 H\left(x^* + t\left(x^k - x^*\right)\right) dt. \tag{4.19}$$

THEOREM 4.10. *Let Assumption* 1 *hold, and assume that*

$$\lambda^E_{\min} := \lambda_{\min}(H_{EE}(x^*)) > 0. \tag{4.20}$$

Then for any $\tau \in (0, 2/\hat{\lambda}_{\max})$, $\{x^k\}$ converges to x^* q-linearly. Moreover, if τ is chosen as in (4.18) *with $\lambda = \lambda^E_{\min}$, then the q_1-factor satisfies*

$$q_1 \leq \frac{\gamma\left(\lambda^E_{\min}\right) - 1}{\gamma\left(\lambda^E_{\min}\right) + 1}. \tag{4.21}$$

Proof. Without loss of generality, we assume that all iteration counts, k, are large enough so that $x^k_i = x^*_i = 0$ for all $i \in L$, and that the spectrum of $\bar{H}^k_{EE}$ falls in the interval $[\lambda^E_{\min} - \epsilon, \hat{\lambda}_{\max}]$ for an arbitrary $\epsilon > 0$. (The first assumption on k is valid because of the finite convergence properties of Theorem 4.5; the second follows from the continuity of the Hessian.) Since $x^k_i = x^*_i = 0$, $\forall i \in L$, the mean-value theorem yields

$$h_E\left(x^k\right) - h_E\left(x^*\right) = x^k_E - x^*_E - \tau\left(g_E\left(x^k\right) - g_E\left(x^*\right)\right) = \left(I - \tau \bar{H}^k_{EE}\right)\left(x^k_E - x^*_E\right).$$

Recall that $x^{k+1} = s_\nu \circ h(x^k)$ and $s_\nu(\cdot)$ is nonexpansive. Hence,

$$\begin{aligned}
\left\|x^{k+1} - x^*\right\| &\equiv \left\|x^{k+1}_E - x^*_E\right\| \\
&\leq \left\|h_E\left(x^k_E\right) - h_E\left(x^*_E\right)\right\| \\
&\leq \left\|I - \tau \bar{H}^k_{EE}\right\| \left\|x^k_E - x^*_E\right\| \\
&\leq \max\left\{\left|1 - \tau\hat{\lambda}_{\max}\right|, \left|1 - \tau\lambda^E_{\min}\right| + \tau\epsilon\right\} \left\|x^k_E - x^*_E\right\| \\
&\equiv \max\left\{\left|1 - \tau\hat{\lambda}_{\max}\right|, \left|1 - \tau\lambda^E_{\min}\right| + \tau\epsilon\right\} \left\|x^k - x^*\right\|.
\end{aligned}$$

Clearly, $\max\{|1-\tau\hat{\lambda}_{\max}|, |1-\tau\lambda^E_{\min}|+\tau\epsilon\}$ is less than one for any $\tau \in (0, 2/\hat{\lambda}_{\max})$ and ϵ sufficiently small; in particular, it equals the right-hand side of (4.21) plus $\tau\epsilon$ when $\tau = \tau(\lambda^E_{\min})$. Since ϵ is arbitrary, (4.21) must hold. ∎

THEOREM 4.11. *Let Assumption* 1 *hold, and also assume that* x^* *satisfies* (i) $\mathrm{supp}(x^*) = E$ *or, equivalently, the strict complementarity condition*

$$|x^*| + (1-\mu|g^*|) > 0, \tag{4.22}$$

and (ii) *the range space* $\mathcal{R}(H_{EE}(x))$ *of* $H_{EE}(x)$ *is invariant in a neighborhood* N^* *of* x^*. *Whenever* $H_{EE}(x^*) \neq \mathbf{0}$, *let*

$$\lambda^{\mathcal{R}}_{\min} := \lambda_{\min}\left(V^\top H_{EE}(x^*)V\right) > 0, \tag{4.23}$$

where V *is any orthonormal basis of* $\mathcal{R}(H_{EE}(x^*))$.

If $H_{EE}(x^*) = \mathbf{0}$, *then* $x^k = x^*$ *for all* k *sufficiently large; otherwise* $\{x^k\}$ *converges to* x^* *q-linearly for any* $\tau \in (0, 2/\hat{\lambda}_{\max})$. *In the latter case, if* τ *is chosen as in* (4.18) *with* $\lambda = \lambda^{\mathcal{R}}_{\min}$, *then the* q_1*-factor satisfies*

$$q_1 \leq \frac{\gamma\left(\lambda^{\mathcal{R}}_{\min}\right)-1}{\gamma\left(\lambda^{\mathcal{R}}_{\min}\right)+1}. \tag{4.24}$$

The proof of this theorem is given in section 5.2. We note that $\mathcal{R}(H_{EE}(x))$ is invariant near x^* if either f is a quadratic function or $H_{EE}(x^*)$ has full rank.

Since Assumption 1 is not required in the proof of global convergence for convex quadratic f, we can directly derive the following results for this case, which is the situation one encounters with compressed sensing. The proof, which is similar to those of Theorems 4.10 and 4.11, is left to the reader.

COROLLARY 4.12. *Let* f *be a convex quadratic function that is bounded below, and* $\{x^k\}$ *be the sequence generated by the fixed-point iterations* (3.5) *with* $\tau \in (0, 2/\lambda_{\max}(H))$.

1. *If* H_{EE} *has full rank, then* $\{x^k\}$ *converges to* x^* *q-linearly. Moreover, if* τ *is chosen as in* (4.18) *with* $\lambda = \lambda_{\min}(H_{EE})$, *then the* q_1*-factor satisfies*

$$q_1 \leq \frac{\gamma(\lambda_{\min}(H_{EE}))-1}{\gamma(\lambda_{\min}(H_{EE}))+1}.$$

2. *Let* x^* *satisfy the strict complementarity condition* (4.22). *Then if* $H_{EE} = \mathbf{0}$, $\{x^k\}$ *converges to* x^* *in a finite number of steps. Otherwise* $\{x^k\}$ *converges to* x^* *q-linearly, and if* τ *is chosen as in* (4.18) *with* $\lambda := \lambda_{\min}(V^\top H_{EE}V)$, *where* V *is an orthonormal basis for the range space of* H_{EE}, *then the* q_1*-factor satisfies*

$$q_1 \leq \frac{\gamma\left(\lambda_{\min}\left(V^\top H_{EE}V\right)\right)-1}{\gamma\left(\lambda_{\min}\left(V^\top H_{EE}V\right)\right)+1}.$$

4.3. Discussion. The assumptions of Theorems 4.10 and 4.11 usually hold for compressed sensing reconstruction problems posed as in (1.3) or (1.2), in which case A is often a Gaussian random matrix or has rows randomly chosen from an orthogonal matrix such as an FFT, DCT, or wavelets transform matrix. It is well known that a randomly generated matrix is full rank with probability one (unless elements of the matrix are generated from a restricted space) [23]. Therefore, when $A \in \mathbb{R}^{m\times n}$

is a random matrix, the reduced Hessian for problem (1.3), i.e., $A_E^\top M A_E$, where A_E consists of columns of A with indices in E, will have full rank with probability one as long as $|E| \leq m$, which is generally the case. A similar argument can be made for partial orthogonal matrices. We believe that the strict complementarity assumption in Theorem 4.11 should also hold for random matrices with a prevailing probability, though we do not currently have a proof for this. We have observed the general convergence behavior predicted by our theorems empirically in computational studies; see section 6.2.

In our convergence theorems, the choice of τ is restricted by the upper bound $2/\hat{\lambda}_{\max}$, where $\hat{\lambda}_{\max}$ is an upper bound for the largest eigenvalue of the Hessian. In compressed sensing applications, the quantity $\hat{\lambda}_{\max}$ is often easily obtained when $M = I$. When A is a partial orthogonal matrix, $\hat{\lambda}_{\max} = \lambda_{\max}(A^\top A) = 1$ and $\tau \in (0, 2)$ will suffice. When A is a Gaussian random matrix (with elements independently drawn from the standard normal distribution), well-known random matrix theory (see [34] or [23], for example) yields

$$n\left(1 - \sqrt{\frac{m}{n}}\right)^2 \leq \lambda_i\left(A^\top A\right) \leq n\left(1 + \sqrt{\frac{m}{n}}\right)^2$$

with prevailing probability for large n. In either case, upper bounding τ is not an issue as long as $M = I$.

For simplicity, we have used a fixed $\tau \in (0, 2/\hat{\lambda}_{\max})$ in our analysis. However, this requirement could be relaxed in the later stages of the iterations when the actions of the mapping $h = I - \tau g$ concentrate on a "reduced space." In this stage, h can remain contractive even if the maximum eigenvalue bound on the Hessian is replaced by that on the reduced Hessian, which will generally increase the upper bound on τ. For example, consider the quadratic problem (1.2) where A is a partial orthogonal matrix. Then $\lambda_{\max}(A^\top A) = 1$, but $\lambda_{\max}(A_E^\top A_E) < 1$, such that h remains contractive even if τ is chosen close to $2/\lambda_{\max}(A_E^\top A_E) > 2$. Such a dynamic strategy, though theoretically feasible, is not straightforward to implement. It should be an interesting topic for further research.

5. Proofs of convergence results. In this section, Theorems 4.5 and 4.11 are proved through several technical results that lead to the final arguments.

5.1. Proof of Theorem 4.5. The lemma below establishes a sufficient condition for $x \in \Omega$ to be a fixed point of $s_\nu \circ h(\cdot)$.

LEMMA 5.1. *Under Assumption* 1, *if*

$$\|s_\nu \circ h(x) - s_\nu \circ h(x^*)\| \equiv \|s_\nu \circ h(x) - x^*\| = \|x - x^*\|, \tag{5.1}$$

then x is a fixed point, and therefore a solution of (1.1)*; that is,*

$$x = s_\nu \circ h(x). \tag{5.2}$$

Proof. Recall that s_ν is component-wise nonexpansive and h is nonexpansive in $\|\cdot\|$. From (5.1),

$$\|x - x^*\| = \|s_\nu \circ h(x) - s_\nu \circ h(x^*)\| \leq \|h(x) - h(x^*)\| \leq \|x - x^*\|. \tag{5.3}$$

Hence, both inequalities hold as equalities. In particular, the no-shrinkage condition (3.8) holds for $y^1 = h(x)$ and $y^2 = h(x^*)$, so Part 1 of Lemma 3.2 yields

$$s_\nu \circ h(x) - s_\nu \circ h(x^*) = h(x) - h(x^*).$$

Rewriting this equation, we get

$$s_\nu \circ h(x) = x - \tau(g(x) - g(x^*)),$$

and since the last inequality in (5.3) also holds as equality, we have $g(x) - g(x^*) = 0$ according to Lemma 4.1, and hence the conclusion. ∎

The next lemma establishes the finite convergence properties stated in Theorem 4.5.

LEMMA 5.2. *Let Assumption* 1 *hold and* $\{x^k\}$ *be generated by the fixed-point iterations* (3.5) *starting from any* $x^0 \in \Omega$. *Then*

1. $x_i^k = 0\ \forall i \in L$ *for all but at most* $\|x^0 - x^*\|^2/\omega^2$ *iterations;*
2. $\operatorname{sgn}(h_i(x^k)) = \operatorname{sgn}(h_i(x^*)) = -\mu g_i^*,\ \forall\ i \in E$, *for all but at most* $\|x^0 - x^*\|^2/\nu^2$ *iterations.*

Proof. We fix any $x^* \in X^*$ and consider $x_i^k \neq 0$ for some $i \in L$. In view of the nonexpansiveness of $s_\nu(\cdot)$ and the related property in Lemma 3.2 part 6, we have

$$\begin{aligned}\left|x_i^{k+1} - x_i^*\right|^2 &= \left|s_\nu \circ h_i\left(x^k\right) - s_\nu \circ h_i(x^*)\right|^2 \\ &\leq \left(\left|h_i\left(x^k\right) - h_i(x^*)\right| - (\nu - h_i(x^*))\right)^2 \\ &\leq \left|h_i\left(x^k\right) - h_i(x^*)\right|^2 - \omega^2,\end{aligned}$$

where the last inequality follows from (4.11). The component-wise nonexpansiveness of $s_\nu(\cdot)$ and the nonexpansiveness of $h(\cdot)$ imply that

$$\left\|x^{k+1} - x^*\right\|^2 \leq \left\|h\left(x^k\right) - h(x^*)\right\|^2 - \omega^2 \leq \left\|x^k - x^*\right\|^2 - \omega^2.$$

Therefore, the number of iterations where $x_i^k \neq 0$ for some $i \in L$ cannot be more than $\|x^0 - x^*\|^2/\omega^2$. This proves the first statement.

For the second statement, we recall (3.1) and note that if $i \in \operatorname{supp}(x^*)$

$$0 \neq x_i^* = \operatorname{sgn}(h_i(x^*)) \max\{|h_i(x^*)| - \nu, 0\},$$

so that $|h_i(x^*)| > \nu$ for $i \in \operatorname{supp}(x^*)$. On the other hand,

$$|h_i(x^*)| = \tau|g^*| = \tau/\mu = \nu,\ \forall i \in E \setminus \operatorname{supp}(x^*).$$

Therefore,

$$|h_i(x^*)| \geq \nu,\ \forall i \in E.$$

Now if $\operatorname{sgn}(h_i(x^k)) \neq \operatorname{sgn}(h_i(x^*))$ for some $i \in E$, then Lemma 3.2, Part 5 implies

$$\begin{aligned}\left|x_i^{k+1} - x_i^*\right|^2 &= \left|s_\nu \circ h_i\left(x^k\right) - s_\nu \circ h_i(x^*)\right|^2 \\ &\leq \left(\left|h_i\left(x^k\right) - h_i(x^*)\right| - \nu\right)^2 \\ &\leq \left|h_i\left(x^k\right) - h_i(x^*)\right|^2 - \nu^2.\end{aligned}$$

Hence, the number of iterations for which $\operatorname{sgn}(h_i(x^k)) \neq \operatorname{sgn}(h_i(x^*))$ for some $i \in E$ cannot be more than $\|x^0 - x^*\|^2/\nu^2$. Moreover, it follows directly from the definitions of E in (4.8), h in (3.3), and g^* in (2.3), and the equation $\tau = \nu\mu$, that $\operatorname{sgn}(h_i(x^*)) = -\mu g_i^*$ for all $i \in E$. ∎

Based on these lemmas, we provide a short proof of Theorem 4.5 for the sake of completeness.

Proof of Theorem 4.5. To show that $\{x^k\}$ converges, we (i) show that $\{x^k\}$ has a limit point, (ii) argue that it must be a fixed point because it satisfies the condition (5.1) of Lemma 5.1, and (iii) prove its uniqueness.

Since $s_\nu \circ h(\cdot)$ is nonexpansive, $\{x^k\}$ lies in a compact subset of Ω and must have a limit point, say,

$$\bar{x} = \lim_{j\to\infty} x^{k_j}.$$

Since for any given fixed point x^* the sequence $\{\|x^k - x^*\|\}$ is monotonically nonincreasing, it has a limit which can be written as

$$\lim_{k\to\infty} \left\|x^k - x^*\right\| = \|\bar{x} - x^*\|, \tag{5.4}$$

where $\bar{x}$ can be any limit point of $\{x^k\}$. That is, all limit points, if more than one exists, must have an equal distance to any given fixed point $x^* \in X^*$.

By the continuity of $s_\nu \circ h(\cdot)$, the image of $\bar{x}$,

$$s_\nu \circ h(\bar{x}) = \lim_{j\to\infty} s_\nu \circ h\left(x^{k_j}\right) = \lim_{j\to\infty} x^{k_j+1},$$

is also a limit point of $\{x^k\}$. Therefore, from (5.4) we have

$$\|s_\nu \circ h(\bar{x}) - s_\nu \circ h(x^*)\| = \|\bar{x} - x^*\|,$$

which allows us to apply Lemma 5.1 to $\bar{x}$ and establish the optimality of $\bar{x}$.

By setting $x^* = \bar{x} \in X^*$ in (5.4), we establish the convergence of $\{x^k\}$ to its unique limit point $\bar{x}$:

$$\lim_{k\to\infty} \left\|x^k - \bar{x}\right\| = 0.$$

Finally, the finite convergence results (4.12)–(4.13) were proved in Lemma 5.2. ∎

5.2. Proof of Theorem 4.11. The next lemma gives a useful update formula for k sufficiently large and $i \in \operatorname{supp}(x^*)$.

LEMMA 5.3. *Under Assumption* 1, *after a finite number of iterations*

$$x_i^{k+1} = x_i^k - \tau\left(g_i\left(x^k\right) - g_i^*\right), \quad \forall i \in \operatorname{supp}(x^*). \tag{5.5}$$

Proof. Since $x^k \to x^* \in X^*$ and $h(\cdot)$ is component-wise continuous, $h_i(x^k) \to h_i(x^*)$. The fact that $|h_i(x^*)| > \nu$ for $i \in \operatorname{supp}(x^*)$ implies that after a finite number of iterations we have $|h_i(x^k)| > \nu$ for $i \in \operatorname{supp}(x^*)$. This gives

$$\begin{aligned} x_i^{k+1} &= \operatorname{sgn}\left(h_i\left(x^k\right)\right)\left(\left|h_i\left(x^k\right)\right| - \nu\right) \\ &= h_i\left(x^k\right) - \nu \operatorname{sgn}\left(h_i\left(x^k\right)\right) \\ &= x_i^k - \tau g_i\left(x^k\right) - (\tau/\mu)\left(-\mu g_i^*\right) \\ &= x_i^k - \tau\left(g_i\left(x^k\right) - g_i^*\right), \end{aligned}$$

for any $i \in \operatorname{supp}(x^*)$. ∎

Proof of Theorem 4.11. Without loss of generality, we can assume that k is large enough so that (5.5) holds and $x^k \in N^*$, where N^* is defined in Theorem 4.11.

Since $x_i^k = 0$ for any $i \in L$, it suffices to consider the rate of convergence of x_i^k for $i \in E = \text{supp}(x^*)$, where equality follows from the strict complementarity assumption on x^*.

Let $\bar{H}^k$ be defined as in (4.19). By assumption, the range and null spaces of $\bar{H}_{EE}^k$ are now invariant for all k. Let $P = VV^\top \in \mathbb{R}^{|E|\times|E|}$ be the orthogonal projection onto the range space of $H_{EE}(x^*)$ such that $I-P$ is the orthogonal projection onto the null space of $H_{EE}(x^*)$. Also recall that x_E denotes the subvector of x corresponding to the index set E.

Since $E = \text{supp}(x^*)$, Lemma 5.3 implies that

$$x_E^{k+1} = x_E^k - \tau \left(g\left(x^k\right) - g(x^*)\right)_E = x_E^k - \tau \bar{H}_{EE}^k \left(x_E^k - x_E^*\right). \tag{5.6}$$

At each iteration, the update, $-\tau \bar{H}_{EE}^k(x_E^k - x_E^*)$, stays in the range space of $H_{EE}(x^*)$. This implies that the null space components of the iterates have converged to the null space components of x^*, namely, for all k sufficiently large,

$$(I-P)\left(x_E^k - x_E^*\right) \equiv \mathbf{0}. \tag{5.7}$$

If $H_{EE}(x^*) = 0$, then the range space is empty and the update vanishes such that $x^k = x^*$ after a finite number of steps.

Now assume that $H_{EE}(x^*) \neq 0$ so that $\lambda_{\min}^{\mathcal{R}} > 0$ exists. It suffices to consider the rate of convergence of $\{Px_E^k\}$ to Px_E^*. It follows from (5.6) and (5.7) that

$$x_E^{k+1} - x_E^* = P\left(x_E^{k+1} - x_E^*\right) = P\left(I - \tau \bar{H}_{EE}^k\right) P\left(x_E^k - x_E^*\right). \tag{5.8}$$

By a routine continuity argument, we know that there exists an arbitrarily small constant $\epsilon > 0$ such that for all k sufficiently large the eigenvalues of $V^\top H_{EE}^k V$ satisfy

$$\hat{\lambda}_{\max} \geq \lambda_i\left(V^\top H_{EE}^k V\right) \geq \lambda_{\min}^{\mathcal{R}} - \epsilon > 0, \quad \forall i.$$

Consequently, given the definition of τ in (4.18) and noting that $P^2 = P = VV^\top$, we calculate from (5.8):

$$\begin{aligned}
\left\|x_E^{k+1} - x_E^*\right\| &\leq \left\|P\left(I - \tau \bar{H}_{EE}^k\right) P\right\| \left\|x_E^k - x_E^*\right\| \\
&= \left\|I - \tau V^\top \bar{H}_{EE}^k V\right\| \left\|x_E^k - x_E^*\right\| \\
&= \max\left\{\left|1 - \tau\hat{\lambda}_{\max}\right|, \left|1 - \tau\lambda_{\min}^{\mathcal{R}}\right| + \tau\epsilon\right\} \left\|x_E^k - x_E^*\right\| \\
&= \left(\frac{\gamma\left(\lambda_{\min}^{\mathcal{R}}\right) - 1}{\gamma\left(\lambda_{\min}^{\mathcal{R}}\right) + 1} + \tau\epsilon\right) \left\|x_E^k - x_E^*\right\|,
\end{aligned} \tag{5.9}$$

which implies (4.24) since ϵ can be arbitrarily small. □

6. A continuation method. Our algorithm for solving (1.1), that is,

$$\min_{x\in\mathbb{R}^n} \|x\|_1 + \mu f(x), \tag{6.1}$$

consists of applying the fixed-point iterations

$$x^{k+1} = s_\nu \circ h\left(x^k\right) := \text{sgn}\left(x^k - \tau g\left(x^k\right)\right) \odot \max\left\{\left|x^k - \tau g\left(x^k\right)\right| - \nu, 0\right\}, \ \mu\nu = \tau$$

(see (3.5) and (4.3)) within the continuation (or path-following) framework described below. Further extensions that may improve our algorithm are certainly possible, but are beyond the scope of this paper.

6.1. Homotopy algorithms in statistics. Statisticians often solve (1.2) (which is (1.1) with $f(x) = \frac{1}{2}\|Ax - b\|^2$) in the context of regression. In Bayesian terminology, this corresponds to maximizing the *a posteriori* probability for recovering the signal x from the measurement $b = Ax + \epsilon$, where the prior on x is Laplacian and ϵ is Gaussian white noise. Practically, such a procedure may be preferred over standard least squares because a sparse solution of (1.2) explicitly identifies the most significant regressor variables.

As intimated in the Introduction, variations on (1.2) may be used in different applications and contexts. For example, problem (1.2) is closely related to this quadratically constrained ℓ_1-minimization problem

$$\min_x \left\{ \|x\|_1 \;\middle|\; \|Ax - b\|^2 \le \sigma^2 \chi^2_{1-\alpha,m} \right\}, \tag{6.2}$$

which is often used when an estimated noise level σ is available. Alternatively, one can constrain the size of $\|x\|_1$ and minimize the sum of squares of the residual $Ax - b$:

$$\min_x \left\{ \frac{1}{2}\|Ax - b\|^2 \;\middle|\; \|x\|_1 \le t \right\}. \tag{6.3}$$

Statisticians often refer to the above problem as the Least Absolute Shrinkage and Selection Operator (LASSO) [58].

Problems (1.2), (6.2), and (6.3) are equivalent in the sense that once the value of one of μ, σ, or t is fixed, there are values for the other two quantities such that all three problems have the same solution. For a detailed explanation, please see [53].

Least Angle Regression (LARS) (see [24], for example) is a method for solving (6.3). LARS starts with the zero vector and gradually increases the number of nonzeros in the approximation x. In fact, it generates the full path of solutions that results from setting the right-hand side of the constraint to every value in the interval $[0, t]$. Thus, LARS is a homotopy algorithm. The construction of the path of solutions is facilitated by the fact that it is piecewise linear, such that any segment can be generated given the solutions at turning points, which are the points at which at least one component changes from zero to nonzero or vice versa. Thus LARS and other homotopy algorithms [43, 49, 63] solve (6.3) by computing the solutions at the turning points encountered as t increases from 0 to a given value. These algorithms require the solution of a least squares problem at every iteration, where the derivative matrix of the residuals consists of the columns of A associated with the nonzero components of the current iterate. For large-scale problems, solving these intermediate least squares problems may prove costly, especially when the solution is only moderately sparse, and/or A is a partial fast transform matrix that is not stored explicitly.

We found it helpful for our algorithm to adopt a continuation strategy similar to homotopy in the sense that we solve (1.1) for an increasing sequence of μ values. However, our algorithm does not track turning points or solve any least squares subproblems, and so only approximately follows the solution path.

6.2. A continuation strategy. The convergence analysis indicates that the speed of the fixed-point algorithm is determined by the values of $\nu = \tau/\mu$ and ω (see Theorem 4.5), and the spectral properties of the Hessian of $f(x)$ (see Theorems 4.10 and 4.11). The signs of $h_i(x^k)$ evolve to agree with those of $h_i(x^*)$ for $i \in E$ faster for larger ν (equivalently, for smaller μ). Similarly, large ω implies fast convergence of the $|x_i^k|$, $i \in L$, to zero. Once the finite convergence properties of Theorem 4.5 are satisfied, all action is directed towards reducing the errors in the E components, and the (worst-

case) convergence rate is dictated by $\|I - \tau \bar{H}_{EE}\|$, which can be considerably smaller than $\|I - \tau \bar{H}\|$, especially when $|E| \ll n$.

In general, we have little or no control over the value of ω, nor the spectral properties of the Hessian. On the other hand, we do have the freedom to choose τ and $\nu = \tau/\mu$. For fixed τ we found $\tau \in [1/\hat{\lambda}_{\max}, 2/\hat{\lambda}_{\max})$ to be superior to $\tau \in (0, 1/\hat{\lambda}_{\max})$. Beyond this, τ does not have much effect on ν and can be chosen empirically or based on considerations concerning $||I - \tau \bar{H}_{EE}||$. The value of μ, on the other hand, while it must eventually be equal to some specified value $\bar{\mu}$, can in the meantime be chosen freely to produce a wide range of ν values. Thus, since larger ν means faster convergence, we propose a continuation strategy for μ. In particular, if problem (6.1) is to be solved with $\bar{\mu}$, we propose solving a sequence of problems (6.1) defined by an increasing sequence $\{\mu_j\}$, as opposed to fixing $\nu = \tau/\bar{\mu}$. When a new problem, associated with μ_{j+1}, is to be solved, the approximate solution for the current (μ_j) problem is used as the starting point. In essence, this framework approximately follows the path $x^*(\mu)$ in the interval $[\mu_1, \bar{\mu}]$, where for any given μ value $x^*(\mu)$ is an optimal solution for (6.1). This path is well defined if the solution to (6.1) is unique for $\mu \in [\mu_1, \bar{\mu}]$. Even if this is not the case, it is reassuring to observe that the algorithm itself is well-defined. A formal statement of our fixed-point continuation method is given in Algorithm 1.

ALGORITHM 1 Fixed-point Continuation (FPC) Algorithm

Require: A, b, x^0, and $\bar{\mu}$

Select $0 < \mu_1 < \mu_2 < \cdots < \mu_L = \bar{\mu}$. Set $x = x^0$.
for $\mu = \mu_1, \mu_2, \ldots, \mu_L$ **do**
  while "not converged" **do**
    Select τ and set $\nu = \tau/\mu$
    $x \leftarrow s_\nu \circ h(x)$
  end while
end for

Our computational experience indicates that the performance of this continuation strategy can be far superior to that of directly applying the fixed-point iterations (3.5) with a specified value $\bar{\mu}$. This is evident in Figure 6.1, where the convergence behavior

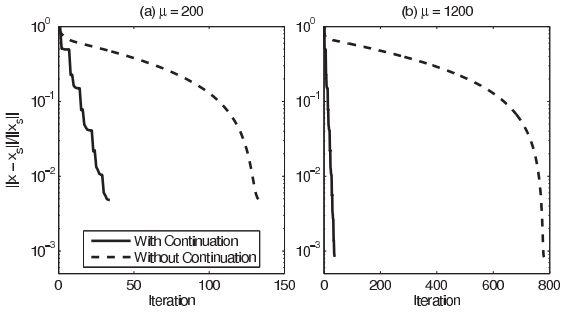


FIG. 6.1. *Convergence acceleration via continuation. The relative error, $\|x^k - x^*\|/\|x^*\|$, and iteration data were obtained by applying the FPC algorithm with or without continuation to two instances of problem* (1.2)*, where in each instance A is a 512×1024 partial DCT matrix, $b = Ax^*$ plus noise for a given sparse x^*, and $\mu = 200$ in the first case and 1200 in the second. The plots show that as μ increases, the advantages of continuation become more pronounced.*

of the two approaches (with and without continuation) is plotted for two values of $\bar{\mu}$, and is in line with the observations of [18, 43, 49, 58]. Moreover, since $x^*(\mu)$ tends to be sparser for smaller μ, the reduced Hessian $\bar{H}_{EE}$ tends to be smaller and better conditioned in this case, such that the continuation strategy should improve the convergence rate for the components with indices in E in addition to the rate of finite convergence of $\mathrm{sgn}(h_i(x^k))$, $i \in E$. Overall, this fixed-point continuation algorithm produces competitive solution times (as compared to other state-of-the-art algorithms) for compressed sensing problems. For instance, we were able to reconstruct signals of length 2,097,152 in 2.5 to 7.5 minutes, depending on the level of sparsity and noise, when A was a $1{,}048{,}576 \times 2{,}097{,}152$ partial DCT matrix (using Matlab 7.3 on a Dell Optiplex GX620 with a 3.2 GHz processor and 4 GB RAM).

In principle, our fixed-point continuation algorithm can be used to solve problems (6.2) and (6.3) in addition to (6.1). Take the LASSO problem (6.3) as an example. When we start our algorithm with a small μ value, the corresponding optimal $\|x\|_1$ is also small; subsequent increases in μ correspond to increases in the optimal $\|x\|_1$. We can stop the process once $\|x\|_1$ approximately equals t, backtracking if necessary. As interesting as such extensions may be, they are not in the scope of the current paper. Indeed, as we observed in our computational study [32], a strength of this algorithmic framework is that a simple implementation is sufficient to obtain good results.

7. Conclusions. We investigated the use of the forward-backward operator splitting technique, combined with a continuation (path-following) strategy, for solving ℓ_1-norm regularized convex optimization problems. Our theoretical analysis yields convergence results stronger than what could be obtained from applying existing general theory to our setting. In particular, we established finite convergence for some quantities and q-linear convergence rates without assuming strict convexity. Interestingly, our rate of convergence results imply, in a general sense, that sparser solutions correspond to faster rates of convergence, which agrees with what has been observed in practice. Our convergence analysis, however, is only for the fixed-point algorithm (3.5) with a fixed μ value. It remains an important, yet more difficult, research issue to study convergence behavior associated with specific continuation strategies.

We have conducted a comprehensive computational study to compare our fixed-point continuation (FPC) algorithm with three recent state-of-the-art compressed sensing recovery algorithms [17, 28, 35]. The numerical results, too lengthy to be included in the present paper, will be reported in a subsequent paper [32]. In brief, these numerical results indicate that FPC's overall performance is competitive with, and is often superior to, these state-of-the-art algorithms. The strong performance of FPC in computing sparse solutions to compressed sensing problems is certainly encouraging. However, it remains a research issue to carefully evaluate and possibly enhance the performance of FPC on other ℓ_1-regularized optimization problems where solutions are not necessarily very sparse.

REFERENCES

[1] D. Baron, M. Wakin, M. Duarte, S. Sarvotham, and R. Baraniuk, *Distributed compressed sensing*, ECE Department, Rice University, preprint, 2005; also available online from http://www.dsp.ece.rice.edu/cs/DCS112005.pdf.

[2] J. Bect, L. Blanc-Feraud, G. Aubert, and A. Chambolle, *A ℓ_1-unified variational framework for image restoration*, European Conference on Computer Vision, Prague, Lecture Notes in Computer Sciences 3024, 2004, pp. 1–13.

[3] J. Bioucas-Dias and M. Figueiredo, *Two-step algorithms for linear inverse problems with non-quadratic regularization*, IEEE International Conference on Image Processing C ICIP' 2007, San Antonio, TX, 2007.
[4] E. Candès, J. Romberg, and T. Tao, *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Inform. Theory, 52 (2006), pp. 489–509.
[5] E. Candès and J. Romberg, *Quantitative robust uncertainty principles and optimally sparse decompositions*, Found. Comput. Math., 6 (2006), pp. 227–254.
[6] E. Candès and T. Tao, *Near optimal signal recovery from random projections: Universal encoding strategies*, IEEE Trans. Inform. Theory, 52 (2006), pp. 5406–5425.
[7] A. Chambolle, R. A. DeVore, N.-Y. Lee, and B. J. Lucier, *Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage*, IEEE Trans. Image Process., 7 (1998), pp. 319–335.
[8] H.-G. Chen and R. T. Rockafellar, *Convergence rates in forward-backward splitting*, SIAM J. Optim., 7 (1997), pp. 421–444.
[9] S. Chen, D. Donoho, and M. A. Saunders, *Atomic decomposition by basis pursuit*, SIAM J. Sci. Comput., 20 (1998), pp. 33–61.
[10] J. Claerbout and F. Muir, *Robust modelling of erratic data*, Geophys., 38 (1973), pp. 826–844.
[11] P. L. Combettes and J.-C. Pesquet, *Proximal thresholding algorithm for minimization over orthonormal bases*, SIAM J. Optim., 18 (2007), pp. 1351–1376.
[12] P. L. Combettes and V. R. Wajs, *Signal recovery by proximal forward-backward splitting*, SIAM J. Multiscale Model. Simul., 4 (2005), pp. 1168–1200.
[13] I. Daubechies, M. Defrise, and C. De Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Commun. Pure Appl. Math., 57 (2004), pp. 1413–1457.
[14] I. Daubechies, M. Fornasier, and I. Loris, *Accelerated projected gradient method for linear inverse problems with sparsity constraints*, arXiv:0706:4297, 2007.
[15] C. De Mol and M. Defrise, *A note on wavelet-based inversion algorithms*, Contemp. Math., 313 (2002), pp. 85–96.
[16] D. Donoho and J. Tanner, *Neighborliness of randomly-projected simplices in high dimensions*, Proc. Nat. Acad. Sci., 102 (2005), pp. 9452–9457.
[17] D. Donoho, Y. Tsaig, I. Drori, and J.-C. Starck, *Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit*, Technical report 2006-02, Department of Statistics, Stanford Univerisity, Stanford, CA, 2006.
[18] D. Donoho and Y. Tsaig, *Fast solutions of ℓ_1-norm minimization problems when the solution may be sparse*, Technical report online, 2006.
[19] D. Donoho, *De-noising by soft-thresholding*, IEEE Trans. Inform. Theory, 41 (1995), pp. 613–627.
[20] D. Donoho, *Compressed sensing*, IEEE Trans. Inform. Theory, 52 (2006), pp. 1289–1306.
[21] J. Douglas and H. H. Rachford, *On the numerical solution of the heat conduction problem in 2 and 3 space variables*, Trans. Amer. Math. Soc., 82 (1956), pp. 421–439.
[22] J. Eckstein, *Splitting methods for monotone operators with applications to parallel optimization*, Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, 1989.
[23] A. Edelman, *Eigenvalues and condition numbers of random matrices*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 543–560.
[24] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, *Least angle regression*, Ann. Stat., 32 (2004), pp. 407–499.
[25] M. Elad, B. Matalon, J. Shtok, and M. Zibulevsky, *A wide-angle view at iterated shrinkage algorithms*, SPIE (Wavelet XII), San Diego, CA, August 26–29, 2007.
[26] M. Elad, B. Matalon, and M. Zibulevsky, *Coordinate and subspace optimization methods for linear least squares with non-quadratic regularization*, J. Appl. Comput. Harmonic Analysis, 23 (2006), pp. 346–367.
[27] M. Elad, *Why simple shrinkage is still relevant for redundant representations?*, IEEE Trans. Inform. Theory, 52 (2006), pp. 5559–5569.
[28] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, *Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems*, reprint, 2007.
[29] M. Figueiredo and R. Nowak, *An EM algorithm for wavelet-based image restoration*, IEEE Trans. Image Process., 12 (2003), pp. 906–916.
[30] D. Gabay, *Applications of the method of multipliers to variational inequalities*, in Augmented Lagrangian Methods: Applications to the Solution of Boundary Value Problems, M. Fortin and R. Glowinski, eds., North–Holland, Amsterdam, 1983.

[31] R. Glowinski and P. Le Tallec, *Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics*, SIAM, Philadelphia, 1989.
[32] E. Hale, W. Yin, and Y. Zhang, *A numerical study on fixed point continuation method applied to compressed sensing*, Rice University CAAM Technical report TR08-24, Rice University, Houston, TX, 2008, submitted.
[33] S. Haubruge, V. H. Nguyen, and J. J. Strodiot, *Convergence analysis and applications of the Glowinski-Le Tallec splitting method for finding a zero of the sum of two maximal monotone operators*, J. Optim. Theory Appl., 97 (1998), pp. 645–673.
[34] D. Jonsson, *Some limit theorems for the eigenvalues of a sample covariance matrix*, J. Multivariate Analysis, 12 (1982), pp. 1–38.
[35] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, *A method for large-scale l_1-regularized least squares*, IEEE J. Selected Topics Signal Process., 1 (2007), pp. 606–617.
[36] S. Kirolos, J. Laska, M. Wakin, M. Duarte, D. Baron, T. Ragheb, Y. Massoud, and R. Baraniuk, *Analog-to-information conversion via random demodulation*, in Proceedings of the IEEE Dallas Circuits and Systems Workshop (DCAS), Dallas, TX, 2006.
[37] J. Laska, S. Kirolos, M. Duarte, T. Ragheb, R. Baraniuk, and Y. Massoud, *Theory and implementaion of an analog-to information converter using random demodulation*, in Proceedings of the IEEE International Symposium on Circuites and Systems (ISCAS), New Orleans, LA, 2007.
[38] J. Laska, S. Kirolos, Y. Massoud, R. Baraniuk, A. Gilbert, M. Iwen, and M. Strauss, *Random sampling for analog-to-information conversion of wideband signals*, in Proceedings of the IEEE Dallas Circuits and Systems Workshop, Dallas, TX, 2006.
[39] S. Levy and P. Fullagar, *Reconstruction of a sparse spike train from a portion of its spectrum and application to high-resolution deconvolution*, Geophys., 46 (1981), pp. 1235–1243.
[40] P. L. Lions and B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM J. Numer. Anal., 16 (1979), pp. 964–979.
[41] Z.-Q. Luo and P. Tseng, *On the linear convergence of descent methods for convex essentially smooth minimization*, SIAM J. Control Optim., 30 (1990), pp. 408–425.
[42] M. Lustig, D. Donoho, and J. Pauly, *Sparse MRI: The application of compressed sensing for rapid MR imaging*, Magnetic Resonance in Medicine, 58 (2007), pp. 1182–1195.
[43] D. Malioutov, M. Çetin, and A. Willsky, *Homotopy continuation for sparse signal representation*, in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 5, Philadelphia, PA, 2005, pp. 733–736.
[44] S. G. Mallat and Z. Zhang, *Matching pursuits with time-frequency dictionaries*, IEEE Trans. Signal Process., 41 (1993), pp. 3397–3415.
[45] B. Mercier, *Inéquations Variationnelles de la Mécanique*, Publications Mathématiques d'Orsay, Université de Paris-Sud, Orsay, France, 80.01 (1980).
[46] A. Miller, *Subset Selection in Regression*, Chapman and Hall, London, 2002.
[47] Y. Nesterov, *Gradient methods for minimizing composite objective function*, www.optimization-online.org, CORE Discussion Paper 2007/76, 2007.
[48] M. A. Noor, *Splitting methods for pseudomonotone mixed variational inequalites*, J. Math. Anal. Appl., 246 (2000), pp. 174–188.
[49] M. Osborne, B. Presnell, and B. Turlach, *A new approach to variable selection in least squares problems*, IMA J. Numer. Anal., 20 (2000), pp. 389–403.
[50] J.-S. Pang, *A posteriori error bounds for the linearly-constrained variational inequality problem*, Math. Methods Oper. Res., 12 (1987), pp. 474–484.
[51] G. B. Passty, *Ergodic convergence to a zero of the sum of monotone operators in Hilbert space*, J. Math. Anal. Appl., 72 (1979), pp. 383–390.
[52] D. H. Peaceman and H. H. Rachford, *The numerical solution of parabolic elliptic differential equations*, SIAM J. Appl. Math., 3 (1955), pp. 28–41.
[53] R. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[54] M. Rudelson and R. Vershynin, *Geometric approach to error-correcting codes and reconstruction of signals*, Int. Math. Res. Not., (2005), pp. 4019–4041.
[55] F. Santosa and W. Symes, *Linear inversion of band-limited reflection histograms*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 1307–1330.
[56] D. Takhar, J. Laska, M. Wakin, M. Duarte, D. Baron, S. Sarvotham, K. Kelly, and R. Baraniuk, *A new compressive imaging camera architecture using optical-domain compression*, in Proceedings of Computational Imaging IV at SPIE Electronic Image, San Jose, CA, 2006.
[57] H. Taylor, S. Bank, and J. McCoy, *Deconvolution with the l_1 norm*, Geophys., 44 (1979), pp. 39–52.
[58] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Royal Statist. Soc. B, 58 (1996), pp. 267–288.

[59] J. Tropp, M. Wakin, M. Duarte, D. Baron, and R. Baraniuk, *Random filters for compressive sampling and reconstruction*, in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Toulouse, France, 2006.

[60] J. Tropp, *Just relax: Convex programming methods for identifying sparse signals*, IEEE Trans. Inform. Theory, 51 (2006), pp. 1030–1051.

[61] Y. Tsaig and D. Donoho, *Extensions of compressed sesning*, Signal Processing, 86 (2005), pp. 533–548.

[62] P. Tseng, *A modified forward-backward splitting method for maximal monotone mappings*, SIAM J. Control Optim., 38 (2000), pp. 431–446.

[63] B. Turlach, *On algorithms for solving least squares problems under an L_1 penalty or an L_1 constraint*, in Proceedings of the American Statistical Association; Statistical Computing Section, Alexandria, VA, 2005, pp. 2572–2577.

[64] E. Van den Berg and M. P. Friedlander, *In pursuit of a root*, UBC Computer Science Technical Report TR-2007-16, 2007.

[65] M. Wakin, J. Laska, M. Duarte, D. Baron, S. Sarvotham, D. Takhar, K. Kelly, and R. Baraniuk, *An architecture for compressiving image*, in Proceedings of the International Conference on Image Processing (ICIP), Atlanta, Georgia, 2006.

[66] M. Wakin, J. Laska, M. Duarte, D. Baron, S. Sarvotham, D. Takhar, K. Kelly, and R. Baraniuk, *Compressive imaging for video representation and coding*, in Proceedings of Picture Coding Symposium (PCS), Beijing, China, 2006.

[67] Y. Zhang, *When is missing data recoverable?*, Rice University CAAM Technical Report TR06-15, 2006.

SIAM J. OPTIM.
Vol. 19, No. 3, pp. 1131–1149

POLYHEDRAL RESULTS FOR 1-RESTRICTED SIMPLE 2-MATCHINGS*

DAVID HARTVIGSEN† AND YANJUN LI‡

Abstract. A simple 2-matching in a graph is a subgraph whose connected components are nontrivial paths and cycles. A simple 2-matching is called 1-restricted if each connected component has two or more edges. In this paper we consider the problem of finding maximum weight 1-restricted simple 2-matchings (which is a relaxation of the traveling salesman problem). We present an integer programming formulation for this problem, characterize the extreme points of the linear programming relaxation, and characterize the graphs for which the linear programming relaxation has all integral extreme points. We show how to recognize these graphs in polynomial time. We also introduce a new class of blossom-type inequalities that tighten the general linear programming relaxation. A complete description of the convex hull of 1-restricted simple 2-matchings is unknown.

Key words. matchings, combinatorial optimization, polyhedral combinatorics, traveling salesman problem

AMS subject classifications. 05C70, 90C27, 90C57

DOI. 10.1137/070697409

1. Introduction.

All graphs $G = (V, E)$ considered in this paper are undirected and (except where noted) have no parallel edges or loops. We associate a real weight with every $e \in E$. A *simple* 2*-matching* in a graph is a subgraph all of whose nodes have degree 1 or 2. (Sometimes in the literature a simple 2-matching is defined as a subset of edges. For convenience in this paper, we use the subgraph definition.) Hence the connected components of a simple 2-matching are nontrivial paths and cycles. The *weight* of a simple 2-matching is the sum of the weights of the edges in the matching. A well-studied problem in the literature is to find a maximum weight simple 2-matching in a graph. A polyhedral characterization and polynomial-time algorithm (due to Edmonds [5] and Johnson [16], respectively; see also [8]), as well as a number of structural theorems (e.g., see Belck [1] and Gallai [9]) are known for this problem, among other things (see [21] for an excellent survey). In this paper we present some analogous results for a restricted version of this problem, which we describe next.

A k*-restricted simple* 2*-matching* is a simple 2-matching such that each connected component has more than k edges. Hence 0-restricted simple 2-matchings are equivalent to simple 2-matchings, and 1-restricted simple 2-matchings are equivalent to simple 2-matchings that contain no "isolated edges." We are interested in the problem of finding maximum weight k-restricted simple 2-matchings. Solutions to this problem on complete graphs, with nonnegative weights, can be shown to provide increasingly accurate approximate solutions for the traveling salesman problem as k increases (see [6]; although k-restricted simple 2-matchings are not discussed in [6], their approach can easily be seen to apply to our problem).

*Received by the editors July 16, 2007; accepted for publication (in revised form) June 2, 2008; published electronically October 31, 2008.
http://www.siam.org/journals/siopt/19-3/69740.html
†Mendoza College of Business, University of Notre Dame, Notre Dame, IN 46556 (David.Hartvigsen.1@nd.edu).
‡Krannert School of Management, Purdue University, West Lafayette, IN 47907 (li14@purdue.edu).

In this paper we begin by introducing an integer programming formulation, call it IP, for the problem of finding a maximum weight 1-restricted simple 2-matching in a graph G. This formulation includes a new class of constraints, which we call *edge-adjacency constraints*. We let $A(G)$ denote the feasible points in the linear programming relaxation of IP, and we let $R(G)$ denote the convex hull of integral points in $A(G)$. (Hence the extreme points of $R(G)$ are precisely the 0,1 incidence vectors of 1-restricted simple 2-matchings of G.)

The paper then contains four main results. The first result characterizes the extreme points of $A(G)$. We find that they have a simple structure with components in $\{0, \frac{1}{2}, 1\}$. This result then yields our second result in which we characterize the graphs G for which $A(G)$ has all integral extreme points. (That is, we characterize when $A(G) = R(G)$.) This characterization has the following form: $A(G) = R(G)$ if and only if G contains no subgraph which is a 1-*restricted odd cycle* (to be defined later). Although 1-restricted odd cycles are fairly simple to describe, they can be arbitrarily large; hence it is not immediately obvious how to polynomially determine if a graph contains one. Our third result addresses this issue. In it we present a second (more technical) characterization of the graphs with no 1-restricted odd cycle that yields a polynomial time test for this property. Our final result is a new class of inequalities (based on a new structure called a 1-*restricted blossom*) that are valid for $R(G)$ and cut off all the fractional extreme points in $A(G)$. These inequalities generalize the edge-adjacency inequalities as well as the well-known blossom inequalities for the simple 2-matching polyhedron (see [5]). We also show that the 1-restricted blossom inequalities, together with the description of $A(G)$, do not yield a complete description of $R(G)$. Finding such a description is an open problem.

Thus we characterize an apparently new class of polyhedra with integral extreme points. (The class has variable coefficients in $\{0, \pm 1\}$ and is easily seen to not be totally unimodular.) Furthermore, the linear system for this polyhedron has a size that is polynomial in the size of the associated graph. It immediately follows that, for graphs containing no 1-restricted odd cycle, we can find a maximum weight 1-restricted simple 2-matching in strongly polynomial time (using the algorithm of Tardos [22]).

Before outlining the paper, let us consider some related work in the literature. The problem of finding maximum weight 1-restricted simple 2-matchings in general graphs, for the special case that all weights equal 1 (i.e., $w \equiv 1$), was considered in [10], where a polynomial-time algorithm and several structural theorems were presented. A related problem involves C_k-*free* 2-*matchings*, which are simple 2-matchings that contain no cycles of length $\leq k$. An algorithm to find a maximum, with $w \equiv 1$, C_3-free 2-matching appears in [11] and an algorithm to find a maximum, with $w \equiv 1$, C_4-free 2-matching in bipartite graphs appears in [12] (see also [19]). The problem of finding a maximum weight C_4-free 2-matching in bipartite graphs has been observed to be NP-hard (see [23], [3], and [20]). The problem of finding a maximum, with $w \equiv 1$, C_5-free 2-matching (in general graphs) is also known to be NP-hard (see [2] for a proof due to Papadimitiou). Related work can be found in [2], [4], and [7].

Let us discuss one other related area of research. A *k-piece* is a connected graph with maximum degree equal to k. A *k-piece packing* in a graph is a subgraph whose connected components are k-pieces. The *node-max k-piece packing problem* is to find a k-piece packing in a graph that contains a maximum number of nodes; the *edge-max k-piece packing problem* is to find a k-piece packing in a graph that contains a maximum number of edges. Observe that the node-max and edge-max 1-piece packing problems are equivalent to the classical matching problem. However, for higher values of k, the node-max and edge-max problems are different from one another. Finally,

note that the edge-max 2-piece packing problem is identical to the maximum, with $w \equiv 1$, 1-restricted simple 2-matching problem.

The node-max 2-piece packing problem was recently studied by Kaneko in [17], where he presented a Tutte-type theorem. This result was extended in [18] by Kano, Katona, and Király, where the authors presented a Tutte–Berge-type theorem. These results were further extended by Hartvigsen, Hell, and Szabó in [13], where the authors presented a polynomial time algorithm and Tutte-type and Tutte–Berge-type theorems for the node-max k-piece packing problem. Finally, a Gallai–Edmonds decomposition theorem for the general problem is presented in [15] by Janata, Loebl, and Szabó. Other research of this type has been recently surveyed in [14] by Hell (see also [21]).

Let us finish this section by outlining the paper. Section 2 contains some basic notation and definitions and section 3 contains our characterization of the extreme points of $A(G)$. Section 4 contains our characterization of the graphs G for which $A(G)$ has all integral extreme points plus several related results. In section 5 we introduce a new class of blossom-like inequalities that are valid for all integral points of $A(G)$. We show that these inequalities cut off all fractional extreme points of $A(G)$, but they do not yield a complete description of $R(G)$.

2. Notation and an IP formulation. In this section we introduce some notation and terminology, and we present an integer programming formulation of the problem of finding a maximum weight 1-restricted simple 2-matching. We begin with a few definitions.

Let $G = (V, E)$ be a graph. For $v \in V$, let $\delta(v)$ denote the subset of edges incident with v. For $uv \in E$, let $adj(uv)$ denote the set of edges incident with exactly one node in $\{u, v\}$; that is, $adj(uv) = \delta(u) + \delta(v) - uv$. For $S \subseteq E$, let $x(S) = \sum_{e \in S} x_e$. Given a subgraph $G' = (V', E')$ of G, a vector $x \in \{0, 1\}^E$ is called the 0-1 *incidence vector* for G' if $x_e = 1$ for every $e \in E'$, and $x_e = 0$ otherwise. We denote this vector $\chi^{E'}$. We use the term *cycle* to refer to a connected subgraph all of whose nodes have degree 2; and we use the term *path* to refer to a connected subgraph such that exactly two nodes have degree 1 and the remaining nodes have degree 2.

Let $A(G)$ denote the set of points $x \in \mathbb{R}^E$ that satisfy the following system:

$$x(\delta(v)) \leq 2 \quad \forall\, v \in V, \tag{1}$$

$$x_e \leq 1 \quad \forall\, e \in E, \tag{2}$$

$$x_e - x(adj(e)) \leq 0 \quad \forall\, e \in E, \tag{3}$$

$$x_e \geq 0 \quad \forall\, e \in E. \tag{4}$$

Observe that the integral points in $A(G)$ are precisely the 0-1 incidence vectors of 1-restricted simple 2-matchings in G. The key to this working is the new type of constraints (3), which we call *edge-adjacency constraints*. These constraints say that if an edge e is in the matching (i.e., $x_e = 1$), then at least one edge adjacent to e must also be in the matching.

Hence, for a weight vector $w \in \mathbb{R}^E$,

$$\max\ wx \ \text{ s.t. } x \in A(G),\ x \text{ integral}$$

is an integer programming formulation for the problem of finding a maximum weight 1-restricted simple 2-matching.

Let $R(G)$ denote the convex hull of the 0-1 incidence vectors of 1-restricted simple 2-matchings of G. Thus, $R(G) \subseteq A(G)$. In the next section we characterize the

extreme points of $A(G)$ and in the section following that we characterize the graphs G for which $A(G) = R(G)$.

3. Characterization of the extreme points of $A(G)$. Our main result in this section is a characterization of the extreme points of $A(G)$ on a general graph $G = (V, E)$. We then present a characterization of the extreme points of (1–2) and (4), whose integer solutions are the incidence vectors of classic simple 2-matchings. An interesting resemblance between these two characterizations is exhibited.

Let x be a feasible solution to $A(G)$ for a graph G. We say x_e and e are *fractional* if $0 < x_e < 1$, and we call x_e the *value* of e. An edge $e = uv$ of G is called 1-*tight* if e has value 1, one edge incident to u has value $\frac{1}{2}$, one edge incident to v has value $\frac{1}{2}$, and the remaining edges adjacent to e have value 0. Observe that 1-tight edges satisfy (3) at equality. Let G' denote a graph obtained from G by contracting a set of edges. (*Contracting* an edge is the operation of deleting the edge and identifying its endnodes. Observe that a graph resulting from a contraction can have parallel edges.) The nodes of G' resulting from contractions are called *shrunk*; the remaining nodes of G' are called *real*. The edges of G' inherit values from x in the obvious manner. Next we give our characterization of the extreme points of $A(G)$.

THEOREM 1. *A point $x \in A(G)$ is extreme if and only if it satisfies both of the following conditions:*

(i) *The point x has only values 0, $\frac{1}{2}$, and 1.*

(ii) *Let G^c be the graph obtained from G by contracting all the 1-tight edges. Then the edges of G^c with value $\frac{1}{2}$ form node-disjoint odd cycles, and every real node in such an odd cycle is incident with an edge with value 1.*

Before we give a proof of Theorem 1, we show in Figure 1 an example of an extreme point of $A(G)$. A subgraph of G is shown in Figure 1(a) with the x values given next to the edges, and the corresponding subgraph of G^c after contraction of the 1-tight edges is shown in Figure 1(b). Observe that edge uv of G is a 1-tight edge. Edge uv is contracted and becomes the shrunk node s in G^c. All the edges of the illustrated subgraph of G^c with value $\frac{1}{2}$ form node-disjoint odd cycles, and every node of such an odd cycle, except the shrunk node, is incident with an edge with value 1.

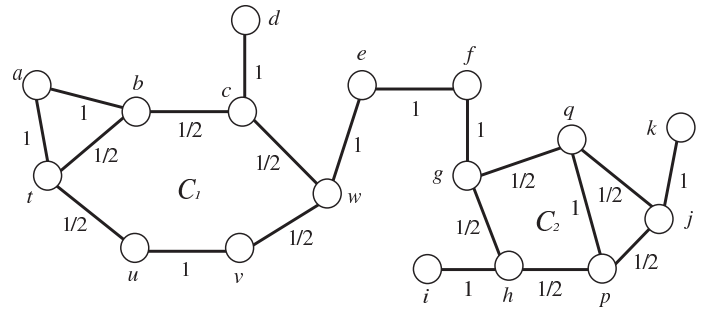
We use the following definition and lemmas in our proof of Theorem 1. Let $x \in A(G)$ for a graph $G = (V, E)$. We say an edge $e \in E$ is *adjacency tight (with respect to x)* if e satisfies inequality (3) at equality and is adjacent to at least one fractional edge.

LEMMA 2. *Let x be an extreme feasible solution to $A(G)$ for a graph G. If e_1 and e_2 are adjacency-tight edges of G, then e_1 is not adjacent to e_2.*

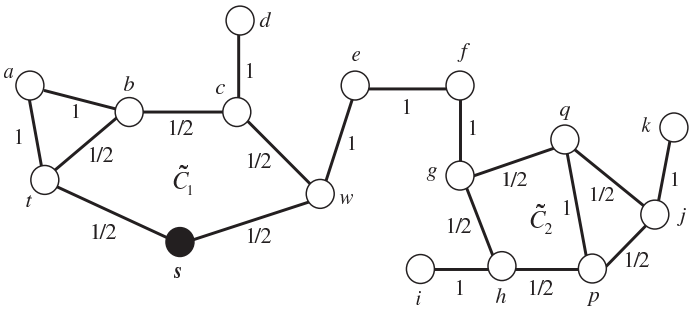
Proof. Let x be an extreme feasible solution to $A(G)$ for a graph G and let e_1 and e_2 be adjacency-tight edges that are adjacent to each other. We derive a contradiction.

First, observe that $x_{e_1} > 0$, since e_1 is adjacency tight. Similarly, $x_{e_2} > 0$. Furthermore, $x_{e_1} \geq x_{e_2}$ because e_1 is adjacency tight. Similarly, $x_{e_2} \geq x_{e_1}$. So $x_{e_1} = x_{e_2}$. Suppose there is an edge $e' \neq e_2$ that is adjacent to e_1 and satisfies $x_{e'} > 0$. Then e_1 being adjacency tight and $x_{e_2} > 0$ imply $x_{e_1} > x_{e_2}$, which is a contradiction. It follows that x_{e_1} and x_{e_2} are fractional, since if they were both equal to 1, then they would not be adjacent to at least one fractional edge. Finally, set $x^1 = x + \epsilon\chi^{\{e_1,e_2\}}$ and $x^2 = x - \epsilon\chi^{\{e_1,e_2\}}$ for some small $\epsilon > 0$. It is easy to see that $x = \frac{1}{2}x^1 + \frac{1}{2}x^2$ and that both x^1 and x^2 are feasible solutions to $A(G)$. Therefore, x is not extreme, which is a contradiction. ∎

LEMMA 3. *Let $x \in A(G)$ for a graph G and let edge $e = uv$ be adjacency-tight.*



(a) subgraph of G



(b) subgraph of G^c

FIG. 1. *An example of an extreme point of* $A(G)$.

(i) *If* u *and* v *are each incident with a fractional edge different from* e*, then* $0 < x(\delta(u)) < 2$ *and* $0 < x(\delta(v)) < 2$.

(ii) *If* e *is adjacent to exactly one fractional edge, then* e *is fractional in* G *and* $0 < x(\delta(u)) < 2$ *and* $0 < x(\delta(v)) < 2$.

Proof. First we prove (i). Clearly, since u and v are incident with fractional edges, $x(\delta(u)) > 0$ and $x(\delta(v)) > 0$. Suppose $x(\delta(u)) = 2$. Then

$$x_e + x(\delta(u) - e) = 2.$$

Because e is adjacency tight, we have that

$$x_e = x(\delta(u) - e) + x(\delta(v) - e).$$

If we solve the first equation for $x(\delta(u) - e)$ and substitute this into the second equation, we get

$$2 - 2x_e = -x(\delta(v) - e).$$

Since the maximum value that x_e can take is 1, $x(\delta(v) - e) \leq 0$, so we have a contradiction and the result follows.

To prove (ii), let e' be the single fractional edge adjacent to e. If e is also adjacent to an edge with value 1, then $x(adj(e)) > 1$ and e cannot be adjacency tight. Hence

e' is the only edge adjacent to e with nonzero value. So e has the same value as e' and the conclusion follows. This completes the proof. □

Proof. (Theorem 1) We first prove that if $x \in A(G)$ is extreme, then it satisfies (i) and (ii) in the statement of the Theorem.

Given an extreme point $x \in A(G)$, let G' be the graph resulting from contracting each adjacency-tight edge of G. Observe that by Lemma 2, each shrunk node of G' corresponds to a unique edge of G. Let F be the subgraph of G' that is induced by the fractional edges of G'. Observe that if F is empty, then there cannot be any adjacency-tight edges; hence the result follows immediately (since all the values of x are 0, 1), so let us assume F is nonempty.

Claim 1. Every connected component of F contains a cycle.

Proof of Claim 1: Suppose not. Then F must contain a connected component that is a tree. Let P be a path in F whose endnodes have degree 1 in F. Observe that if an endnode, say v, of P was obtained by shrinking an edge, say e, then e is fractional: If $x_e = 1$, then v cannot have degree 1 in F, since e is adjacency-tight; if $x_e = 0$, then e was not adjacency-tight. Partition the edges of P into two sets, say B (blue) and R (red), so that no two adjacent edges have the same color. If an endnode v of P is shrunk, then add the corresponding edge of G to the same set, B or R, that contains the edge of P incident with v. Define x^1 and x^2 on E as follows: $x^1 = x + \epsilon(\chi^B - \chi^R)$ and $x^2 = x + \epsilon(\chi^R - \chi^B)$, where $\epsilon > 0$. Observe that $x = \frac{1}{2}x^1 + \frac{1}{2}x^2$. It remains to show that, for sufficiently small ϵ, x^1 and x^2 are feasible solutions to $A(G)$, which contradicts x being extreme.

Clearly this is true for inequalities (2) and (4), for sufficiently small $\epsilon > 0$, since all the edges whose values are changed are fractional. It is also easy to see that this property holds for inequalities (3) for all edges of P (since (3) is not tight for them), for all edges inside shrunk nodes of P, and for all edges not in P, but adjacent to edges in P (which cannot be adjacency tight). So let us consider inequalities (1). Again, the property clearly holds for all real nodes of P. So consider a shrunk node v of P that was obtained by shrinking the edge v_1v_2 of G. Suppose v is an interior node of P. If v_1 is adjacent in G to the two edges of P that are adjacent to v, then $x(\delta(v_1))$ and $x(\delta(v_2))$ have the same values in x^1 and x^2. Otherwise, using Lemma 3(i), we have that $0 < x(\delta(v_1)) < 2$ and $0 < x(\delta(v_2)) < 2$. Suppose v is an endnode of P. Lemma 3(ii) implies that $0 < x(\delta(v_1)) < 2$ and $0 < x(\delta(v_2)) < 2$. Claim 1 now follows.

Claim 2. F does not contain an even cycle.

Proof of Claim 2: Again, suppose this is not the case. Let C be an even cycle of F. Partition the edges of C into two sets, say B (blue) and R (red), and define x^1 and x^2 on E, as in the proof of Claim 1. Using the same logic as in the proof of Claim 1, one can show that x is not an extreme feasible solution, which is a contradiction.

Claim 3. None of the nodes on an odd cycle in F is incident with a fractional edge other than the two edges incident to it in the cycle.

Proof of Claim 3: By contradiction. Let C be an odd cycle in F, let v be a node on F, and let vv' be an edge of F not in C. Let us assume v' is not on C. (If v' were on C, then vv' would form an even cycle with a portion of F, which has been ruled out in Claim 2.) To begin, let P be the path vv'. Let us extend the graph P by repeatedly adding an edge of F to the endnode of P that is different from v, until one of the following happens:

Case 1: The new edge of P has an endnode with degree 1 in F.

Case 2: The new edge of P contains a node in C, which is different from v.

Case 3: The new edge of P contains a node of P (possibly v).

Case 1: Let w denote the endnode of P not on C. Partition the edges of $C \cup P$ into two sets, say B (blue) and R (red), so the two edges of C incident with v have the same color and no other two adjacent edges of $C \cup P$ have the same color. If w is shrunk, then add the corresponding edge of G to the same set, B or R, that contains the edge of P incident with w. Let B^C, B^P, R^C, R^P denote the edges of B in C, the edges of B in P, and so on. Define x^1 and x^2 on E as follows: $x^1 = x + \epsilon(\chi^{B^C} + 2\chi^{B^P} - \chi^{R^C} - 2\chi^{R^P})$ and $x^2 = x + \epsilon(\chi^{R^C} + 2\chi^{R^P} - \chi^{B^C} - 2\chi^{B^P})$, where $\epsilon > 0$. Observe that $x = \frac{1}{2}x^1 + \frac{1}{2}x^2$. It remains to show that, for sufficiently small ϵ, x^1 and x^2 are feasible solutions to $A(G)$, which contradicts x being extreme. (The following logic is quite close to that used in Claim 1.)

Clearly this is true for inequalities (2) and (4), for sufficiently small $\epsilon > 0$, since all the edges whose values are changed are fractional. It is also easy to see that this property holds for inequalities (3) for all edges of $C \cup P$ (since (3) is not tight for them), for all edges inside shrunk nodes of $C \cup P$, and for all edges not in $C \cup P$, but adjacent to edges in $C \cup P$ (which cannot be adjacency tight). So let us consider inequalities (1). Again, the property clearly holds for all real nodes of $C \cup P$. So consider a shrunk node u of $C \cup P$ that was obtained by shrinking the edge u_1u_2 of G. Suppose $u \neq w$. If u_1 is adjacent in G to all the edges of $C \cup P$ that are adjacent to u, then $x(\delta(u_1))$ and $x(\delta(u_2))$ have the same values in x^1 and x^2. Otherwise, using Lemma 3(i), we have that $0 < x(\delta(u_1)) < 2$ and $0 < x(\delta(u_2)) < 2$. Suppose $u = w$. Lemma 3(ii) implies that $0 < x(\delta(u_1)) < 2$ and $0 < x(\delta(u_2)) < 2$. Case 1 now follows.

Case 2: It is evident that $C \cup P$ contains an even cycle, and hence we can apply Claim 2.

Case 3: $C \cup P$ must consist of two odd cycles, say C and C', connected by a path, say Q (possibly of length 0), that shares only its endnodes, say v and v', with C and C', respectively.

Partition the edges of $C \cup Q \cup C'$ into two sets, say B (blue) and R (red), so the two edges of C incident with v have the same color, the two edges of C' incident with v' have the same color, and no other two adjacent edges of $C \cup Q \cup C'$ have the same color. Let $D = C \cup C'$ and let B^D, B^Q, R^D, R^Q denote the edges of B in D, the edges of B in Q, and so on. Define x^1 and x^2 on E as follows: $x^1 = x + \epsilon(\chi^{B^D} + 2\chi^{B^Q} - \chi^{R^D} - 2\chi^{R^Q})$ and $x^2 = x + \epsilon(\chi^{R^D} + 2\chi^{R^Q} - \chi^{B^D} - 2\chi^{B^Q})$, where $\epsilon > 0$. The argument then proceeds as above, which finishes the proof of the claim.

Claim 4. For any odd cycle C in F, x has value 1 on every contracted edge that corresponds to a shrunk node of C, and $x(\delta(v)) = 2$ on every real node v of C.

Proof of Claim 4: Let C be an odd cycle in F and let v be a shrunk node on C obtained from contracting edge $e = v_1v_2$ in G. Let e_1 and e_2 be the two edges of C incident with v. Since e_1 and e_2 are fractional and e is adjacency tight, $x_e > 0$. Let us assume that e is fractional in G and then derive a contradiction.

Partition the edges of C into two sets, say B (blue) and R (red), so the two edges of C incident with v have the same color, say B, and no other two adjacent edges of C have the same color. Define x^1 and x^2 on E as follows: $x^1 = x + \epsilon(\chi^R - \chi^B + 2\chi^e)$ and $x^2 = x + \epsilon(\chi^B - \chi^R - 2\chi^e)$, where $\epsilon > 0$. Observe that $x = \frac{1}{2}x^1 + \frac{1}{2}x^2$. It remains to show that, for sufficiently small ϵ, x^1 and x^2 are feasible solutions to $A(G)$, which contradicts x being extreme.

Arguing as in the proof of Claim 1, it is not difficult to show that this is true for inequalities (2), (3), and (4). So let us consider inequality (1).

Case 1: e_1 and e_2 are adjacent to v_1 in G. Observe that v_2 cannot be adjacent to a fractional edge in G by Claim 3. Furthermore, v_2 cannot be adjacent to an edge at

value 1, since e is fractional and adjacency tight. Hence $x(\delta(v_2)) < 1$. Since $x(\delta(v_1))$ remains unchanged in x^1 and x^2, the result follows.

Case 2: e_1 and e_2 are adjacent to v_1 and v_2, respectively, in G. Then by Lemma 3(i), $0 < x(\delta(v_1)) < 2$ and $0 < x(\delta(v_2)) < 2$; so, again, the result follows.

Now we prove that $x(\delta(v)) = 2$ on every real node v of C. Let us assume that v is a real node on C satisfying $x(\delta(v)) < 2$. Proceed just as we did for the previous case that v is shrunk (except we don't have the edge e). The result follows as before, except showing that x^1 and x^2 satisfy (1) for sufficiently small ϵ is immediate because $x(\delta(v)) < 2$.

Claim 5. Every odd cycle in F consists of edges with value $\frac{1}{2}$.

Proof of Claim 5: Consider an odd cycle C in F. Let $u_1, u_2, \ldots, u_{2m+1}$ be the cycle nodes and let $u_1u_2, u_2u_3, \ldots, u_{2m}u_{2m+1}, u_{2m+1}u_1$ be the cycle edges. Let $x_{u_1u_2} = \alpha$, where $0 < \alpha < 1$. By Claim 3 and Claim 4, every real node on C is incident with an edge of G with value 1, and every shrunk node on C is incident with no other nonzero-value edge of G' except the two adjacent cycle edges. Therefore, we have $x_{u_2u_3} = 1-\alpha$, $x_{u_3u_4} = \alpha, \ldots, x_{u_{2m}u_{2m+1}} = 1 - \alpha$, and $x_{u_{2m+1}u_1} = \alpha$. Since $x_{u_1u_2} + x_{u_{2m+1}u_1} = 1$, $\alpha = \frac{1}{2}$. Claim 5 is proved.

It follows from the above five claims that if a feasible solution x is extreme, then it satisfies (i) and (ii).

Now we prove that if a feasible solution x satisfies (i) and (ii), then it is extreme. Assume, by contradiction, that there are two different feasible solutions x^1 and x^2 such that their convex combination is x. We will derive a contradiction by showing $x = x^1 = x^2$.

First, we clearly see that $x = x^1 = x^2$ on the edges of G that have value 0 or 1. Now consider an odd cycle C in F whose edges have value $\frac{1}{2}$. For every node v of C, (ii) implies that $x_C(\delta(v)) = 1$. This follows from the fact that the inequality (1) is tight at v if v is a real node, and the inequality (3) is tight on edge e if v was obtained by shrinking e. Hence, the inequality (1) or (3) is also tight for x^1 and x^2 at v, which means that $x^1_C(\delta(v)) = x^2_C(\delta(v)) = 1$ at every node v of C. Let $u_1, u_2, \ldots, u_{2m+1}$ be the nodes of C and let $u_1u_2, u_2u_3, \ldots, u_{2m}u_{2m+1}, u_{2m+1}u_1$ be the cycle edges. Let $x^1_{u_1u_2} = \alpha$, where $0 < \alpha < 1$. By a similar argument as in Claim 5, we see that x^1 has value $\frac{1}{2}$ on all the cycle edges of C, and so does x^2. Therefore, $x = x^1 = x^2$ on all the odd cycles in F, which yields a contradiction. □

As an interesting analog to the result of Theorem 1, the extreme points of (1–2) and (4) are characterized by the following theorem.

THEOREM 4. *A feasible solution x to* (1–2) *and* (4) *is extreme if and only if it satisfies both of the following conditions:*

(i) *The solution x has only values* 0, $\frac{1}{2}$, *and* 1*;*

(ii) *The edges with value $\frac{1}{2}$ form node-disjoint odd cycles, and every node of such an odd cycle is incident with an edge with value* 1.

The proof of Theorem 4 is similar to but simpler than that of Theorem 1, so it is not given in this paper. For comparison purposes, we provide in Figure 2 an example of an extreme point of (1–2) and (4). Some similarities between this figure and Figure 1 can be observed.

4. Characterization and recognition of the graphs for which $A(G) = R(G)$. The main purpose of this section is to characterize, in two ways, the graphs G for which the extreme points of $A(G)$ are incidence vectors of 1-restricted simple 2-matchings. That is, we characterize the graphs G for which $A(G) = R(G)$. The first characterization is conceptually simple and is in terms of an excluded subgraph. Its

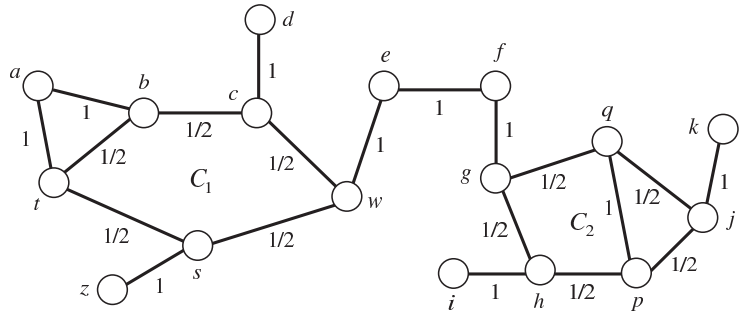


FIG. 2. *An example of an extreme point of* (1–2) *and* (4).

proof follows easily from the extreme point result in the previous section. The second characterization is a bit more technical, but it immediately leads to a polynomial time recognition algorithm for the graphs G such that $A(G) = R(G)$. These two characterizations are stated in the first subsection. The algorithm and some related results and observations appear in the second subsection. A proof of the second characterization appears in the third subsection.

4.1. Characterization theorems. We begin this section by defining a class of graphs that yields our first characterization of the graphs G such that $A(G) = R(G)$. After a few more definitions, we next state our second characterization result.

DEFINITION 5. *A graph $G = (V, E)$ is a* 1-restricted odd cycle *if it contains a cycle C and a set of edges M such that*

(i) *Every node of C is incident with exactly one edge of M;*

(ii) *Every edge of M is incident with one or two nodes of C;*

(iii) *Every node not in C is incident with one or two edges of M;*

(iv) $E = E(C) \cup M$*;*

(v) $|E(C)\backslash M|$ *is odd.*

An edge of M with exactly one node in C is called a petal *of C. An edge of M that has two nodes in C but is not an edge of C is called a* chord *of C. A node not in C is called a* tip*. A node of C that occurs in a petal or chord is called an* attachment node*. Two attachment nodes of C, say u and v, are called* adjacent *if there exists a path from u to v on C that contains no other attachment nodes. Such a path is called an* attachment node path*.*

In Figure 3, an example of 1-restricted odd cycle with 15 nodes is shown. There are 13 nodes in cycle C: $a, b, \ldots, m$. The set M has eight edges (shown as bold), while $E(C)\backslash M$ has nine edges (shown as non-bold). There are three petals an, bn and co, one chord hk, two tips n and o, and five attachment nodes a, b, c, h and k. If this 1-restricted odd cycle is a subgraph of a graph G, by Theorem 1, a fractional extreme point of $A(G)$ can be constructed by setting each bold edge to 1, each non-bold edge to $\frac{1}{2}$, and any remaining edges in G to 0.

As a consequence of Theorem 1, the following necessary and sufficient conditions hold for $A(G) = R(G)$.

THEOREM 6. *For a graph G, $A(G) = R(G)$ if and only if G does not contain a* 1*-restricted odd cycle.*

Proof. If G contains a 1-restricted odd cycle consisting of a cycle C and a set of edges M, by Theorem 1, we can construct a fractional extreme point of $A(G)$ by assigning value 1 to each edge in M, assigning value $\frac{1}{2}$ to each edge in $E(C)\backslash M$, and assigning value 0 to the remaining edges of G. Hence, $A(G) \neq R(G)$.

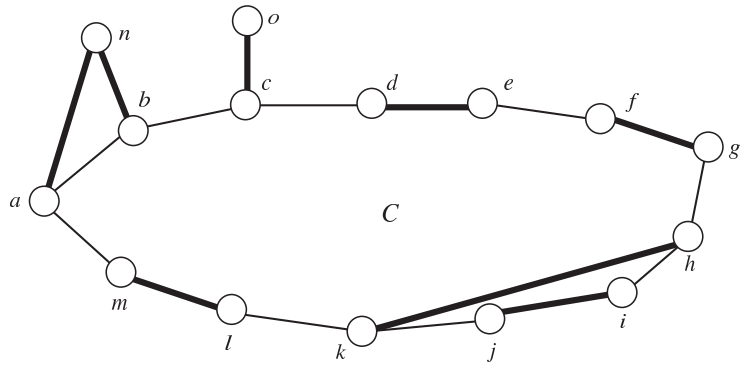


FIG. 3. *An example of a* 1*-restricted odd cycle.*

If $A(G) \neq R(G)$, $A(G)$ must contain a fractional extreme point that is characterized by Theorem 1. Choose an arbitrary odd cycle of G^c, say C_0, whose edges all have value $\frac{1}{2}$. Let S_1 be the set of 1-tight edges of G with exactly one endnode incident with an edge of C_0. Let S_2 be the set of 1-tight edges of G with each endnode incident with an edge of C_0. Let R be the set of edges of G that have value 1 and are incident with the real nodes of C_0. By Definition 5, we find a 1-restricted odd cycle of G defined as follows: Let C be the cycle that consists of the edges of C_0 and S_2, and let M be the edge set that is the union of S_1, S_2, and R. □

Next we define a special class of 1-restricted odd cycles that plays an important role in our second characterization theorem and the corresponding recognition algorithm for the class of graphs that have no 1-restricted odd cycles.

DEFINITION 7. *Let G be a* 1*-restricted odd cycle with a cycle C and a set of edges M. G is called a* fundamental 1-restricted odd cycle *if C does not have a chord in M and it is one of the following types:*

(i) *C contains* 0 *mod* 4 *edges and exactly two petals.*

(ii) *C contains* 1 *mod* 4 *edges and exactly one petal.*

(iii) *C contains* 2 *mod* 4 *edges and no petals.*

(iv) *C contains* 3 *mod* 4 *edges and exactly three petals.*

Note that, for the fundamental 1-restricted odd cycles of types (i) and (iv), the attachment node paths are odd; for the fundamental 1-restricted odd cycles of types (ii), C has odd length. The following proposition describes a relationship between the 1-restricted odd cycles and the fundamental 1-restricted odd cycles in a graph.

PROPOSITION 8. *A graph contains a* 1*-restricted odd cycle if and only if it contains a fundamental* 1*-restricted odd cycle.*

Proof. Since every fundamental 1-restricted odd cycle is a 1-restricted odd cycle, it is sufficient to show that if a graph G contains a 1-restricted odd cycle, then it contains a fundamental 1-restricted odd cycle. Let C be the cycle and M be the set of edges for the 1-restricted odd cycle.

Case 1: The 1-restricted odd cycle in G has no chord in M.

If $|E(C)|$ is 0 mod 4, then, since $|E(C)\backslash M|$ is odd, C must have at least two petals in M. Let e_1 and e_2 be any two petals whose respective attachment nodes, v_1 and v_2, are adjacent. Then we know that the two paths P_1 and P_2 in C between v_1 and v_2 are odd. We define a subgraph with a cycle C' and a set of edges M' as follows: $C' = C$, M' consists of e_1 and e_2 and some edges in P_1 and P_2 such that every node of C is incident with exactly one edge of M'. Since $|E(C)|$ is 0 mod 4, $|E(C')\backslash M'|$ is odd. Hence, the constructed subgraph is a fundamental 1-restricted odd cycle of type (i).

If $|E(C)|$ is 1 mod 4, C must have at least one petal in M. Let e be any such petal. Now we define a subgraph with a cycle C' and a set of edges M': $C' = C$; M' consists of e and some edges in C such that every node of C is incident with exactly one edge of M'. Since $|E(C)|$ is 1 mod 4, $|E(C')\backslash M'|$ is odd. So the constructed subgraph is a fundamental 1-restricted odd cycle of type (ii).

If $|E(C)|$ is 2 mod 4, we define a subgraph with a cycle C' and a set of edges M' as follows: $C' = C$; M' consists of some edges in C such that every node of C is incident with exactly one edge of M'. Since $|E(C)|$ is 2 mod 4, $|E(C')\backslash M'|$ is odd. The constructed subgraph is a fundamental 1-restricted odd cycle of type (iii).

If $|E(C)|$ is 3 mod 4, since $|E(C)\backslash M|$ is odd, C must have at least three petals in M. Let e_1, e_2 and e_3 be any three petals with respective attachment nodes v_1, v_2, and v_3. Suppose v_1 and v_2 are adjacent and that v_2 and v_3 are adjacent. Let P_1 be the attachment node path for v_1 and v_2 and let P_2 be the attachment node path for v_2 and v_3. Let P_3 be the path from v_1 to v_3 in C that does not contain v_2. Observe that P_1, P_2, and P_3 are each odd. We define a subgraph with a cycle C' and a set of edges M' as follows: $C' = C$, M' consists of e_1, e_2, and e_3 and some edges in P_1, P_2, and P_3 such that every node of C is incident with exactly one edge of M'. Since $|E(C)|$ is 3 mod 4, $|E(C')\backslash M'|$ is odd. Therefore, the constructed subgraph is a fundamental 1-restricted odd cycle of type (iv).

Case 2: The 1-restricted odd cycle in G has at least one chord in M.

It suffices to show that we can construct another 1-restricted odd cycle with a cycle C' satisfying $V(C') \subset V(C)$. Let e be a chord of C in M, let v_1 and v_2 be the attachment nodes of e, and let P_1 and P_2 be the two paths in C between v_1 and v_2. Since $|E(C)\backslash M|$ is odd, one of $|E(P_1)\backslash M|$ and $|E(P_2)\backslash M|$ is even. Assume that $|E(P_1)\backslash M|$ is even. We define a subgraph with a cycle C' and a set of edges M' as follows: C' contains P_1 and e, M' consists of the two edges on P_2 incident with v_1 or v_2 and the edges in M incident with the nodes in P_1. Apparently, $|E(C')\backslash M'|$ is odd and $V(C') \subset V(C)$. Also, one can easily verify that the constructed subgraph is a 1-restricted odd cycle. □

We next consider how to determine if a graph has a fundamental 1-restricted odd cycle. We would like to find a nice characterization of the graphs without fundamental 1-restricted odd cycles that leads to a polynomial-time recognition algorithm of such graphs. In order to do so, we need to define the following special graphs.

DEFINITION 9. *A k mod 4 ear, for $k = 0, 1, 2$ or 3, in a graph G is either*

(i) *a nontrivial path P with k mod 4 edges whose two endnodes have degree more than 2 in G and whose interior nodes, if any, have degree 2 in G; or*

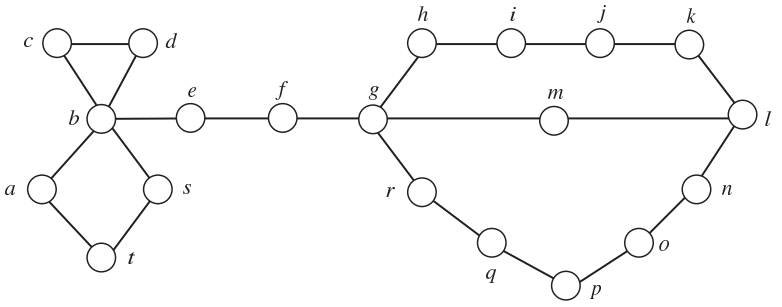
(ii) *a nontrivial cycle with k mod 4 edges that contains at most one node of degree more than 2 in G.*

DEFINITION 10. *The reduced graph G^R of a given graph G is defined as follows:*

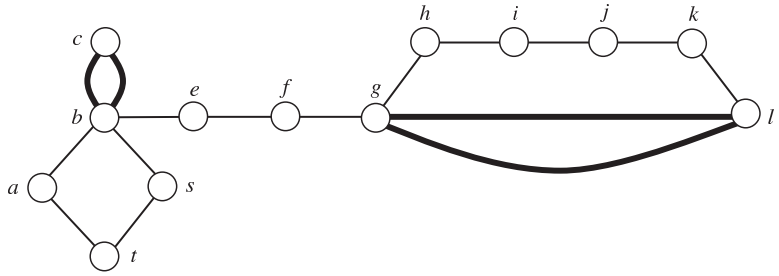
(i) *For every 2 mod 4 ear in G, replace it with a single edge (which is a loop if the ear is a cycle) so that all nodes in G of degree more than 2 maintain their degrees in G^R;*

(ii) *For every 3 mod 4 ear in G, replace it with a path of length 2 or two parallel edges (if the ear is a cycle) so that all the nodes in G of degree more than 2 maintain their degrees in G^R.*

In Figure 4, a graph G and its reduced graph G^R are given to illustrate Definitions 9 and 10. G has three maximal 2-connected subgraphs: the subgraph induced by b, c, and d; the subgraph induced by a, b, s, and t; and the subgraph induced by $g, h, i, j, k, l, m, n, o, p, q$, and r. G has a 0 *mod* 4 ear with edges ab, bs, st, and ta; a 1 *mod* 4 ear with gh, hi, ij, jk, and kl; two 2 *mod* 4 ears: one with gm and ml, the



Graph G



Reduced graph G^R

FIG. 4. *An example of G and G^R.*

other with ln, no, op, pq, qr, and rg; a 3 *mod* 4 ear with bc, cd, and db. The 3 *mod* 4 ear in G is shrunk to two parallel (bold) edges bc in G^R, and two 2 *mod* 4 ears are shrunk to two parallel (bold) edges gl.

One can easily verify that the graph G in Figure 4 has no subgraph that is a fundamental 1-restricted odd cycle. We also observe that G contains at most one odd-length ear in each maximal 2-connected subgraph and G^R is bipartite. The following theorem characterizes the graphs without fundamental 1-restricted odd cycles, which results in a polynomial time algorithm for recognizing the graphs that contain no 1-restricted odd cycles. The algorithm is given in section 4.2 and the proof of Theorem 11 is given in section 4.3.

THEOREM 11. *A connected graph G contains no fundamental* 1*-restricted odd cycle as a subgraph if and only if either G is K_4 (a complete graph with four nodes), or G is a cycle of length* 1 *mod* 4*, or G contains at most one odd-length ear in each maximal* 2*-connected subgraph and G^R is bipartite.*

4.2. Some related results and observations. In this subsection we gather some consequences of the characterization theorems stated in the previous subsection. Chief among them are polynomial time algorithms for recognizing graphs without 1-restricted odd cycles and for finding maximum weight 1-restricted simple 2-matchings in such graphs. We also state some related results for the problem of finding maximum weight (0-restricted) simple 2-matchings.

We begin by observing that the characterization theorems in the previous section yield two simple-to-describe classes of graphs that contain no 1-restricted odd cycles. The first class is trees. The second class contains any graph constructed as follows:

Take an arbitrary graph and replace every edge with a nontrivial path of length 0 *mod* 4. It is easy to see that any such graph does not contain odd-length ears and its reduced graph is bipartite.

We have characterized the graphs for which the system (1–4) has all integral extreme points. Since all entries in the constraint matrix are in $\{0, \pm 1\}$, it is natural to ask if, in this case, this matrix is totally unimodular. Consider a graph G with two adjacent edges: ab and bc. The two rows formed by the constraints 1 for node b and the constraints 3 for edge ab contain the submatrix $\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$, whose determinant is -2. Hence the constraint matrix is not totally unimodular.

We next observe that Theorem 11 immediately yields a polynomial-time algorithm to determine if a connected graph G contains a 1-restricted odd cycle, that is, to determine if $A(G) = R(G)$. Here is an outline of the algorithm: If G is K_4 or a cycle of length 1 *mod* 4, then $A(G) = R(G)$. Otherwise, identify the maximal 2-connected subgraphs of G and the odd-length ears in each maximal 2-connected subgraph. If a maximal 2-connected subgraph has more than one odd-length ear, then $A(G) \neq R(G)$. Otherwise, construct G^R. If G^R is bipartite, then $A(G) = R(G)$. Otherwise, $A(G) \neq R(G)$.

Another observation is that for graphs that contain no 1-restricted odd cycles, we can find a maximum weight 1-restricted simple 2-matching in strongly polynomial time. Here is how: For any such graph $G = (V, E)$, construct the system (1–4). Now apply the algorithm of Tardos [22] to this system. Tardos's algorithm finds an optimal solution to a linear program in a number of steps which is polynomially bounded in the size of the variable coefficient matrix; and, for our problem, the size of this matrix is polynomial since the system (1–4) has a number of entries which is polynomial in $|V|$ and $|E|$ and those entries are all in $\{0, \pm 1\}$.

Let us end this section by stating some analogous results to those in the previous section for the case of simple 2-matchings (i.e., 0-restricted simple 2-matchings). First, a graph G is called a 0-*restricted odd cycle* if it satisfies conditions (i)–(v) for a 1-restricted odd cycle (see Definition 5) plus the following condition: (vi) $E(C) \cap M = \emptyset$. Figure 2 contains two 0-restricted odd cycles: The first has an odd cycle C_1 (with nodes b, c, w, s, and t) and $M = \{at, ab, cd, ew, sz\}$; the second has an odd cycle C_2 (with nodes g, q, j, p, and h) and $M = \{fg, pq, jk, hi\}$. The following result is analogous to Theorem 6, and can be proved in a similar manner.

THEOREM 12. *For a graph G, the inequalities* (1–2) *and* (4) *define the convex hull of incidence vectors of simple* 2*-matchings of G if and only if G does not contain a* 0*-restricted odd cycle.*

THEOREM 13. *A connected graph G contains no* 0*-restricted odd cycle as a subgraph if and only if either G is K_4 or every odd cycle in G contains at least one node of degree* 2*.*

Proof. The sufficiency follows immediately, so let us prove the necessity. We assume G contains no 0-restricted odd cycle as a subgraph. An odd cycle in G is called *full* if every node of the cycle has a degree greater than 2 in G. If G contains no odd cycles or if all the odd cycles of G are not full, then the result follows. So let us assume G contains a full odd cycle. We show that G is K_4 or we derive a contradiction. Let C be the shortest full odd cycle in G. Clearly C can have no chord, since a chord implies that G has a shorter full odd cycle than C; contradiction. Since C does not induce a 0-restricted odd cycle, there must exist three edges of G, say v_1w, v_2w, and v_3w, where $v_1, v_2, v_3 \in V(C)$, and $w \notin V(C)$. If the edges v_1v_2, v_2v_3, v_1v_3 are not edges of C, then it is easy to see that there exists a shorter full cycle than C; contradiction. Hence C is a triangle and C together with w induces a

K_4. So, either G is K_4, or, since G is connected, there is an edge of G, not in the K_4, incident with an edge of the K_4, implying the existence of a 0-restricted odd cycle; again, a contradiction. □

Theorem 13 suggests the following polynomial-time algorithm for recognizing the graphs that contain no 0-restricted odd cycle: Given a graph G that is not K_4, remove all the ears of G that have length more than 1 and check if the resulting graph is bipartite. If yes, then G contains no 0-restricted odd cycle; otherwise, any odd cycle in the resulting graph yields a 0-restricted odd cycle.

4.3. Proof of Theorem 11. In this section we prove Theorem 11. To do this we present three lemmas from which the theorem immediately follows.

LEMMA 14. *If a connected graph G contains no fundamental 1-restricted odd cycle, then G contains at most one odd-length ear in each maximal 2-connected subgraph, or $G = K_4$.*

Proof. Assume that G contains no fundamental 1-restricted odd cycle, but G contains a maximal 2-connected subgraph B that contains two or more odd-length ears. We show that this leads to either a contradiction or that $G = K_4$.

Let us choose two odd-length ears in B, one with distinct endnodes u and v and the other with distinct endnodes w and z. Since B is a maximal 2-connected subgraph, there exist two node-disjoint paths in B, say from u to w and from v to z, that form a cycle C with two odd-length ears. Let us assume that we have found such a cycle C that has a minimum number of edges. It follows that C cannot contain any chord, otherwise we would have a shorter such cycle. Hence every node of u, v, w, and z is incident with an edge, the other endnode of which is not in C. We call such an edge a *petal-edge*.

If C has length 0 *mod* 4, then we obtain a fundamental 1-restricted odd cycle by adding to C the petal-edges at u and v, a contradiction. If C has length 1 *mod* 4, then we obtain a fundamental 1-restricted odd cycle by adding to C the petal-edge at u, a contradiction. If C has length 2 *mod* 4, then C is a fundamental 1-restricted odd cycle, a contradiction.

Finally, suppose that C has length 3 *mod* 4. Because the paths from u to v and from w to z in C have odd length and because C has odd length, the path from u to w in C is odd and the path from v to z is even, or vice versa. Without loss of generality, assume that the path from u to w is odd. Add to C three petal-edges from u, v, and w. Let this subgraph be G'. If G' has two or three nodes not in C, then it is a fundamental 1-restricted odd cycle. So let us assume that G' has only one node a that is not in C. Thus the degrees of u, v and w in G' are 3.

Observe that C has three odd paths: from u to v; from u to w; and from v to z to w. If one of these paths has length more than one, then we have a fundamental 1-restricted odd cycle. To see this, suppose without loss of generality that the path from v to z to w has length more than 1. Arbitrarily choose one of the other paths, say the one from u to v, and consider the cycle C' that consists of au, av, and the path from u to v. The length of C' is either 1 *mod* 4 or 3 *mod* 4, and in either case we can construct a fundamental 1-restricted odd cycle based on it, a contradiction. Thus all three of our paths in C have length 1. It follows that the graph induced by C and the node a is K_4. It is now easy to see that either $G = K_4$ or we can construct a fundamental 1-restricted odd cycle based on it (since G is connected). □

LEMMA 15. *If a graph G contains no fundamental 1-restricted odd cycle, if G contains at most one odd-length ear in each maximal 2-connected subgraph, and if G is not a cycle of length 1 mod 4, then G^R is bipartite.*

Proof. By contradiction, suppose that the conclusion does not hold for some graph G; i.e., G^R contains an odd cycle C.

If C is a loop in G^R or the maximal 2-connected subgraph containing C has no odd-length ear in G or the maximal 2-connected subgraph has an odd-length ear but C does not contain the corresponding ear in G^R, then it is not hard to see that the cycle in G corresponding to C has length 2 *mod* 4, which implies that G has a fundamental 1-restricted odd cycle, a contradiction.

Now we assume that C contains an ear that corresponds to an odd-length ear in G. So this odd-length ear in G is an odd-length path in G.

Case 1: The odd-length path in G has length 1 *mod* 4.

So the remaining path in C has even length. We see that the path in G that corresponds to the remaining path in C has an even number of ears with length 2 *mod* 4. Therefore, the corresponding path in G has length 0 *mod* 4, and C corresponds to cycle of length 1 *mod* 4 in G. By assumption, this cycle cannot be the entire graph G; hence G contains a fundamental 1-restricted odd cycle with cycle C, a contradiction.

Case 2: The odd-length path in G has length 3 *mod* 4.

So the corresponding path in G^R has length 2, and hence the remaining path in C has odd length. It follows that the path in G that corresponds to the remaining path in C has an odd number of ears with length 2 *mod* 4. This implies that the corresponding path in G has length 2 *mod* 4, and C corresponds to a cycle of length 1 *mod* 4 in G, which again yields a fundamental 1-restricted odd cycle, a contradiction. ∎

LEMMA 16. *If $G = K_4$ or G contains at most one odd-length ear in each maximal 2-connected subgraph and G^R is bipartite, then G contains no fundamental 1-restricted odd cycle.*

Proof. If $G = K_4$, then we see by inspection that G contains no fundamental 1-restricted odd cycle. By contradiction, now suppose that $G \neq K_4$ and G contains a fundamental 1-restricted odd cycle G' with cycle C and G^R is bipartite.

Case 1: C has 0 *mod* 4 edges.

By definition, C has two petal edges whose attachment nodes in C are separated by two odd-length paths. Observe that each such odd-length path must contain an odd-length ear. Hence, G has at least two odd-length ears, a contradiction.

Case 2: C has 2 *mod* 4 edges.

Then C contains an even number of odd-length ears. If this number is greater than or equal to 2, then we get the contradiction. So let us assume that the number of odd-length ears in C is 0. Then C must contain an odd number of ears of length 2 *mod* 4; hence the cycle in G^R that corresponds to C is odd, a contradiction.

Case 3: C has 3 *mod* 4 edges.

Then C is the union of three odd-length paths between the attachment nodes. Therefore, C must contain at least three odd-length ears, a contradiction.

Case 4: C has 1 *mod* 4 edges.

Then C contains an odd number of odd-length ears. If the number of odd-length ears is greater than or equal to 3, then we get a contradiction. So we assume that C contains exactly one odd-length ear. If this odd-length ear is a cycle, then C corresponds to an odd cycle in G^R, a contradiction.

Suppose the odd-length ear is a path P. If P has length 1 *mod* 4, then $C \backslash P$ has length 0 *mod* 4 and hence contains an even number of ears of length 2 *mod* 4. Therefore, C corresponds to an odd cycle in G^R, a contradiction. If P has length 3 *mod* 4, then $C \backslash P$ has length 2 *mod* 4 and hence contains an odd number of ears of length 2 *mod* 4. Again, C corresponds to an odd cycle in G^R, a contradiction. ∎

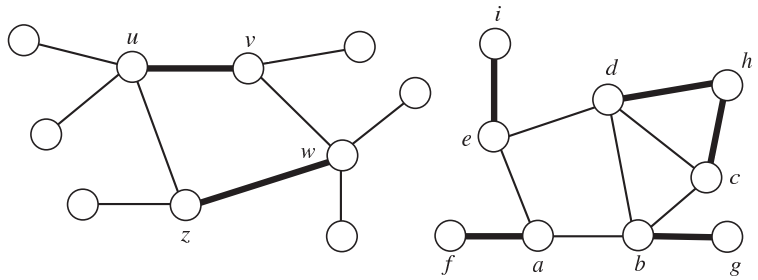


FIG. 5. *An example of a* 1*-restricted blossom.*

5. A class of blossom-type valid inequalities. In this section, we introduce a new class of inequalities, called the 1-restricted blossom inequalities, and we show they are valid for $R(G)$. These inequalities are similar to the classical blossoms both in structure and in that they have Chvátal rank of 1. We show that no fractional extreme solution to $A(G)$ satisfies all the 1-restricted blossom inequalities. However, we also show, with a small example, that the 1-restricted blossom inequalities together with inequalities (1–4) do not yield a complete description of $R(G)$. We begin with a few definitions.

A pair (B, T) is called a 1-*restricted blossom* of a graph G if the following two conditions are satisfied:

1. B is a set of nodes of G (called the *center* of the blossom).
2. T is an odd cardinality set of edges of G such that
   - Each edge in T is incident with one or two nodes of B; and
   - Each node of B is incident with exactly one edge in T.

   As a matter of notation:
   - Let T_i, for $i = 1, 2$, be the edges in T incident with exactly i nodes of B;
   - Let B_i, for $i = 1, 2$, be the nodes in B incident with an edge in T_i;
   - Let E^+ be the edges of T plus the edges with both ends in B_1; and
   - Let E^- be the edges with both ends in B_2, but not in T_2; plus the edges with one end in B_2 and one end not in B.

See Figure 5 for an example of a 1-restricted blossom. The pair (B, T) of the blossom given in Figure 5 is defined as follows:

$$\begin{aligned}
&T \equiv T_1 \cup T_2, \quad B \equiv B_1 \cup B_2,\\
&T_1 \equiv \{af, bg, ch, dh, ei\}, \quad T_2 \equiv \{uv, wz\},\\
&B_1 \equiv \{a, b, c, d, e\}, \quad B_2 \equiv \{u, v, w, z\},\\
&E^+ \equiv \{\text{all the edges of the right component plus } uv \text{ and } wz\},\\
&E^- \equiv \{\text{all the edges of the left component except } uv \text{ and } wz\}.
\end{aligned}$$

We associate the following inequality, called a 1-*restricted blossom inequality*, with each 1-restricted blossom (B, T):

$$\sum_{e \in E^+} x_e - \sum_{e \in E^-} x_e \leq |B_1| + \frac{|T| - 1}{2}. \tag{5}$$

Observe that when T contains a single edge, say uv, and $B = \{u, v\}$, then inequality (5) is the edge-adjacency constraint for uv. Furthermore, it is not hard to see that

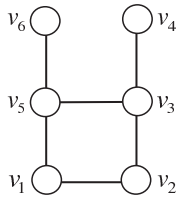


FIG. 6. *An example that has a facet-defining inequality with variable coefficient* 2.

(B,T) and the inequality (5) define a standard 2-matching blossom and inequality (see [5]) when $B_2, T_2 = \emptyset$ (which implies $E^- = \emptyset$). Hence the 1-restricted blossom inequalities are a common generalization of the edge-adjacency constraints and the standard 2-matching blossom inequalities.

PROPOSITION 17. *Any inequality* (5) *is valid for* $R(G)$.

Proof. Every 1-restricted simple 2-matching satisfies the following inequalities:

$$x(\delta(v)) \leq 2 \quad \forall\, v \in B_1, \tag{6}$$

$$x_e - x(adj(e)) \leq 0 \quad \forall\, e \in T_2, \tag{7}$$

$$x_e \leq 1 \quad \forall\, e \in T. \tag{8}$$

Adding up all the inequalities (6–8), dividing both sides by 2, and rounding down the variable coefficients and the right-hand side, we obtain inequality (5) that is valid for all the 1-restricted simple 2-matchings. □

PROPOSITION 18. *There is no fractional extreme point of* $A(G)$ *that satisfies all the inequalities* (5).

Proof. Consider an arbitrary fractional extreme point x for $A(G)$, as described in Theorem 1. Let C' be an odd cycle of G^c whose edges have value $\frac{1}{2}$ in x. Let C be the cycle in G that corresponds to C', and let M be the edges in G that have value 1 in x and are adjacent to one or two nodes of C. It is easy to see that C and M define a 1-restricted odd cycle in G as in Definition 5. If we let B denote the nodes of C and let $T = M$, then (B,T) is a 1-restricted blossom. It is easy to check that x satisfies all the inequalities (6–8) for (B,T) at equality. Since (5) is obtained by adding the inequalities (6–8) and rounding down, it follows that x violates (5) for (B,T). □

We next claim that the system (1–4, 5) is not sufficient to characterize $R(G)$ for all graphs G. To see this, first notice that the variable coefficients of every inequality in (1–4, 5) are 0, 1, or -1. To prove the claim we now present, for a specific graph G, a complete description of the facet-defining inequalities of $R(G)$, where one of the inequalities has a variable coefficient 2.

Consider the graph G consisting of a cycle of length 4 with two petals as shown in Figure 6. By Theorem 1, $A(G)$ has a unique fractional extreme point $f = (x_{v_1v_2}, x_{v_2v_3}, x_{v_3v_4}, x_{v_3v_5}, x_{v_5v_1}, x_{v_5v_6}) = (1, \frac{1}{2}, 1, \frac{1}{2}, \frac{1}{2}, 1)$. Now we show that we only need (1–4) and inequality

$$2x_{v_1v_2} - x_{v_2v_3} + x_{v_3v_4} - x_{v_5v_1} + x_{v_5v_6} \leq 2 \tag{9}$$

to give a complete description of $R(G)$.

First, we prove the validity of (9). Observe that G is a 1-restricted blossom (B,T), where $B = \{v_1, v_2, v_3, v_5\}$ and $T = \{v_1v_2, v_3v_4, v_5v_6\}$. The 1-restricted blossom inequality for (B,T) is

$$x_{v_1v_2} + x_{v_3v_4} + x_{v_3v_5} + x_{v_5v_6} \leq 3. \tag{10}$$

The inequalities (3) on edges v_1v_2, v_3v_4, and v_5v_6 are

$$x_{v_1v_2} - x_{v_2v_3} - x_{v_5v_1} \leq 0, \tag{11}$$

$$x_{v_3v_4} - x_{v_2v_3} - x_{v_3v_5} \leq 0, \tag{12}$$

$$x_{v_5v_6} - x_{v_3v_5} - x_{v_5v_1} \leq 0. \tag{13}$$

The inequality (2) on edge v_1v_2 is

$$x_{v_1v_2} \leq 1. \tag{14}$$

By doing $2 \times (10) + 2 \times (11) + (12) + (13) + 2 \times (14)$, we get a valid equality $6x_{v_1v_2} - 3x_{v_2v_3} + 3x_{v_3v_4} - 3x_{v_5v_1} + 3x_{v_5v_6} \leq 8$. Dividing both sides of this inequality by 3 and then rounding down the variable coefficients and right-hand side, we obtain the valid inequality (9) for $R(G)$. By a simple combinatorial argument, one can also show that all the 1-restricted simple 2-matchings of G satisfy this inequality.

Next, we prove that inequality (9) induces a facet that, together with (1–4), are enough to describe $R(G)$. Inequality (9) is facet-defining for $R(G)$, because it is satisfied as equality by the following linearly independent incidence vectors $(x_{v_1v_2}, x_{v_2v_3}, x_{v_3v_4}, x_{v_3v_5}, x_{v_5v_1}, x_{v_5v_6})$ of six 1-restricted simple 2-matchings of G:

$$x^1 = (1,0,0,0,1,1), \quad x^2 = (1,1,1,0,0,0), \quad x^3 = (1,0,1,1,1,0),$$
$$x^4 = (1,1,0,1,0,1), \quad x^5 = (1,1,1,0,1,1), \quad x^6 = (0,0,1,1,0,1).$$

Consider the following six inequalities from (1–4):

$$x_{v_1v_2} \leq 1, \tag{15}$$

$$x_{v_2v_3} + x_{v_3v_4} + x_{v_3v_5} \leq 2, \tag{16}$$

$$x_{v_3v_5} + x_{v_5v_1} + x_{v_5v_6} \leq 2, \tag{17}$$

$$x_{v_1v_2} - x_{v_2v_3} - x_{v_5v_1} \leq 0, \tag{18}$$

$$-x_{v_2v_3} + x_{v_3v_4} - x_{v_3v_5} \leq 0, \tag{19}$$

$$-x_{v_3v_5} - x_{v_5v_1} + x_{v_5v_6} \leq 0. \tag{20}$$

The equations from (15–20) are independent and give a unique solution $f = (1, \frac{1}{2}, 1, \frac{1}{2}, \frac{1}{2}, 1)$, which is the only fractional extreme point of $A(G)$. Observe that x^1, x^2, x^3, x^4, and x^5 satisfy (15) at equality, but x^6 does not; x^2, x^3, x^4, x^5, and x^6 satisfy (16) at equality, but x^1 does not; x^1, x^3, x^4, x^5, and x^6 satisfy (17) at equality, but x^2 does not; x^1, x^2, x^3, x^4, and x^6 satisfy (18) at equality, but x^5 does not; x^1, x^2, x^3, x^5, and x^6 satisfy (19) at equality, but x^4 does not; x^1, x^2, x^4, x^5 and x^5 satisfy (20) at equality, but x^3 does not. Hence, (15–20) are facet-defining inequalities of $A(G)$. Also observe that x^6 satisfies (16–20) at equality, but not (15); x^1 satisfies (15, 17–20) at equality, but not (16); x^2 satisfies (15–16, 18–20) at equality, but not (17); x^5 satisfies (15–17, 19–20) at equality, but not (18); x^4 satisfies (15–18, 20) at equality, but not (19); x^3 satisfies (15–19) at equality, but not (20). Geometrically, we see that the simplex defined by the convex hull of f, x^1, x^2, x^3, x^4, x^5, and x^6 is contained in $A(G)$, and the facets of the simplex are induced by inequalities (9, 15–20). Therefore, we conclude that inequalities (1–4, 9) give a complete description of $R(G)$.

6. Conclusion and future research. In this paper, we have presented a natural integer programming formulation for the problem of finding maximum weight

1-restricted simple 2-matchings, we have characterized the extreme points of the linear programming relaxation, and we have characterized when the relaxation has all integral extreme points. We have introduced a new class of inequalities that tightens this linear programming relaxation, but we know that this system is not sufficient to completely characterize the integral hull. So an obvious direction for future research is to look for a complete polyhedral description and a polynomial-time algorithm for the problem of finding maximum weight 1-restricted simple 2-matchings (or to show the problem is NP-hard). Another research direction would be to consider similar questions for the problem of finding maximum weight k-restricted simple 2-matchings, when $k > 1$.

REFERENCES

[1] H.-B. Belck, *Reguläre Faktoren von Graphen*, J. Reine Angew. Math., 188 (1950), pp. 228–252.

[2] G. Cornuéjols and W.R. Pulleyblank, *A matching problem with side conditions*, Discrete Math., 29 (1980), pp. 135–159.

[3] W.H. Cunningham, *Matching, matroids, and extensions*, Math. Program. B, 91 (2002), pp. 515–542.

[4] W.H. Cunningham and Y. Wang, *Restricted 2-factor polytopes*, Math. Program., 87 (2000), pp. 87–111.

[5] J. Edmonds, *Maximum matching and a polyhedron with* $0,1$*-vertices*, J. Research National Bureau of Standards Section B, 69 (1965), pp. 125–130.

[6] M.L. Fisher, G.L. Nemhauser, and L.A. Wolsey, *An analysis of approximations for finding a maximum weight Hamiltonian circuit*, Oper. Res., 27 (1979), pp. 799–809.

[7] A. Frank, *Restricted t-matchings in bipartite graphs*, Discrete Appl. Math., 131 (2003), pp. 337–346.

[8] H.N. Gabow, *An efficient reduction technique for degree-constrained subgraph and bidirected network flow problems*, in Proceedings of the Fifteenth Annual ACM Symposium on Theory of Computing, The Association for Computing Machinery, New York, 1983, pp. 448–456.

[9] T Gallai, *On factorisation of graphs*, Acta Math. Acad. Sci. Hung., 1 (1950), pp. 133–153.

[10] D. Hartvigsen, *Maximum cardinality* 1*-restricted simple* 2*-matching*, Electron. J. Combin., 14 (2007), #R73.

[11] D. Hartvigsen, *Extensions of Matching Theory*, Ph.D. thesis, 1984, Department of Mathematics, Carnegie-Mellon University, Pittsburgh, PA; under the supervision of G. Cornuéjols.

[12] D. Hartvigsen, *Finding maximum square-free* 2*-matchings in bipartite graphs*, J. Combin. Theory, Ser. B, 96 (2006), pp. 693–705.

[13] D. Hartvigsen, P. Hell, and J. Szabó, *The k-piece packing problem*, J. Graph Theory, 52 (2006), pp. 267–293.

[14] P. Hell, *Packing in graphs*, Electron. Notes Discrete Math., 5 (2000).

[15] M. Janata, M. Loebl, and J. Szabó, *A Gallai-Edmonds type theorem for the k-piece packing problem*, Electron. J. Combin., 12 (2005), #R8.

[16] E.L. Johnson, *Network Flows, Graphs and Integer Programming*, Ph.D. thesis, 1965, Operations Research Center, University of California, Berkeley, CA.

[17] A. Kaneko, *A necessary and sufficient condition for the existence of a path factor every component of which is a path of length at least two*, J. Combin. Theory Ser. B, 88 (2003), pp. 195–218.

[18] M. Kano, G.Y. Katona, and Z. Király, *Packing paths of length at least two*, Discrete Math., 283 (2005), pp. 129–135.

[19] Y. Nam, *Matching Theory: Subgraphs with Degree Constraints and other Properties*, Ph.D. thesis, 1994, Department of Mathematics, University of British Columbia, Canada; under the supervision of R.P. Anstee.

[20] M. Russell, *Restricted* 2*-factors*, Master's Thesis, 2001, Department of Mathematics, University of Waterloo; under the supervision of W.H. Cunningham

[21] A. Schrijver, *Combinatorial Optimization, Polyhedra and Efficiency*, Springer, New York, 2003.

[22] É. Tardos, *A strongly polynomial algorithm to solve combinatorial linear programs*, Oper. Res., 34 (1986), pp. 250–256.

[23] O. Vornberger, *Easy and hard cycle covers*, Preprint, Universitat Paderborn, 1980.

SIAM J. OPTIM.
Vol. 19, No. 3, pp. 1150–1170

PARALLEL SPACE DECOMPOSITION OF THE MESH ADAPTIVE DIRECT SEARCH ALGORITHM*

CHARLES AUDET†, J. E. DENNIS JR.‡, AND SÉBASTIEN LE DIGABEL†

Abstract. This paper describes a parallel space decomposition (PSD) technique for the mesh adaptive direct search (MADS) algorithm. MADS extends a generalized pattern search for constrained nonsmooth optimization problems. The objective of the present work is to obtain good solutions to larger problems than the ones typically solved by MADS. The new method (PSD-MADS) is an asynchronous parallel algorithm in which the processes solve problems over subsets of variables. The convergence analysis based on the Clarke calculus is essentially the same as for the MADS algorithm. A practical implementation is described, and some numerical results on problems with up to 500 variables illustrate the advantages and limitations of PSD-MADS.

Key words. parallel space decomposition, mesh adaptive direct search (MADS), asynchronous parallel algorithm, nonsmooth optimization, convergence analysis.

AMS subject classifications. 90C56, 90C30, 65K05, 68W10

DOI. 10.1137/070707518

1. Introduction.

This paper considers optimization problems of the form

$$(\mathcal{P}) \qquad \min_{x\in\Omega} f(x),$$

with the objective function $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$. Our motivation is to treat $\mathcal{P}$ when n grows large. The feasible region Ω is assumed to satisfy a nonsmooth constraint qualification, which we will discuss later, and we assume only the presence of an oracle to tell whether or not a given $x \in \mathbb{R}^n$ is feasible. We are concerned primarily with cases where $f(x)$ or the oracle are given by black-box computer simulations, which are assumed to evaluate in finite time. This is common in engineering design. Indeed, the reason we allow $f(x)$ to take on the value ∞ is that, for many such problems, no value of $f(x)$ is returned, even for some $x \in \Omega$, because of the internal workings of the simulation used to drive the design. See [2, 3, 10, 13, 21, 27, 32, 42].

There are other useful derivative-free direct search methods designed for problems similar to $\mathcal{P}$. These include the Nelder–Mead simplex [43], the DIRECT algorithm [20, 24, 30], frame-based methods [16, 44], the generalized pattern search (GPS) [7, 14, 49], the asynchronous parallel pattern search (APPS) approach [25, 29, 36, 34, 35], and the mesh adaptive direct search (MADS) [1, 8]. Related is the implicit filter method [31], though it does use a coarse difference gradient approximation. The reader may consult [31, 33, 37] for a survey of some of these direct search methods.

*Received by the editors November 6, 2007; accepted for publication (in revised form) May 28, 2008; published electronically October 31, 2008. Work of the first author was supported by FCAR grant NC72792 and NSERC grant 239436-05. The second author was supported by LANL 94895-001-04 34. Both were supported by AFOSR FA9550-07-1-0302, the Boeing Company, ExxonMobil Upstream Research Company, and the first and third authors were supported by the Consortium for Research and Innovation in Aerospace in Québec (CRIAQ).

http://www.siam.org/journals/siopt/19-3/70751.html

†GERAD and Département de mathématiques et de génie industriel, École Polytechnique de Montréal, C.P. 6079, Succ. Centre-ville, Montréal (Québec), H3C 3A7, Canada (Charles.Audet@gerad.ca, www.gerad.ca/Charles.Audet, Sebastien.Le.Digabel@gerad.ca).

‡Computational and Applied Mathematics Department, Rice University, Seattle, WA 98136 (dennis@rice.edu, http://www.caam.rice.edu/∼dennis).

Using these methods to solve expensive problems with more than a few dozen variables may be impractical, since they may need a large number of costly black-box evaluations. Dennis and Wu [18] reviewed different parallel methods for continuous optimization and concluded that a combination of GPS and the parallel variable distribution (PVD) of Ferris and Mangasarian [19] should be considered:

> "... parallel variable distribution and parallel direct searches seem an interesting pairing...."

The present paper is based on this remark.

PVD is an evolution of the block-Jacobi technique of [11], which optimizes in parallel a series of reduced subproblems on the subspaces of the original variables of $\mathcal{P}$. Dennis and Torczon [17] described a first parallel version of GPS, which evaluates the black-box in parallel and synchronizes at each iteration to compare solutions and update the current iterates. The APPS [25, 36], removes this synchronization barrier. In APPS, each process explores the space of variables using its own set of directions and does not wait for the other processes to terminate. APPS is expected to be more efficient than the synchronous version of [17], especially if the black-box has heterogeneous behavior that depends on the point where it is evaluated. A convergence analysis is presented in [35] for the smooth case.

Our work applies a decomposition of the variables of $\mathcal{P}$ based on the block-Jacobi technique of [11] that inspired the PVD method of [19]. This allows a natural parallel application of MADS to smaller subproblems, in an asynchronous way. The new algorithm, called PSD-MADS (parallel space decomposition-MADS, can be interpreted as a particular instance of MADS, thus inheriting the main results of the MADS convergence analysis. The paper focuses on the definition of the PSD-MADS frameworks and on its convergence analysis, and not on the choice of the subproblem variables. In our practical implementation of the algorithm, a simple random strategy is used, and it performs well.

The paper is divided as follows: section 2 gives an overview of the PSD and MADS methods. Section 3 presents the new asynchronous parallel algorithm PSD-MADS, and section 4 ensures that the main convergence results of MADS are maintained by showing that the entire PSD-MADS algorithm may be interpreted as a specific MADS instance. An implementation of PSD-MADS is described in section 5, with some numerical results on problems with a number of variables ranging from 20 to 500. Finally, section 6 gives some conclusions and proposes possible extensions of PSD-MADS.

2. Relevant literature. This section presents an overview of PSD methods. The MADS, its convergence analysis, and a practical implementation are also described in detail.

2.1. PSD methods. PSD methods decompose $\mathcal{P}$ into a finite number of smaller dimension subproblems, which can be solved in parallel with one process assigned to each subproblem.

Define $N = \{1, 2, \ldots, n\}$, where n is the number of variables of the optimization problem $\mathcal{P}$, and $Q = \{1, 2, \ldots, q\}$, where q is the number of available processes. Each process $p \in Q$ works on a nonempty subset $N_p \subseteq N$ of the variables. The other variables are fixed, based on the incumbent solution $x^* \in \Omega$, the current best known solution. More precisely, process $p \in Q$ works on the optimization subproblem

$$(\mathcal{P}_p(x^*)) \qquad \min_{x \in \Omega_p(x^*)} f(x),$$

with $\Omega_p(x^*) = \left\{x \in \Omega \ : x_i = x_i^* \ \forall i \in \overline{N}_p\right\}$ and $\overline{N}_p = N \setminus N_p$. The subproblem $\mathcal{P}_p(x^*)$ contains $n_p = |N_p|$ free variables, indexed by N_p. In section 5, we propose a simple and random strategy to build the subsets N_p.

The block-Jacobi method in [11] is an iterative two-step algorithm and may be described in a very general way as follows. At each iteration, the first step, the *parallelization*, consists of solving the subproblems in parallel, and the second step, the *synchronization*, gathers the subproblem solutions and constructs the next iterate. Similar methods are described in [26, 41, 50].

A variant of the method was introduced by Ferris and Mangasarian [19], as the PVD for a differentiable objective function f with continuous partial derivatives. In order to solve the subproblems more efficiently, the PVD method allows a priori fixed variables to change in a limited fashion, along directions typically based on ∇f. These variables are denoted as "forget-me-not" terms.

The convergence analysis in [19] requires that subproblems be solved to optimality. In the unconstrained case, if ∇f exists and is Lipschitz, then the accumulation points of the generated sequences are stationary points. In addition, if f is assumed to be convex, the convergence rate is shown to be linear. When Ω is nonempty, closed, convex, block-separable, and the functions defining it are also continuously differentiable, convergence results are still available. When there are general constraints, Ferris and Mangasarian recommend transforming the problem into unconstrained problems via penalty functions. This strategy is untested as far as we know, and we prefer to avoid estimating penalty constants.

These are parallel synchronous algorithms because the synchronization step waits for all of the processes to end. The conclusion of [19] states that an asynchronous version of the algorithm would increase efficiency. This is done in [40] for unconstrained problems, where the synchronization step is dropped at the expense of the convergence analysis.

The extensions of the PVD method are given in [45, 46, 47] with similar convergence results to those in [19] under less restrictive conditions. For example, subproblems do not need to be solved to complete optimality, as, for example, when one Newton-like iteration is used. A convergence analysis for the constrained case is given with either block-separability or convexity assumptions on the structure of Ω.

In the above references, no practical and generic strategy is given concerning the choice of the subproblem variables (sets N_p). However, the sets do need to form a partition of N, and they are fixed throughout the entire process. In the PSD [22] the subspaces can be chosen differently at each iteration.

Fukushima [23] extends the PVD method to a more general framework for unconstrained problems. The sets of subproblem variables are not fixed through the iterations and are not required to form a partition of N, but they must span N. In particular, an overlapping of the subproblem variables is allowed. Some experiments with such methods are given in [51].

More recently, the multidisciplinary optimization via adaptive response surfaces (MOVARS) algorithm [12] combines the GPS method with the synchronous PVD framework (including the "forget-me-not" terms from [19]) on fixed subsets N_p, but there is no convergence analysis.

In most of the references of this section, f is assumed to be at least differentiable, and constraints, if they are considered, are block-separable or convex. These are not reasonable assumptions for our target class of engineering design problems, and thus our convergence analysis does not rely on the analysis of [19] or its extensions. Rather, by incorporating MADS with its weaker hypotheses, we will inherit the MADS con-

vergence analysis. It will also give us greater flexibility concerning the way to handle constraints, the amount of work devoted to the subproblems, the lack of necessity for a synchronization step, and for the choice of the subsets N_p. Concerning this last issue, we remind the reader that we will not propose an elaborated strategy for this, as the focus of the paper is first to define the new method.

2.2. MADS. We now summarize the MADS algorithm [8] for problem $\mathcal{P}$, which extends the GPS algorithm for linearly constrained optimization [14, 49].

The constraints defining Ω are handled by the extreme barrier approach, as in [8, 38, 39]. This means that trial points outside Ω are simply rejected by setting their objective function value to ∞. Of course, this requires that the user provides a feasible initial point $x_0 \in \Omega$. We make the standard assumption that all of the trial points generated by the algorithm lie in a compact set.

MADS is an iterative algorithm where the black-box functions are evaluated at some trial points that are either accepted as new iterates because they are feasible and decrease the objective or are rejected.

All trial points generated by these algorithms are constructed to lie on a mesh

$$M(\Delta) \;=\; \left\{x + \Delta D z \;:\; x \in V,\; z \in \mathbb{N}^{n_D}\right\} \subset \mathbb{R}^n, \tag{1}$$

where the set V, called the *cache*, is a data structure memorizing all previously evaluated points so that no double evaluations occur, $\Delta \in \mathbb{R}^+$ represents a mesh size parameter, and D is an $n \times n_D$ matrix representing a fixed finite set of n_D directions in $\mathbb{R}^n$. More precisely, D is called the set of mesh directions and is chosen so that $D = GZ$, where G is a nonsingular $n \times n$ matrix and Z is an $n \times n_D$ integer matrix. The definition given by (1) differs slightly from the one in [8]. There the mesh was indexed by the iteration number instead of being parameterized by Δ. The reason for this difference is that our parallel algorithm will be working simultaneously on different size meshes originally generated at different iterations. Note also that in order to simplify the notation, the mesh size parameter Δ used here is the equivalent of Δ^m in [8].

Each iteration is divided into three steps: the search, the poll, and an update step determining the success of the iteration and producing the next iterate. The search and poll are treated specially in that the poll need not be carried out at an iteration if the search finds a better point. At each iteration, the algorithm attempts to generate an improved incumbent solution on the current mesh $M(\Delta_k)$, where Δ_k is the mesh size parameter at iteration k. The search step is very flexible and allows for trial points anywhere on the mesh. The way of generating these points is free of any rules, as long as they remain on the current mesh $M(\Delta_k)$ and that the search terminates in finite time. Some search strategies can be tailored for a specific application, while others are generic, such as the use of Latin hypercube sampling [48], or variable neighborhood search [4]. In summary, if one wants to define a MADS algorithm with a specific search, all that needs to be done to ensure convergence is to show that the search requires finite time and generates a finite number of trial points lying on the mesh.

The poll step explores the mesh $M(\Delta_k)$ near the current iterate x_k, and its rules ensure theoretical convergence of the algorithm. The way of choosing the directions used to generate the poll points is the difference between GPS and MADS. In GPS, the set of normalized potential poll directions must be chosen from a finite set that is fixed across all iterations. In MADS, the normalized directions may be chosen to be asymptotically dense in the unit sphere, which allows better coverage. We use the

MADS
[0] INITIALIZATION
$x_0 \in \Omega$, $\Delta_0 > 0$, $k \leftarrow 0$
[1] POLL AND SEARCH STEPS
objective: find a $y \in M(\Delta_k) \cap \Omega$ *such that* $f(y) < f(x_k)$
SEARCH STEP (optional)
evaluate the functions on a finite number of $M(\Delta_k)$ points
POLL STEP (optional if the search step succeeded)
generate n_{dir} MADS directions $d_i \in \mathbb{R}^n$
evaluate the functions on the MADS frame
$P_k = \{x_k + \Delta_k d_i : i = 1, 2, \ldots, n_{dir}\} \subseteq M(\Delta_k)$
[2] UPDATES
$x_{k+1} \leftarrow y$ (iteration success) OR x_k (iteration failure)
$\Delta_{k+1} \leftarrow \tau^{\omega_k} \Delta_k$ (reduced if iteration fails)
$k \leftarrow k + 1$
GOTO **[1]** IF no stopping condition is verified

FIG. 1. *High-level description of the* MADS *algorithm. The directions* d_i *are positive integer combinations of the columns of* D*. The search or poll steps can be stopped before all evaluations are terminated (opportunistic strategy).*

terminology of [16, 44] and say that at iteration k, the set of trial poll points is called the frame P_k. The set of directions used to construct P_k is denoted D_k, and it is not a subset of D.

In the last step of the kth iteration, the mesh size parameter is updated according to $\Delta_{k+1} \leftarrow \tau^{\omega_k} \Delta_k$, where $\tau > 1$ is a fixed rational number and ω_k is an integer that depends on the success of the iteration. When no improvement is made, the iteration is said to fail, and ω_k is taken to be an integer in the interval $[\omega^-; -1]$ with $\omega^- \leq -1$, forcing the next trial poll points to be closer to the current iterate. When a new best iterate is found, the iteration is said to succeed, and Δ_k is possibly increased with ω_k in $[0; \omega^+]$, with the integer $\omega^+ \geq 0$. Specific values for τ, ω^-, and ω^+ are suggested in section 2.4.

A high-level description of the algorithm is summarized in Figure 1. We encourage the reader to consult [8] for a complete description.

2.3. MADS convergence analysis. We will summarize the main convergence results for MADS given in [8]. These results assume that constraints are treated by the extreme barrier approach, and they constitute a hierarchical series of results relying on the Clarke calculus [15] for nonsmooth functions.

The main theorem is that, under a local Lipschitz assumption on f and under the assumption that the set of all normalized poll directions is dense in the unit sphere, the algorithm produces a Clarke stationary point. More precisely, MADS generates a point $\hat{x} \in \Omega$ at which the Clarke generalized directional derivatives of f in all of the directions in the Clarke tangent cone at $\hat{x}$ are nonnegative. The only assumptions needed are that f is Lipschitz near $\hat{x}$ and the constraint qualification that the hypertangent cone of Ω at $\hat{x}$ is nonempty. A corollary to this result in the unconstrained case is that if f is strictly differentiable near $\hat{x}$, then $\nabla f(\hat{x}) = 0$.

The convergence result that requires the least assumptions on f and Ω, the zeroth order result, is that MADS generates a limit point $\hat{x}$, which is the limit of mesh local minimizers on meshes that get infinitely fine. The notion of local optimality is, with respect to the current poll set, defined using a positive spanning set of directions.

More formally, MADS generates a convergent subsequence of iterates $\{x_k\}_{k\in K} \subset \Omega$ such that $x_k \to \hat{x}$, and $f(x_k) \leq f(x_k + \Delta_k d_k)$ for all directions d_k in a positive spanning set D_K, and $\|\Delta_k d_k\| \to 0$.

The price to pay for our new capability to handle a large number of variables is that this last convergence result will be lost. We will consider a MADS algorithm whose poll set contains a single element instead of being built using a positive spanning set of directions. We will refer to this as a *single-poll* MADS algorithm, and it still retains the property of generating asymptotically dense polling directions.

The next section discusses the LTMADS (lower-triangular MADS) implementation of the MADS algorithm. LTMADS uses positive bases to construct the poll sets. It is stated that the union of these normalized directions forms a dense set because if one looks closely at the proof in [8], one sees that it is the subset of single-poll normalized MADS directions that grows dense in the unit sphere. Thus, with the assumption of local Lipschitz continuity, the main convergence result guaranteeing a Clarke stationary point holds.

2.4. The LTMADS implementation of MADS. MADS is a general class of algorithms, where the search and poll steps need to satisfy certain conditions for the convergence results to hold. In particular, one of these conditions is that the total set of normalized poll directions used by the algorithm be dense in the unit sphere. In [8], after the definition of the MADS framework, a practical implementation is given. This implementation is named LTMADS, since it implies the random construction of a lower-triangular matrix. At this moment, LTMADS is the only published MADS implementation, and all MADS codes in section 5.2 correspond to LTMADS.

LTMADS fixes τ to 4, $\omega^- = -1$, $\omega^+ = 1$, and the set of mesh directions $D = [-I_n \; I_n]$, where I_n represents the $n \times n$ identity matrix. The mesh is based on the nonnegative integer value $\ell = -\log_4(\Delta_k)$, $\Delta_k = 4^{-\ell}$, and directions are constructed randomly using a lower-triangular matrix. One of these directions is a special case and is fixed for each value of ℓ. This direction, called $b(\ell)$, has one coordinate (the largest in absolute value) set to $\pm 2^\ell$ so that poll points are within $\sqrt{\Delta_k}$ of the poll center x_k in the ℓ_∞ norm.

The result stated in [6, 8] is that with probability one, the series of normalized directions $b(\ell)$ grows dense in the unit sphere. In LTMADS, the direction $b(\ell)$ is augmented at each iteration with other directions to form a positive spanning set of polling directions. We can, as explained in the preceding section, construct a single-poll MADS algorithm with dense polling directions using only the $b(\ell)$ directions, but the zeroth order convergence result of MADS is lost. Also, because we are not polling at each iteration in a positive spanning set of directions, the mesh size might drop too quickly with this single-poll version of MADS, and so the search step is of extra importance. This is the key to the PSD-MADS algorithm described in the next section: one process executes a single-poll MADS algorithm, while the work of the other processes may be interpreted as a search step.

3. PSD of MADS (PSD-MADS). This section describes the combination of MADS with a PSD method. The resulting algorithm is called PSD-MADS. It is an asynchronous parallel algorithm where a master process decides on the subsets $N_p \subseteq N$ and assigns the resulting optimization subproblems $\mathcal{P}_p(x^*)$ to slaves. The slaves apply MADS to attempt to improve the incumbent solution x^*. No synchronization step is performed. When a slave completes its assigned task, the master assigns a new subproblem with a possibly new N_p and x^*.

3.1. General description of PSD-MADS. Although PSD-MADS is an asynchronous parallel algorithm, the notion of iteration is kept, and it corresponds to two successive calls by the master to one special slave, called the *pollster* slave, described more precisely in section 3.2. The pollster slave executes a single-poll MADS algorithm on the entire problem $\mathcal{P}$, while the other slaves, called the *regular* slaves, work on the subproblems $\mathcal{P}_p(x^*)$. This task partition between the pollster and the regular slaves allows the convergence analysis of section 4, where it is shown that the pollster slave executes a valid MADS algorithm, thus inheriting the convergence results of [8]. Note that the pollster slave's task requires the fewest function values of any of the poll steps.

Each subproblem $\mathcal{P}_p(x^*)$ is a subproblem of $\mathcal{P}$ with a reduced number of variables indexed by the set N_p. When an optimization process terminates, the slave communicates its progress to the master. If it has found an improved solution, then that becomes the new incumbent solution. The slave immediately starts work on a new subproblem assigned by the master. There is no need to synchronize all of the slaves.

With several MADS instances executing in parallel, it is necessary to define different mesh size parameters. First, Δ_j^p corresponds to the mesh $M(\Delta_j^p)$ used at iteration j of the MADS algorithm performed by a regular slave s_p. The mesh size parameter is denoted differently for the pollster slave, with Δ_k^1 (notice the same iteration counter k used both for the pollster slave and PSD-MADS). The number Δ_k^1 is called the *pollster mesh size parameter* at iteration k of PSD-MADS. Finally, an additional mesh size parameter Δ_k^M is called the *master mesh size parameter*. The mesh $M(\Delta_k^M)$ is never used explicitly, but it is useful for comparing the two other meshes $M(\Delta_k^1)$ and $M(\Delta_j^p)$. At iteration k of PSD-MADS and at iteration j of the MADS algorithm performed on a subproblem $\mathcal{P}_p(x^*)$ by a regular slave s_p for $p \in \{2, 3, \ldots, q-2\}$, the PSD-MADS construction ensures that

$$\Delta_k^1 \leq \Delta_k^M \leq \Delta_j^p. \tag{2}$$

Inequalities (2) are formally proved in the convergence analysis of section 4, where PSD-MADS is interpreted as a valid single-poll MADS instance performed by the pollster slave. An additional hypothesis on the different meshes $M(\Delta_k^M)$, $M(\Delta_k^1)$, and $M(\Delta_j^p)$ is necessary.

Hypothesis 3.1. If two mesh size parameters Δ and Δ' satisfy $\Delta = \tau^\omega \Delta'$, where $\omega \in \mathbb{N}$, then $M(\Delta) \subseteq M(\Delta')$.

This assumption holds for the PSD-MADS implementation given in section 5.

The q processes are partitioned into a master, $q-2$ slaves, and a cache server (process number $q-1$), which memorizes all points that have been evaluated. The $q-2$ slaves include the pollster slave (process number 1) and $q-3$ regular slaves. The notation s_p, with $p \in Q \setminus \{q-1, q\}$, is used to identify the $q-2$ processes assigned as slaves, and $Q_{reg} = \{2, 3, \ldots, q-2\}$ is the set of the indices of the $q-3$ regular slaves. The qth process is used as the master, which defines the lower-dimensional subproblems $\mathcal{P}_p(x^*)$ and communicates them to the slaves.

An advantage of applying the PSD method to MADS instead of another optimization method is that most of the conditions necessary for convergence in the other PSD methods mentioned in section 2.1 can be relaxed (the smoothness of the functions, the conditions on the constraints, no synchronization step, and no restrictions on the choice of the sets N_p).

This new algorithm is not a particular case of the method in [23], which generalizes many parallel variable decomposition methods, since general constraints are allowed,

POLLSTER ($p = 1$)
Inputs : *pollster mesh size* Δ_k^1
starting point x_0
Output : *pollster solution* x_p
solve problem $\mathcal{P}$: MADS(pollster)
terminate after a single evaluation
send x_p to master

FIG. 2. *Pseudocode for pollster slave.* MADS*(pollster) considers all n variables with a single-poll direction and terminates after one iteration.*

and f is not assumed to be smooth. PSD-MADS also differs from the recent MOVARS algorithm [12], which does require N_p to partition the variables, because it provides a convergence analysis, it dynamically changes the sets N_p, and it is an asynchronous parallel method. The next sections describe precisely the role of each process.

3.2. The pollster slave s_1, on $M(\Delta_k^1)$. The pollster slave s_1 has a special role; its set of variables is always fixed to $N_1 = N$, so that it works on the original problem $\mathcal{P}$. Due to its greater impact on the algorithm and to distinguish s_1 from the other slaves, we call it the pollster slave, or, simply, the pollster.

To reduce the expected high number of evaluations done by the successive pollster instances, a single-poll MADS algorithm is used (the poll directions are reduced to a single element), with the conditions that the union of all of the normalized directions used throughout the algorithm are dense in the unit sphere and that the norms of those directions are in proper relation with the mesh size parameter.

Moreover, the pollster is limited to only one MADS iteration, with no search step and one poll step. It follows that, at most, one function evaluation will be performed (zero function evaluation if the unique poll trial point is found in the cache), and the pollster mesh size parameter Δ_k^1 will not be updated (this is done by the master).

The notation MADS(pollster) or MADS(s_1) refers to the single-poll MADS algorithm performed by the pollster. MADS(pollster) is defined so that its mesh size parameter Δ_k^1 cannot be larger than the master mesh size Δ_k^M at iteration k of PSD-MADS (see (2)).

The pollster pseudocode is shown in Figure 2. The pollster mesh size is updated by the master. The best obtained solution corresponds to x_p, which is sent to the master. The convergence analysis in section 4 is based on the pollster and on the fact that consecutive runs of MADS(s_1) form a valid single-poll MADS instance on $\mathcal{P}$.

3.3. The regular slaves s_2 to s_{q-2}, on $M(\Delta_j^p)$. The regular slaves s_p, $p \in Q_{reg}$ work on subsets N_p of N and use the positive spanning sets of directions. The MADS algorithm working on problem $\mathcal{P}_p(x^*)$ and performed by slave s_p is designated by MADS(s_p).

Subproblem $\mathcal{P}_p(x^*)$ is defined as an $|N_p|$-variable problem, since all of the variables in $N \setminus N_p$ are fixed. Trial points generated by MADS(s_p) are then in $\mathbb{R}^n$, with some coordinates fixed. The values of these fixed coordinates are directly taken from the starting point for MADS(s_p), i.e., x^*, the incumbent solution. The user supplies a parameter $bbe_{\max} > 0$ that indicates the maximum allowed number of black-box calls for the application of MADS to the optimization of a subproblem.

The pseudocode for the regular slaves is shown in Figure 3. MADS(s_p) generates the trial points on meshes of sizes Δ_j^p, where j is the iteration counter of the subprob-

SLAVE s_p ($p \in Q_{reg}$)
Inputs : *initial mesh size* Δ_0^p
minimum mesh size $\Delta_{\min}^p$
starting point x_0
set of variables N_p
Outputs : *slave solution* x_p
final mesh size Δ_{stop}^p
solve subproblem $\mathcal{P}_p(x^*)$: MADS(s_p)
terminate when $\Delta_j^p < \Delta_{\min}^p$ OR after $bbe_{\max}$ evaluations
send x_p and Δ_{stop}^p to master

FIG. 3. *Pseudocode for slaves processes. Does not include pollster slave, which is specifically described in Figure* 2.

lem algorithm. The initial mesh size Δ_0^p for MADS(s_p) is set by the master. The value of the parameter $\Delta_{\min}^p$ also is supplied by the master and equals Δ_k^M, where k is the PSD-MADS iteration at which MADS(s_p) started. Finally, we impose that no mesh size for MADS(s_p), $p \in Q_{reg}$ exceeds the PSD-MADS initial mesh size Δ_0^{user} provided by the user. MADS(s_p) terminates if $bbe_{\max}$ evaluations are made, or if a minimal mesh size $\Delta_{\min}^p$ is reached. The final mesh size (Δ_{stop}^p) and the best solution found (x_p) are sent to the master.

The union of all regular slaves MADS(s_p) instances is interpreted as a search step for the total problem single-poll MADS algorithm. This is important to the convergence analysis in section 4.

3.4. The cache server—$(q-1)$th process. The cache server is a specialized process that simply memorizes all evaluated points. Each time a process generates a trial point, the cache server is interrogated. This is done to avoid unnecessary expensive function evaluations in case this point has already been evaluated. The cache server provides the global availability of any improvement made by any slave. This is interpreted in section 5 as a search step (the *cache search*) by the regular slaves on their subproblems.

3.5. The master—qth process. The master process coordinates the work of the $q-2$ slaves. It waits for slave results, updates data, and assigns work to slaves. It evaluates only the black-box functions at the starting point x_0.

The master process provides the master mesh size Δ_k^M at iteration k of PSD-MADS, which is the link between the mesh sizes Δ_k^1 and Δ_j^p on which the different MADS(s_p), $p \in Q_{reg}$ work. The initial master mesh size $\Delta_0^M = \Delta_0^{user}$ is set by the user.

The master process updates the pollster mesh size Δ_k^1, after a pollster instance terminates. If no improvement is made by any slave s_1 to s_{q-1} during iteration k, the iteration is a failure, and the pollster mesh size is reduced. If the iteration succeeds, then the pollster mesh size is increased. In all cases, the pollster mesh size is smaller than the master mesh size (2). The value of the pollster mesh size is also kept less than or equal to Δ_0^{user}.

For all regular slaves s_2 to s_{q-2}, the minimal mesh size $\Delta_{\min}^p$ is set to the current value of Δ_k^M. This, as is explained in more detail in the convergence analysis, leads to the fact that, at iteration k of PSD-MADS, no regular slave can generate trial points

MASTER
[0] INITIALIZATION
$x^* \leftarrow x_0 \in \Omega$, $\Delta_0^1 \leftarrow \Delta_0^M \leftarrow \Delta_0^{user} > 0$, $k \leftarrow 0$, $\omega^- \leq -1$, $\omega^+ \geq 0$
start MADS(pollster) with (Δ_0^{user}, x_0) (Figure 2)
FOR ALL ($p \in Q_{reg}$)
construct N_p and set $\Delta_{\min}^p \leftarrow \Delta_0^M$
start MADS(s_p) with (Δ_0^{user}, $\Delta_{\min}^p$, x_0, N_p) (Figure 3)
[1] ITERATIONS
given values from a slave s_p (x_p, Δ_{stop}^p)
IF $\big(f(x_p) < f(x^*)\big)$ (success)
$x^* \leftarrow x_p$
IF $(p = 1)$ (pollster, Δ_{stop}^p corresponds to Δ_k^1)
$\Delta_{k+1}^M \leftarrow \tau^{\alpha_k}\Delta_k^1 \leq \min_{p \in Q_{reg}} \Delta_{\min}^p$, with $\alpha_k \in [0; \omega^+]$
$\Delta_{k+1}^1 \leftarrow \tau^{\omega_k}\Delta_k^1$ (Figure 5)
$k \leftarrow k + 1$
start MADS(pollster) with (Δ_k^1, x^*) (Figure 2)
ELSE (regular slave)
construct N_p
$\Delta_{\min}^p \leftarrow \Delta_k^M$
$\Delta_0^p \leftarrow \tau^\gamma \Delta_{stop}^p$, with $\gamma \in \mathbb{Z}$ and so that $\Delta_k^M \leq \Delta_0^p \leq \Delta_0^{user}$
start MADS(s_p) with (Δ_0^p, $\Delta_{\min}^p$, x^*, N_p) (Figure 3)
GOTO **[1]** IF no stopping condition is verified

FIG. 4. *Pseudocode for master process. Δ_k^M and Δ_k^1 are the master and pollster mesh sizes at iteration k, and Δ_{stop}^p is the last mesh size of a slave s_p. If $p = 1$, $\Delta_{stop}^p = \Delta_k^1 \leq \Delta_k^M$, else $\Delta_{stop}^p \geq \Delta_k^M$. The master evaluates the black-box just once for x_0.*

POLLSTER MESH SIZE UPDATE $\Delta_{k+1}^1 \leftarrow \tau^{\omega_k}\Delta_k^1$
IF (iteration success)
$\omega_k = \alpha_k \in [0; \omega^+]$ $\big(\Delta_{k+1}^1 \leftarrow \Delta_{k+1}^M\big)$
$\big($pollster mesh size increase, $\Delta_{k+1}^1 \geq \Delta_k^1\big)$
ELSE
$\omega_k \in [\omega^-; -1]$
$\big($pollster mesh size decrease, $\Delta_{k+1}^1 < \Delta_k^1\big)$

FIG. 5. *An update of the next pollster mesh size Δ_{k+1}^1. In any case, the pollster mesh size verifies $\Delta_k^1 \leq \Delta_k^M$.*

on meshes finer than $M(\Delta_k^M)$ and that all of the slaves work, in fact, on the pollster mesh of size Δ_k^1.

The master process pseudocode is described in Figure 4, and the pollster mesh size update is detailed in Figure 5. The pseudocode for the master process implies that, when the master mesh size is updated, it is always possible to find an integer $\alpha_k \in [0; w^+]$ such that $\tau^{\alpha_k}\Delta_k^1 \leq \min_{p \in Q_{reg}} \Delta_{\min}^p$. The next proposition shows that $\alpha_k = 0$ is always a candidate.

PROPOSITION 3.2. *At iteration k of the* PSD-MADS *algorithm, there exists a nonnegative integer α_k such that $\tau^{\alpha_k}\Delta_k^1 \leq \min_{p \in Q_{reg}} \Delta_{\min}^p$.*

APPARENT POLLSTER
[0] INITIALIZATION
$x_0 \in \Omega$, $\Delta_0^M \leftarrow \Delta_0^1 \leftarrow \Delta_0^{user} > 0$, $k \leftarrow 0$
[1] POLL AND SEARCH STEPS
SEARCH STEP (by other slaves, opportunistic)
ask cache server for $x_s \in M(\Delta_k^M) \subseteq M(\Delta_k^1)$
SINGLE-POLL STEP
construct and evaluate $P_k = \{x_{poll}\} \subseteq M(\Delta_k^1)$
[2] UPDATES
determine type of success of iteration k
$\Delta_{k+1}^1 \leftarrow \tau^{\omega_k}\Delta_k^1$ (cannot be larger than Δ_{k+1}^M)
$x_{k+1} \leftarrow (x_s$ or x_{poll} or $x_k)$
$k \leftarrow k+1$
GOTO **[1]** IF no stopping condition is verified

FIG. 6. *A detailed pseudocode of the apparent pollster algorithm, the algorithm from the point of view of the pollster slave. At every moment, a finite number of $M(\Delta_k^1)$ points are evaluated in parallel by other slaves. These evaluations are considered within the opportunistic search step. Δ_k^M is updated by the master after the poll step.*

Proof. At iteration 0, $\Delta_0^1 = \Delta_0^M = \Delta_0^{user} = \min_{p\in Q_{reg}} \Delta_{\min}^p$ so $\alpha_0 = 0$, and therefore it exists. Then $\Delta_1^M = \Delta_0^{user}$ and $\min_{p\in Q_{reg}} \Delta_{\min}^p$ at iteration 1 is equal to Δ_0^{user}. Figure 5 ensures that Δ_1^1 is bounded above by Δ_0^{user}, and therefore $\alpha_1 = 0$ is a possible value.

Suppose, by way of induction, that, for some $k \geq 2$, the proposition is true at iteration $k-1$. It follows that $\Delta_k^M = \tau^{\alpha_{k-1}}\Delta_{k-1}^1 \leq \min_{p\in Q_{reg}} \Delta_{\min}^p$, and as it corresponds to new values for $\Delta_{\min}^p$, $p \in Q_{reg}$, and the new smaller possible value of $\min_{p\in Q_{reg}} \Delta_{\min}^p$ at iteration k remains Δ_k^M. The largest value that Δ_k^1 may take is also Δ_k^M, which shows $\alpha_k = 0$ validates the result. □

This proof allows all values of α_k to be zero, but, in practice, nonzero values are likely. For example, if iteration 1 failed and $\Delta_1^1 = \Delta_0^{user}$, then the following mesh updates are possible: $\Delta_2^M \leftarrow \Delta_0^{user}$ ($\alpha_1 = 0$) and $\Delta_2^1 \leftarrow \Delta_0^{user}/4$. $\min_{p\in Q_{reg}} \Delta_{\min}^p$ is still equal to Δ_0^{user} at iteration 2, and so α_2 can be either 0 or 1.

4. Convergence analysis of PSD-MADS. It is shown here that the entire algorithm may be interpreted as a single-poll MADS algorithm applied to the original problem $\mathcal{P}$ and that conditions are met so that the main convergence results from [8] hold. These conditions are that the regular slaves generate a finite number of trial points lying on the pollster mesh and that all of these trial points can be interpreted as a search step with the pollster slave providing the poll step. This is detailed in Figure 6, and we refer to it as the *apparent pollster algorithm*. This algorithm is another way of interpreting the PSD-MADS algorithm described by the pseudocodes in Figures 2, 3, 4, and 5. Iteration k of the apparent pollster algorithm corresponds to the iteration k of PSD-MADS (used by the master process), and the notions of iteration success and failure remain the same.

The convergence analysis in this section proves that the apparent pollster algorithm is a single-poll MADS algorithm with the following components:

- A search step performed by regular slaves $s_2, s_3, \ldots, s_{q-2}$ on meshes of coarseness larger than or equal to Δ_k^M;

- A poll step at iteration k (the same k used by the master process in Figure 4) performed by one call to the pollster slave s_1 on a mesh of size $\Delta_k^1 \leq \Delta_k^M$;
- A mesh update performed by the master process with $\Delta_{k+1}^1 \leftarrow \tau^{\omega_k}\Delta_k^1$ and the integer $\omega_k \in \begin{cases} [0;\omega^+] & \text{iteration success,} \\ [\omega^-;-1] & \text{iteration failure.} \end{cases}$

The master mesh size parameter Δ_k^M at iteration k is the link described by inequalities (2) between the mesh size of MADS(pollster) and the different mesh sizes of MADS(s_p). It is updated by the master with the MADS(pollster) mesh (via $\Delta_{stop}^p = \Delta_k^1$), in such a way that, at every iteration k of the apparent pollster algorithm, Δ_k^1 satisfies $\Delta_k^1 \leq \Delta_k^M$. This Δ_k^M update by the master in the apparent pollster algorithm occurs when the mesh size Δ_k^1 is updated, and while its value does not change during the poll step, it can possibly be updated during the search step, since that is performed in parallel. This possible change of the Δ_k^M value within the search step of the apparent pollster algorithm is governed by the fact that Δ_k^M cannot be exceeded by any regular slave mesh size ($\Delta_k^M \leq \min_{p\in Q_{reg}} \Delta_{\min}^p$).

To show that the apparent pollster algorithm is a valid single-poll MADS algorithm applied to the original problem $\mathcal{P}$ and that the convergence conditions of [8] hold, the search trial points, whose evaluations are performed at any time in parallel by the other slaves, must remain finite in number and on the current pollster mesh at iteration k, Δ_k^1. This will be shown via the following propositions.

PROPOSITION 4.1. *The mesh size parameter at iteration j of the* MADS *algorithm performed by a slave s_p, $p \in Q_{reg}$ on a subproblem $\mathcal{P}_p(x^*)$ satisfies $\Delta_j^p = \tau^{-\eta_j}\Delta_0^{user}$ for some integer $\eta_j \geq 0$. This can be extended to the pollster slave at iteration k, with $\Delta_k^1 = \tau^{-\eta_k}\Delta_0^{user}$.*

Proof. We first show that the proposition is true for the first optimization subproblem solved by a regular slave s_p, $p \in Q_{reg}$. The initial mesh size parameter used for this MADS instance is Δ_0^{user}, and with the standard MADS mesh update rules, at iteration j, $\Delta_j^p = \tau^{\omega_{j-1}}\Delta_{j-1}^p = \cdots = \tau^{\sum_{i=0}^{j-1}\omega_i}\Delta_0^{user}$. Then $\eta_j = -\sum_{i=0}^{j-1}\omega_i \geq 0$, because no mesh size can be larger than Δ_0^{user}.

Suppose now that the proposition is true for the rth MADS instance performed by s_p. In particular, the last mesh size parameter of this instance can be written $\Delta_{stop}^p = \tau^{-\eta_{stop}}\Delta_0^{user}$, where η_{stop} is a nonnegative integer. From the algorithm described in Figure 4, the first mesh size parameter of the $(r+1)$th MADS instance performed by s_p is $\Delta_0^p = \tau^\gamma\Delta_{stop}^p$, with $\gamma \in \mathbb{Z}$. Then, at iteration j of the $(r+1)$th instance, $\Delta_j^p = \tau^{\sum_{i=0}^{j-1}\omega_i}\Delta_0^p$ and $\eta_j = -\sum_{i=0}^{j-1}\omega_i - \gamma + \eta_{stop} \geq 0$ because $\Delta_j^p \leq \Delta_0^{user}$. The proposition can be extended to the pollster slave with the same induction proof on k. ∎

PROPOSITION 4.2. *At iteration k of* PSD-MADS*, and at iteration j of the* MADS *algorithm performed by s_p ($p \in Q_{reg}$) on a subproblem $\mathcal{P}_p(x^*)$, there exists a nonnegative integer β_j such that $\Delta_j^p = \tau^{\beta_j}\Delta_k^M$.*

Proof. From the algorithm in Figure 4, the master mesh size parameter, at iteration k of PSD-MADS, can be written $\Delta_k^M = \tau^{\alpha_k - 1}\Delta_{k-1}^1$, with $\alpha_{k-1} \in \mathbb{N}$, and $\Delta_{k-1}^1 = \tau^{-\eta_{k-1}}\Delta_0^{user}$, with $\eta_{k-1} \in \mathbb{N}$, from Proposition 4.1. From the same proposition, the mesh size parameter at iteration j of MADS(s_p), $p \in Q_{reg}$ can be written $\Delta_j^p = \tau^{-\eta_j}\Delta_0^{user}$, $\eta_j \in \mathbb{N}$. Then, $\Delta_j^p = \tau^{\beta_j}\Delta_k^M$, with $\beta_j = \eta_{k-1} - \eta_j - \alpha_{k-1}$. The minimal mesh size parameter $\Delta_{\min}^p$ considered by MADS(s_p) corresponds to Δ_i^M, where $i \leq k$ is an anterior iteration of PSD-MADS. The current value of Δ_k^M was chosen to be smaller than $\min_{p\in Q_{reg}} \Delta_{\min}^p \leq \Delta_i^M$. Then, $\Delta_k^M \leq \Delta_i^M \leq \Delta_j^p$, and β_j is a nonnegative integer. ∎

An immediate corollary, with Hypothesis 3.1, is that at iterations k of PSD-MADS and j of MADS(s_p), $p \in Q_{reg}$, $M(\Delta_j^p) \subseteq M(\Delta_k^M)$.

PROPOSITION 4.3. *At iteration k of* PSD-MADS*, every trial point generated by the* MADS *algorithm performed by s_p, $p \in Q_{reg}$ on any subproblem $\mathcal{P}_p(x^*)$, lies on the pollster mesh $M(\Delta_k^1)$.*

Proof. From the algorithm in Figure 4, the pollster and master mesh size parameters at iteration k of PSD-MADS are linked with $\Delta_k^M = \tau^{\alpha_k}\Delta_k^1$, $\alpha_k \in \mathbb{N}$. With Hypothesis 3.1 and Proposition 4.2, at iteration j of MADS(s_p), $M(\Delta_j^p) \subseteq M(\Delta_k^M) \subseteq M(\Delta_k^1)$. Since all MADS($s_p$) trial points are constructed on $M(\Delta_j^p)$, they also lie on $M(\Delta_k^1)$. ∎

This series of propositions ensures that all of the trial points of the search step of the apparent pollster at iteration k, performed in parallel by regular slaves, lie on the current pollster mesh Δ_k^1. In addition, their number remains finite as the time between two iterations, corresponding to a single-point poll, is finite (with the hypothesis that the black-box evaluates or is terminated to return ∞, in finite time). The PSD-MADS algorithm, viewed from the perspective of the pollster slave, thus executes a valid single-poll MADS search, and the main convergence results of [8] remain valid. Let $\hat{x}$ be the limit of a subsequence of PSD-MADS incumbents at unsuccessful iterations. Then

- If f is Lipschitz near $\hat{x} \in \Omega$, then the Clarke derivative satisfies $f^\circ(\hat{x}; v) \geq 0$ for all $v \in T_\Omega^H(\hat{x})$, the hypertangent cone to Ω at $\hat{x}$;
- In the unconstrained case and if f is strictly differentiable at $\hat{x}$, $\nabla f(\hat{x}) = 0$.

As mentioned in section 2.3, the fact that the single-poll version of MADS is used sacrifices the zeroth order result of [8], i.e., $\hat{x}$ cannot be said to be the limit of local optima on meshes that get infinitely fine.

5. A practical implementation of PSD-MADS. This section proposes a practical implementation of the PSD-MADS algorithm described in section 3 based on the LTMADS implementation proposed in [8] and summarized in section 2.4. Numerical tests complete the implementation description.

5.1. PSD-MADS implementation.

Verification of Hypothesis 3.1. The above convergence analysis relies on Hypothesis 3.1. An easy way to satisfy this hypothesis is to simply choose τ to be an integer. Indeed, consider the mesh point $x \in M(\Delta)$ and mesh size $\Delta \in \mathbb{R}$. From the mesh definition (1), x can be written as $y + \Delta \sum_{i=1}^{n_D} z_i d_i$, where y belongs to V, the set of currently evaluated points, and z_i are nonnegative integers. Now, if $\Delta' = \tau^\omega \Delta$, where $\omega \in \mathbb{N}$ and $1 \leq \tau \in \mathbb{N}$, then x can be rewritten as $x = y + \Delta' \sum_{i=1}^{n_D} \tau^\omega z_i d_i$. It follows that $\tau^\omega z_i \in \mathbb{N}$, $i = 1, 2, \ldots, n_D$, and therefore $x \in M(\Delta')$. We have shown that $M(\Delta) \subseteq M(\Delta')$, and thus Hypothesis 3.1 is satisfied. In the proposed PSD-MADS implementation, the LTMADS fixed value of $\tau = 4$ is used.

Directions used by the pollster. The LTMADS direction $b(\ell)$ is used in the single-poll MADS algorithm executed by the pollster slave. The union of normalized directions $b(\ell)$, $\ell = 1, 2, \ldots$, is dense in the unit sphere with probability one, and MADS(pollster) with the $b(\ell)$ direction respects the conditions for a valid single-poll MADS algorithm.

Sets N_p of subproblem variables. This paper does not focus on the choice of the subproblem variables. Rather, we use this very simple strategy: let the sets N_p, $p \in Q_{reg} = \{2, 3, \ldots, q-2\}$, be randomly generated by the master using a

uniform distribution before each subproblem parameter set is sent to a regular slave process. In order to keep an easy parametrization of this PSD-MADS implementation, the number of variables for each subproblem is fixed throughout the entire algorithm $|N_2| = |N_3| = \cdots = |N_{q-2}| = ns$, where ns is a parameter chosen by the user (recall that, for the pollster, $N_1 = N$). Notice also that ns is not required to satisfy $(q-3)ns \geq N$. Furthermore, when MADS(s_p), $p \in Q_{reg}$ succeeds in improving the incumbent, the same set N_p is kept for the next run performed by the slave s_p.

Mesh update rules. The mesh directions of definition (1) are the standard LTMADS $2n$ directions $D = [-I_n \; I_n]$. The following mesh size parameter updates are in accordance with the LTMADS mesh update rules:

- **Regular slaves mesh size Δ_j^p (at iteration j of MADS(s_p), $p \in Q_{reg}$):** After an iteration fails, the mesh size is updated with $\Delta_{j+1}^p \leftarrow \Delta_j^p/4$ ($\omega_j = -1$ in Figure 1). If a poll step is successful, $\Delta_{j+1}^p \leftarrow 4\Delta_j^p$ ($\omega_j = 1$). In the next search step, if a successful point is found in the cache server, set $\Delta_{j+1}^p \leftarrow 4\Delta_{cache}$, where Δ_{cache} is the mesh size used to generate this point. Equation (3) summarizes these updates as follows:

$$\Delta_{j+1}^p \leftarrow \begin{cases} \min\left\{\Delta_0^{user}, 4\Delta_j^p\right\} & \text{poll success,} \\ \min\left\{\Delta_0^{user}, 4\Delta_{cache}\right\} & \text{cache search success,} \\ \Delta_j^p/4 & \text{iteration failure.} \end{cases} \tag{3}$$

  If $\Delta_{j+1}^p < \Delta_{\min}^p$ or if the number of new function evaluations exceeds $bbe_{\max}$, MADS(s_p) terminates and communicates $\Delta_{stop}^p = \Delta_j^p$ to the master. The next optimization performed by this slave will start with an initial mesh size parameter Δ_0^p equal to $4^\gamma \Delta_{stop}^p$, with $\gamma = 1$ if at least one success was achieved since the beginning of the current optimization (even by another slave), or else $\gamma = -1$. However, this may lead to a value smaller than $\Delta_{\min}^p = \Delta_k^M$, and, in this case, set $\Delta_0^p \leftarrow \Delta_k^M$.
  The Δ_0^p choice for the next MADS(s_p) is summarized by

$$\Delta_0^p \text{ (next MADS}(s_p)) \leftarrow \begin{cases} \min\left\{\Delta_0^{user}, 4\Delta_{stop}^p\right\} & \text{success,} \\ \max\left\{\Delta_k^M, \Delta_{stop}^p/4\right\} & \text{else.} \end{cases} \tag{4}$$

- **Master mesh size Δ_k^M at iteration k of PSD-MADS:** The update of the master mesh size is performed by the master after a pollster instance terminates. Δ_{k+1}^M is bounded below by the mesh size parameter of the terminated pollster Δ_k^1 and above by the minimum of all $\Delta_{\min}^p$ values currently used by regular slaves. These $\Delta_{\min}^p$ values correspond to previous master mesh sizes. It would be possible to choose the parameter α_k in Figure 4 at each update so that Δ_{k+1}^M is fixed to Δ_0^{user}, with α_k equal to the η_k from Proposition 4.1. However, such a strategy would not be efficient, as regular slaves would always generate trial points on the same mesh $M(\Delta_0^{user})$. The master mesh size has then to be reduced somehow through the PSD-MADS evolution. However, it should not be reduced too rapidly, or the algorithm would progress slowly or even terminate prematurely in practice.
  We propose the following strategy: From Figure 4, Δ_k^M is updated by $\Delta_{k+1}^M \leftarrow 4^{\alpha_k}\Delta_k^1$, with $\alpha_k \in \mathbb{N}$, and from Proposition 4.1, $\Delta_k^1 = 4^{-\eta_k}\Delta_0^{user}$, with some $\eta_k \in \mathbb{N}$. If iteration k succeeded, set $\alpha_k = \eta_k = \log_4\left(\Delta_0^{user}/\Delta_k^1\right)$ (maximal Δ_k^M increase), and else $\alpha_k = \eta_k - \lfloor(\eta_k + 1)/3\rfloor$ (attenuated Δ_k^M increase). In both cases, if Δ_{k+1}^M is greater than at least one of the regular slave's mesh

size $\Delta^p_{\min}$, then Δ^M_{k+1} is set to the least $\Delta^p_{\min}$ values. This can be summarized by the following:

$$\Delta^M_{k+1} \leftarrow \begin{cases} \min\left\{\Delta_0^{user}, \min\limits_{p\in Q_{reg}} \Delta^p_{\min}\right\} & \text{iteration success,} \\ \min\left\{4^{-\lfloor(\eta_k+1)/3\rfloor}\Delta_0^{user}, \min\limits_{p\in Q_{reg}} \Delta^p_{\min}\right\} & \text{iteration failure.} \end{cases} \tag{5}$$

For example, if $\Delta_0^{user} = \Delta^p_{\min} = 1$ for each $p \in Q_{reg}$ and if the pollster instance fails with a pollster mesh size of $\Delta^1_k = 1/16$, then the master mesh size Δ^M_{k+1} is set to $1/4$ ($\eta_k = 2$, $\alpha_k = 1$).

- **Pollster mesh size $\boldsymbol{\Delta^1_k}$ at iteration $\boldsymbol{k}$ of PSD-MADS:** In the case of an iteration success, Δ^1_{k+1} is set to Δ^M_{k+1} ($\omega_k = \alpha_k \in \mathbb{N}$), or else $\Delta^1_{k+1} = \Delta^1_k/4$ ($\omega_k = -1$):

$$\Delta^1_{k+1} \leftarrow \begin{cases} \Delta^M_{k+1} = \min\left\{\Delta_0^{user}, \min\limits_{p\in Q_{reg}} \Delta^p_{\min}\right\} & \text{iteration success,} \\ \Delta^1_k/4 & \text{iteration failure.} \end{cases} \tag{6}$$

MADS parameters for MADS(s_p), $p \in Q_{reg}$. The regular slaves $p \in Q_{reg}$ solve MADS(s_p) using the standard MADS $2|N_p|$ directions. All polls are opportunistic, meaning that a subproblem optimization terminates as soon as a better point is found. The one-point dynamic search strategy of [8] is also performed: it consists, after a successful poll step, in evaluating, within a single-point search, the black-box functions at a mesh point located further along the same successful direction.

In addition to the poll and the one-point dynamic search, MADS(s_p) performs a specialized search step, which simply consists in querying the cache server for the best available feasible point. This special search step generates no additional function evaluation and allows every regular slave to know the best points eventually obtained by other slaves. Note that this search step has no obligation to give a point lying on the current mesh of MADS(s_p), but this does not influence the convergence analysis as it is based on the pollster s_1, and as the point given by this search must come from another slave, thus lying on $M(\Delta^M_k)$.

Practical termination criteria. The regular slaves $p \in Q_{reg}$ terminate MADS(s_p) as soon as the mesh size parameter Δ^p_j drops below $\Delta^p_{\min} = \Delta^M_k$ (where k is the PSD-MADS iteration at which MADS(s_p) was started) or after a finite number of $bbe_{\max}$ black-box function evaluations are made. The PSD-MADS algorithm is stopped after an overall limit of $bbe^{global}_{\max}$ black-box evaluations is reached.

5.2. Numerical experiments. The PSD-MADS implementation described in section 5.1 is tested here, on two different problems. The implementation of MADS used to optimize subproblems corresponds to LTMADS and is the research version of the NOMAD C++ code [5]. The parallel master/slaves paradigm is achieved with MPI with $q = 6$ or 14 processes.

PSD-MADS is compared to three other parallel algorithms, on the same number q of processes: First, the pGPS method described in [17], which corresponds to the unmodified GPS method where evaluations are made in parallel. Second, pMADS, which is the trivial adaptation of pGPS that uses LTMADS instead of GPS. pGPS and pMADS are both synchronous parallel algorithms. The third method is APPS version 5.0.1 [25, 36], the only available GPS asynchronous parallel algorithm.

The first problem (referred as Problem A) considered for the tests is the G2 example from [28]. It has been chosen for its difficulty and for its variable size: our tests involve $n = 20, 50, 250$, and 500 variables. Problem A is written as follows:

$$\min_{x\in\mathbb{R}^n} f(x) = -\left|\frac{\sum_{i=1}^{n}\cos^4 x_i - 2\prod_{i=1}^{n}\cos^2 x_i}{\sqrt{\sum_{i=1}^{n} i x_i^2}}\right|$$

$$\text{subject to}\begin{cases} -\prod_{i=1}^{n} x_i + 0.75 \le 0, \\ \sum_{i=1}^{n} x_i - 7.5n \le 0, \\ 0 \le x_i \le 10, \quad i = 1, 2, \ldots, n. \end{cases}$$

The problem is treated as a black-box, and an upper limit of $100n$ function evaluations is imposed. The feasible starting point for all methods is the center of the bound constrained domain $x_0 = [5\ 5\ \ldots\ 5]^T \in \Omega$. The best known value from [28], for $n = 20$, is $f(x) = -0.803619$. In [28], various genetic algorithms gave good solutions, after several hundred thousand evaluations. Here, after a maximum of 2000 evaluations, PSD-MADS achieved $f(x) \simeq -0.76$.

The second test problem (Problem B) was designed for the MOVARS algorithm [12]. It has $n = 60$ variables and one constraint with two different versions: $G \ge 250$, or $G \ge 500$ (see [12] for a more complete description). An infeasible starting point is provided in [12], but cannot be used in the present work, since constraints are treated with the extreme barrier approach. The feasible starting points considered here for the two versions of Problem B have been obtained by minimizing the constraint violation $(\max\{0, 250 - G\})^2$ or $(\max\{0, 500 - G\})^2$, from the starting point of [12], with the pMADS algorithm. These optimizations required three evaluations for $G \ge 250$, with the resulting feasible point x_0 giving $f(x_0) = 3678.35$ and 74 evaluations for $G \ge 500$, and $f(x_0) = 3014$. These evaluations costs are considered in Figure 8. The feasible starting points, our source code for Problem B, and our best points are available on the website www.gerad.ca/Charles.Audet (see [5]). Our results for Problem B are not compared with the MOVARS algorithm results because numerical values are not given in [12]. The best solutions found gave $f(x) = 13.565$ for $G \ge 250$, and $f(x) = 245.866$ for $G \ge 500$.

The various results of this section are measured considering two quantities: z represents the best value of the objective function of problem $\mathcal{P}$, and *bbe* represents the total number of black-box evaluations. One evaluation is counted for the calls to both the objective f and the constraints of Ω.

The most representative cost of a black-box optimization algorithm is the number of black-box evaluations. For this reason, no speedup curves are given, and q is kept constant for each problem ($q = 14$ for Problem A and $q = 6$ for Problem B). Still, the durations of executions are given. The PSD-MADS method was not conceived in order to reduce the time to obtain a solution. Instead, we seek to obtain better solutions than a nondecomposing algorithm for problems with a large number of variables ($20 \le n \le 500$).

For all of our tests, the termination criteria is the maximum total number of black-box evaluations, which is $bbe_{\max}^{global} = 100n$ for Problem A and $bbe_{\max}^{global} = 3000$ for Problem B (as in [12]).

TABLE 1

Numerical results for problems A *and* B: z_{best}, z_{worst}, *and* z_{avg} *give information on the* 30 *runs performed for each* pMADS *and* PSD-MADS *test series,* S_{avg} *gives a measure of the area below the curves in Figures* 7 *and* 8, *and* t_{avg} *represents the average wall clock time, in seconds. Best values appear in bold.*

Algo.	Prob.	z_{best}	z_{worst}	z_{avg}	S_{avg}	t_{avg}	Prob.	z_{best}	z_{worst}	z_{avg}	S_{avg}	t_{avg}
pGPS		-0.450	-0.450	-0.450	1,002	7		-0.277	-0.277	-0.277	3,400	14
APPS	A	-0.519	**-0.519**	-0.519	782	**3**	A	-0.461	-0.461	-0.461	2,355	**6**
pMADS	n=20	**-0.775**	-0.434	-0.592	670	19	n=50	-0.498	-0.430	-0.457	1,939	33
PSD-MADS		-0.761	-0.430	**-0.666**	**595**	8		**-0.727**	**-0.528**	**-0.663**	**1,553**	29
pGPS		-0.089	-0.089	-0.089	18,336	**77**		-0.073	-0.073	-0.073	37,392	**179**
APPS	A	-0.196	-0.196	-0.196	16,934	137	A	-0.129	-0.129	-0.129	35,797	1,300
pMADS	n=250	-0.449	-0.438	-0.444	9,703	95	n=500	-0.447	-0.439	-0.443	19,380	275
PSD-MADS		**-0.698**	**-0.464**	**-0.603**	**8,568**	83		**-0.688**	**-0.461**	**-0.576**	**17,660**	277
pGPS		764.741	764.741	764.741	2,731,920	11		869.559	869.559	869.559	3,552,910	11
APPS	B	813.216	813.216	813.216	3,868,460	**6**	B	1,097.560	1,097.560	1,097.560	4,519,510	**6**
pMADS	$G \geq$	32.700	317.167	112.522	1,071,870	14	$G \geq$	417.049	948.768	662.841	2,892,140	14
PSD-MADS	250	**13.565**	**307.305**	**70.121**	**965,553**	14	500	**245.866**	**731.023**	**463.969**	**2,603,480**	19

The initial (and maximal) mesh size parameter is $\Delta_0^{user} = 2$ for Problem A. For Problem B, due to scaling reasons, the value of Δ_0^{user} differs for each variable and is set to be 0.2 times the range of the variables (i.e., $\Delta_0^{user} = 0.3$ for the 15 first variables, 0.35 for the next 30 variables, and 0.44 for the last 15 variables). These values have been decided empirically to give good results with standard MADS and APPS runs. The linear nature of the second constraint of Problem A is exploited by APPS. Since PSD-MADS and pMADS involve randomness in the polling directions, 30 runs are made for each test. The parallel execution of pGPS and APPS can affect their determinism; however, this effect was ignored, and one run was performed for each test.

To measure the quality of the solutions found, the best (z_{best}), worst (z_{worst}), and average (z_{avg}) values of the objective function value z at the $100n$th evaluation are reported. Another measure is S_{avg}, representing the area between a curve z versus *bbe* and the line $z = -0.8$ for Problem A (no run gave $z < -0.8$), and $z = 0$ for Problem B. Wall clock time expressed in seconds are reported in the column t_{avg}. Best runs are obtained with small values for all of these quantities.

PSD-MADS was tested on Problem A with $n = 20$ and 50 by varying $bbe_{\max}$, the maximum number of black-box evaluations for each regular subproblem, and ns, the number of variables in each subproblem. The number of processes has been set to $q = 14$ in order to fully exploit 12 processors. Good results were obtained by setting $bbe_{\max} = 10$ and having the regular slaves working on small dimensional subspaces $ns = 2$. These values are kept for $n > 50$. For Problem B, $bbe_{\max}$ is kept to 10. The best results have been obtained by distributing the 60 variables amongst 3 regular slaves with $q = 6$ and $ns = 20$.

Table 1 and Figures 7 and 8 summarize the numerical results. For all instances of Problem A, APPS outperforms pGPS, but neither does as well as PSD-MADS. In the three larger instances of Problem A, the worst f value produced by PSD-MADS is always better than all of the other methods'f values. For Problem B, pGPS outperforms APPS, and better results are obtained with pMADS and PSD-MADS, with a small advantage to PSD-MADS. In all of the curves in Figures 7 and 8, one can notice that pMADS is always the fastest to descend, but PSD-MADS overtakes it and produces better solutions. Finally, we remark that APPS terminates in the least wall clock time on smaller problems, albeit with a less optimal function value. However, for problems with 250 and 500 variables, the wall clock time grows significantly worse. This is in accordance with the remark in [25] stating that APPS targets problems with less than 100 variables.

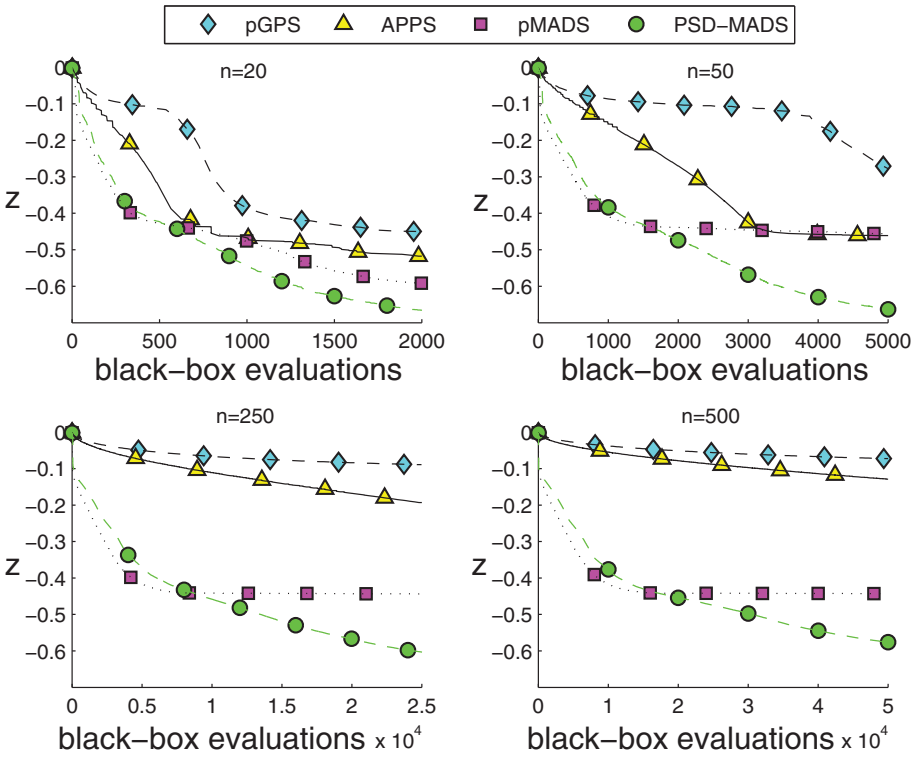


FIG. 7. *Problem* A*: graphs of the objective function value versus the number of evaluations for all test results.* PSD-MADS *and* pMADS *plots correspond to average values of the* 30 *runs performed for each test.*

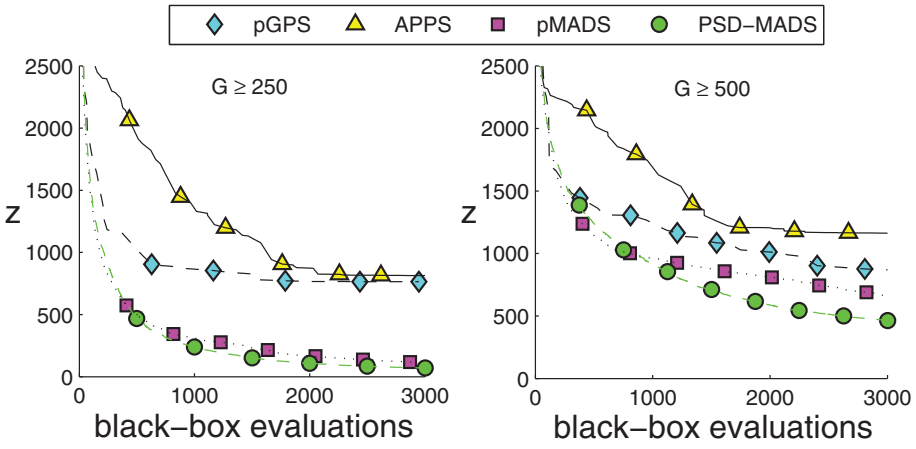


FIG. 8. *Problem* B*: graphs of the objective function value versus the number of evaluations for all test results.* PSD-MADS *and* pMADS *plots correspond to average values of the* 30 *runs performed for each test.*

We conclude this section with some advice for readers interested in testing Psd-Mads. First, we think that the Psd-Mads decomposition is beneficial for problems with more than 20 variables. For these problems, at least 3 processors are necessary. Furthermore, since the master and cache server processes are not demanding in terms of Cpu, 5 processes can be executed on the 3 processors, whose work will be mainly devoted to two regular slaves and the pollster. Two regular slaves is the minimum number to benefit from the decomposition. So, even if only a few processors are available, it is still worthwhile to try this method. Finally, if the user has no particular strategy to choose the subsets of variables in each subproblem, we recommend to equally distribute the variables to the regular slaves. If the user knows that some of the variables are more likely to produce descent than others, then some subproblems can be devoted to these variables, while single-poll Mads can be used on the subproblems of less important variables.

6. Discussion and possible extensions. This paper introduced Psd-Mads, a new Psd technique with less restrictive conditions than usual Psd methods on the functions to be optimized and on the sets of variables in the subproblems. It is shown that the algorithm, from any starting point, produces a subsequence of iterates converging to a solution satisfying local optimality conditions (global convergence to local solutions), based on Clarke calculus and on the Mads convergence analysis. A practical implementation is described, with a small number of parameters ($bbe_{\max}$ and ns), and very encouraging results have been obtained on a difficult problem from the literature, with up to 500 variables.

We presented a first basic implementation of Psd-Mads with a very simple and generic strategy to choose the sets of variables. An obvious extension is a better strategy to decide on the sets of variables in the subproblems. Of course, it is not clear how to do this, in general, or we would have done it here. However, for some applications, the user may have special knowledge that would help in this task. For example, the user might put similarly scaled variables in the same subproblem.

It would also be interesting to incorporate the Pvd idea of the "forget-me-not" terms and allow some basic changes in the subproblems for fixed variables. A third possibility would be to perform some additional search steps in the slave subspaces. Another possible extension would be to reintroduce the synchronization step of the original block-Jacobi method but without the parallel barrier. This "recomposition" step could be performed in parallel by one of the regular slaves, from a pool of successful points, in order to create a problem similar to the one in [19]. Finally, the constraints of Ω could be treated with the progressive barrier [9], instead of the extreme barrier approach. This would allow for infeasible iterates, including the starting point.

Acknowledgments. We would like to thank anonymous referees for their constructive remarks and comments.

REFERENCES

[1] M. A. Abramson and C. Audet, *Convergence of mesh adaptive direct search to second-order stationary points*, SIAM J. Optim., 17 (2006), pp. 606–619.

[2] M. A. Abramson, *Mixed variable optimization of a load-bearing thermal insulation system using a filter pattern search algorithm*, Optim. Eng., 5 (2004), pp. 157–177.

[3] P. Alberto, F. Nogueira, H. Rocha, and L. N. Vicente, *Pattern search methods for user-provided points: Application to molecular geometry problems*, SIAM J. Optim., 14 (2004), pp. 1216–1236.

[4] C. Audet, V. Béchard, and S. Le Digabel, *Nonsmooth optimization through mesh adaptive direct search and variable neighborhood search*, J. Global Optim., 41 (2008), pp. 299–318.
[5] C. Audet, G. Couture, and J. E. Dennis, Jr., *The* Nomad *Project*, www.gerad.ca/nomad.
[6] C. Audet, A. L. Custódio, and J. E. Dennis, Jr., *Erratum: Mesh adaptive direct search algorithms for constrained optimization*, SIAM J. Optim., 18 (2008), pp. 1501–1503.
[7] C. Audet and J. E. Dennis, Jr., *Analysis of generalized pattern searches*, SIAM J. Optim., 13 (2002), pp. 889–903.
[8] C. Audet and J. E. Dennis, Jr., *Mesh adaptive direct search algorithms for constrained optimization*, SIAM J. Optim., 17 (2006), pp. 188–217.
[9] C. Audet and J. E. Dennis, Jr., *A* Mads *Algorithm with a Progressive Barrier for Derivative-Free Nonlinear Programming*, SIAM J. Optim., submitted.
[10] C. Audet and D. Orban, *Finding optimal algorithmic parameters using derivative-free optimization*, SIAM J. Optim., 17 (2006), pp. 642–664.
[11] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall, Inc., Upper Saddle River, NJ, 1989.
[12] A. J. Booker, E. J. Cramer, P. D. Frank, J. M. Gablonsky, and J. E. Dennis, Jr., MoVars*: Multidisciplinary optimization via adaptive response surfaces*, Aiaa paper 2007–1927, in Proceedings of the 48th AIAA/ASME/ASCE/AHS/ASC Conference on Structures, Structural Dynamics, and Materials, Honolulu, 2007.
[13] A. J. Booker, J. E. Dennis, Jr., P. D. Frank, D. W. Moore, and D. B. Serafini, *Managing surrogate objectives to optimize a helicopter rotor design—further experiments*, Aiaa paper 1998–4717, in Proceedings of the 7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization, St. Louis, MO, 1998.
[14] A. J. Booker, J. E. Dennis, Jr., P. D. Frank, D. B. Serafini, V. Torczon, and M. W. Trosset, *A rigorous framework for optimization of expensive functions by surrogates*, Struct. Optim., 17 (1999), pp. 1–13.
[15] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983. Reissued in 1990 by SIAM, Philadelphia, as Vol. 5 in the series Classics Appl. Math.
[16] I. D. Coope and C. J. Price, *Frame-based methods for unconstrained optimization*, J. Optim. Theory Appl., 107 (2000), pp. 261–274.
[17] J. E. Dennis, Jr. and V. Torczon, *Direct search methods on parallel machines*, SIAM J. Optim., 1 (1991), pp. 448–474.
[18] J. E. Dennis, Jr. and Z. Wu, *Parallel continuous optimization*, Sourcebook of Parallel Computing, Morgan Kaufmann Publishers, San Francisco, CA, 2003, pp. 649–670.
[19] M. C. Ferris and O. L. Mangasarian, *Parallel variable distribution*, SIAM J. Optim., 4 (1994), pp. 815–832.
[20] D. Finkel and C. Kelley, *Convergence Analysis of the* Direct *Algorithm*, Technical report CRSC-TR04-28, Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC, 2004.
[21] K. R. Fowler, C. T. Kelley, C. T. Miller, C. E. Kees, R. W. Darwin, J. P. Reese, M. W. Farthing, and M. S. C. Reed, *Solution of a well-field design problem with implicit filtering*, Optim. Eng., 5 (2004), pp. 207–234.
[22] A. Frommer and R. A. Renaut, *A unified approach to parallel space decomposition methods*, J. Comput. Appl. Math., 110 (1999), pp. 205–233.
[23] M. Fukushima, *Parallel variable transformation in unconstrained optimization*, SIAM J. Optim., 8 (1998), pp. 658–672.
[24] J. Gablonsky and C. T. Kelley, *A locally-biased form of the direct algorithm*, J. Global Optim., 21 (2001), pp. 27–37.
[25] G. A. Gray and T. G. Kolda, *Algorithm* 856: *Appspack* 4.0: *A synchronous parallel pattern search for derivative-free optimization*, ACM Trans. Math. Software, 32 (2006), pp. 485–507.
[26] S.-P. Han, *Optimization by updated conjugate subspaces*, in Numerical Analysis, Pitman Res. Notes Math. Ser. 140, D. Griffiths and G. Watson, eds., Longman Sci. Tech., Harlow, 1986, pp. 82–97.
[27] R. E. Hayes, F. H. Bertrand, C. Audet, and S. T. Kolaczkowski, *Catalytic combustion kinetics: Using a direct search algorithm to evaluate kinetic parameters from light-off curves*, Canad. J. Chem. Eng., 81 (2003), pp. 1192–1199.
[28] A. Hedar and M. Fukushima, *Derivative-free filter simulated annealing method for constrained continuous global optimization*, J. Global Optim., 35 (2006), pp. 521–549.
[29] P. D. Hough, T. G. Kolda, and V. J. Torczon, *Asynchronous parallel pattern search for nonlinear optimization*, SIAM J. Sci. Comput., 23 (2001), pp. 134–156.

[30] D. R. Jones, C. D. Perttunen, and B. E. Stuckman, *Lipschitzian optimization without the Lipschitz constant*, J. Optim. Theory Appl., 79 (1993), pp. 157–181.
[31] C. T. Kelley, *Iterative Methods for Optimization*, Frontiers Appl. Math. 18, SIAM, Philadelphia, 1999.
[32] M. Kokkolaras, C. Audet, and J. E. Dennis, Jr., *Mixed variable optimization of the number and composition of heat intercepts in a thermal insulation system*, Optim. Eng., 2 (2001), pp. 5–29.
[33] T. G. Kolda, R. M. Lewis, and V. Torczon, *Optimization by direct search: New perspectives on some classical and modern methods*, SIAM Rev., 45 (2003), pp. 385–482.
[34] T. G. Kolda and V. Torczon, *Understanding asynchronous parallel pattern search*, in High Performance Algorithms and Software for Nonlinear Optimization, G. DiPillo and A. Murli, eds., Kluwer Academic Publishers, Norwell, MA, 2003, pp. 316–335.
[35] T. G. Kolda and V. J. Torczon, *On the convergence of asynchronous parallel pattern search*, SIAM J. Optim., 14 (2004), pp. 939–964.
[36] T. G. Kolda, *Revisiting asynchronous parallel pattern search for nonlinear optimization*, SIAM J. Optim., 16 (2005), pp. 563–586.
[37] R. M. Lewis, V. Torczon, and M. W. Trosset, *Direct search methods: Then and now*, J. Comput. Appl. Math., 124 (2000), pp. 191–207.
[38] R. M. Lewis and V. Torczon, *Pattern search algorithms for bound constrained minimization*, SIAM J. Optim., 9 (1999), pp. 1082–1099.
[39] R. M. Lewis and V. Torczon, *Pattern search methods for linearly constrained minimization*, SIAM J. Optim., 10 (2000), pp. 917–941.
[40] C.-S. Liu and C.-H. Tseng, *Parallel synchronous and asynchronous space-decomposition algorithms for large-scale minimization problems*, Comput. Optim. Appl., 17 (2000), pp. 85–107.
[41] L. O. Mangasarian, *Parallel gradient istribution in unconstrained optimization*, SIAM J. Control Optim., 33 (1995), pp. 1916–1925.
[42] A. L. Marsden, M. Wang, J. E. Dennis, Jr., and P. Moin, *Optimal aeroacoustic shape design using the surrogate management framework*, Optim. Eng., 5 (2004), pp. 235–262.
[43] J. A. Nelder and R. Mead, *A simplex method for function minimization*, Comput. J., 7 (1965), pp. 308–313.
[44] C. J. Price and I. D. Coope, *Frames and grids in unconstrained and linearly constrained optimization: A nonsmooth approach*, SIAM J. Optim., 14 (2003), pp. 415–438.
[45] C. A. Sagastizábal and M. V. Solodov, *Parallel variable distribution for constrained optimization*, Comput. Optim. Appl., 22 (2002), pp. 111–131.
[46] M. V. Solodov, *New inexact parallel variable distribution algorithms*, Comput. Optim. Appl., 7 (1997), pp. 165–182.
[47] M. V. Solodov, *On the convergence of constrained parallel variable distribution algorithms*, SIAM J. Optim., 8 (1998), pp. 187–196.
[48] B. Tang, *Orthogonal array-based latin hypercubes*, J. Amer. Statist. Assoc., 88 (1993), pp. 1392–1397.
[49] V. Torczon, *On the convergence of pattern search algorithms*, SIAM J. Optim., 7 (1997), pp. 1–25.
[50] P. Tseng, *Dual coordinate ascent methods for non-strictly convex minimization*, Math. Program., 59 (1993), pp. 231–247.
[51] E. Yamakawa and M. Fukushima, *Testing parallel variable transformation*, Comput. Optim. Appl., 13 (1999), pp. 253–274.

SIAM J. OPTIM.
Vol. 19, No. 3, pp. 1171–1183

SMOOTH OPTIMIZATION WITH APPROXIMATE GRADIENT*

ALEXANDRE D'ASPREMONT†

Abstract. We show that the optimal complexity of Nesterov's smooth first-order optimization algorithm is preserved when the gradient is computed only up to a small, uniformly bounded error. In applications of this method to semidefinite programs, this means in some instances computing only a few leading eigenvalues of the current iterate instead of a full matrix exponential, which significantly reduces the method's computational cost. This also allows sparse problems to be solved efficiently using sparse maximum eigenvalue packages.

Key words. smooth optimization, first-order methods, semidefinite programming

AMS subject classifications. 90C60, 65K05, 90C22

DOI. 10.1137/060676386

1. Introduction.

In [13] it was shown that smooth convex minimization problems of the form:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in Q, \end{array}$$

where f is a convex function with Lipschitz continuous gradient and Q is a sufficiently simple compact convex set, could be solved with a complexity of $O(1/\sqrt{\epsilon})$, where ϵ is the precision target. Furthermore, it can be shown that this complexity bound is optimal for that class of problems (see [14] for a discussion). More recently, [15] showed that this method could be combined with a smoothing argument to produce an $O(1/\epsilon)$ complexity bound for nonsmooth problems where the objective has a saddle-function format. In particular, this meant that a broad class of semidefinite optimization problems could be solved with significantly lower memory requirements than interior point methods and a better complexity bound than classic first-order methods (bundle, subgradient, etc.).

Here, we show that substituting an approximate gradient, which may allow significant computation and storage savings, does not affect the optimal complexity of the algorithm in [13]. It is somewhat intuitive that an algorithm which exhibits good numerical performance in practice should be robust to at least some numerical error in the objective function and gradient computations since all implementations are necessarily computing these quantities up to some multiple of machine precision. Our objective here is to make that robustness explicit in order to design optimal schemes using only approximate gradient information.

For nonsmooth problems, when the objective function $f(x)$ can be expressed as a saddle function on a compact set, the method in [15] starts by computing a smooth (i.e., with Lipschitz continuous gradient), uniform ϵ-approximation of the objective function $f(x)$; it then uses the smooth minimization algorithm in [13] to solve the approximate problem. When this smoothing technique is applied to semidefinite

*Received by the editors November 30, 2006; accepted for publication (in revised form) May 9, 2008; published electronically October 31, 2008.
http://www.siam.org/journals/siopt/19-3/67638.html
†ORFE Department, Princeton University, Princeton, NJ 08544 (aspremon@princeton.edu). This author received support from NSF grant DMS-0625352, ONR grant N00014-07-1-0150, a Peek junior faculty fellowship, and a gift from Google, Inc.

optimization, computing exact gradients requires forming a matrix exponential, which is often the dominant numerical step in the algorithm.

Although there are many different methods for computing this matrix exponential (see [11] for a survey), their complexity is comparable to that of a full eigenvalue decomposition of the matrix. In problem instances where only a few leading eigenvalues suffice to approximate this exponential, the per iteration complexity of the algorithm described here becomes comparable to that of classical first-order methods such as the bundle method (see [6]) or subgradient methods (see [19], for example), which have a global complexity bound of $O(1/\epsilon^2)$ (see [14]), while keeping the optimal complexity of $O(1/\epsilon)$ of the algorithm in [15].

We apply this result to a maximum eigenvalue minimization problem (or semidefinite program with constant trace). We first recall the complexity bound derived in [16] based on a smoothing argument, using exact gradients. We produce a rough theoretical estimate of the number of eigenvalues required for convergence when approximate gradients are used. We then derive an explicit condition on the quality of the gradient approximation to guarantee convergence and compute a bound on the number of iterations. We show both on randomly generated problem instances and on problems generated from biological data sets that actual computational savings vary significantly with problem structure but can be substantial in some cases.

The paper is organized as follows. In the next section, we prove convergence of the algorithm in [13] when only an approximate gradient is used. In section 3 we describe how these results can be applied to semidefinite optimization. Finally, in the last section we test their performance on semidefinite relaxation and maximum eigenvalue minimization problems.

2. Smooth optimization with approximate gradient. Following the results and notations in [15, section 3], we study the problem:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in Q, \end{array} \tag{2.1}$$

where $Q \subset \mathbf{R}^n$ is a compact convex set and f is a convex function with Lipschitz continuous gradient, such that

$$\|\nabla f(x) - \nabla f(y)\|^* \leq L\|x-y\|, \quad x, y \in Q,$$

for some $L > 0$, which also means

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2}L\|y-x\|^2, \quad x, y \in Q. \tag{2.2}$$

The key difference here is that the *oracle* information we obtain for ∇f is *noisy*. Note that the function values are not required to compute iterates in the algorithm described here, so even if our knowledge of function values $f(x)$ is noisy, we will always use exact values in the proofs that follow. At each iteration, we obtain $\tilde{\nabla} f(x)$ satisfying

$$\left|\left\langle \tilde{\nabla} f(x) - \nabla f(x), y - z\right\rangle\right| \leq \delta \quad x, y, z \in Q, \tag{2.3}$$

for some precision level $\delta > 0$. Throughout the paper, we assume that Q is simple enough so that this condition can be checked efficiently. As in [13], we also assume that certain projection operators on Q can be computed efficiently, and we refer the

reader to the end of this section for more details. Here, $d(x)$ is a prox-function for the set Q, i.e., continuous and strongly convex on Q with parameter σ (see [14] or [7] for a discussion of regularization techniques using strongly convex functions). We let x_0 be the center of Q for the prox-function $d(x)$ so that

$$x_0 \triangleq \operatorname{argmin}_{x\in Q} d(x),$$

assuming without loss of generality that $d(x_0) = 0$, we then have

$$d(x) \geq \frac{1}{2}\sigma\|x - x_0\|^2. \tag{2.4}$$

We denote by $\tilde{T}_Q(x)$ a solution to the following subproblem:

$$\tilde{T}_Q(x) \triangleq \operatorname{argmin}_{y\in Q} \left\{ \langle \tilde{\nabla} f(x), y - x\rangle + \frac{1}{2}L\|y - x\|^2 \right\}. \tag{2.5}$$

We let $y_0 = \tilde{T}_Q(x_0)$, where x_0 is defined above. We recursively define three sequences of points: the current iterate $\{x_k\}$, the corresponding $y_k = \tilde{T}_Q(x_k)$, together with

$$z_k \triangleq \operatorname{argmin}_{x\in Q} \left\{ \frac{L}{\sigma} d(x) + \sum_{i=0}^{k} \alpha_i \big[f(x_i) + \langle \tilde{\nabla} f(x_i), x - x_i\rangle \big] \right\}, \tag{2.6}$$

and a step size sequence $\{\alpha_k\} \geq 0$ with $\alpha_0 \in (0, 1]$ so that

$$\begin{aligned} x_{k+1} &= \tau_k z_k + (1 - \tau_k) y_k, \\ y_{k+1} &= \tilde{T}_Q(x_{k+1}), \end{aligned} \tag{2.7}$$

where $\tau_k = \alpha_{k+1}/A_{k+1}$ with $A_k = \sum_{i=0}^{k} \alpha_i$. We implicitly assume here that the two subproblems defining y_k and z_k can be solved very efficiently (in the examples that follow, they amount to Euclidean projections). We will show recursively that for a good choice of step sequence α_k, the iterates x_k and y_k satisfy the following relationship (denoted by $\mathcal{R}_k$):

$$A_k f(y_k) \leq \psi_k + A_k g(k, \delta) \quad (\mathcal{R}_k),$$

where $g(k, \delta)$ measures the accumulated gradient approximation error and will be bounded in Lemma 2.1, and

$$\psi_k \triangleq \min_{x\in Q} \left\{ \frac{L}{\sigma} d(x) + \sum_{i=0}^{k} \alpha_i \big[f(x_i) + \langle \tilde{\nabla} f(x_i), x - x_i\rangle \big] \right\}.$$

First, using $d(x) \geq \frac{1}{2}\sigma\|x - x_0\|^2$, then inequality (2.2) and condition (2.3), we have

$$\psi_0 = \min_{x\in Q} \left\{ \frac{L}{\sigma} d(x) + \alpha_0 \big[f(x_0) + \langle \tilde{\nabla} f(x_0), x - x_0\rangle \big] \right\} \geq \alpha_0 f(y_0) - \alpha_0 \delta,$$

which is $\mathcal{R}_0$. We can then bound the approximation error in the following result.

LEMMA 2.1. *Let α_k be a step sequence satisfying:*

$$0 < \alpha_0 \leq 1 \quad \textit{and} \quad \alpha_k^2 \leq A_k, \quad k \geq 0; \tag{2.8}$$

suppose that $(\mathcal{R}_k)$ *holds with* x_{k+1} *and* y_{k+1} *are defined as in* (2.7)*, and then* $(\mathcal{R}_{k+1})$ *holds with*

$$g(k+1,\delta) = (1-\tau_k)g(k,\delta) + \tau_k 3\delta,$$

where $\tau_k \in [0,1]$ *and* $g(0,\delta) = \alpha_0\delta$.

Proof. Let us assume that $(\mathcal{R}_k)$ holds. Because $d(x)$ is strongly convex with parameter σ, the function

$$\frac{L}{\sigma}d(x) + \sum_{i=0}^{k} \alpha_i \big[f(x_i) + \langle \tilde{\nabla} f(x_i), x - x_i\rangle\big]$$

is strongly convex with parameter L. Using this property and the definition of z_k, we obtain

$$\begin{aligned}\psi_{k+1} &= \min_{x\in Q}\left\{\frac{L}{\sigma}d(x) + \sum_{i=0}^{k+1} \alpha_i \big[f(x_i) + \langle \tilde{\nabla} f(x_i), x - x_i\rangle\big]\right\}\\ &\geq \min_{x\in Q}\left\{\psi_k + \frac{1}{2}L\|x-z_k\|^2 + \alpha_{k+1}\big[f(x_{k+1}) + \langle \tilde{\nabla} f(x_{k+1}), x - x_{k+1}\rangle\big]\right\}.\end{aligned}$$

Now, using $(\mathcal{R}_k)$ and then the convexity of $f(x)$, we get

$$\begin{aligned}&\psi_k + A_k g(k,\delta) + \alpha_{k+1}\big[f(x_{k+1}) + \langle \tilde{\nabla} f(x_{k+1}), x - x_{k+1}\rangle\big]\\ &\geq A_k f(y_k) + \alpha_{k+1}\big[f(x_{k+1}) + \langle \tilde{\nabla} f(x_{k+1}), x - x_{k+1}\rangle\big]\\ &\geq A_k[f(x_k) + \langle \nabla f(x_{k+1}), y_k - x_{k+1}\rangle] + \alpha_{k+1}\big[f(x_{k+1}) + \langle \tilde{\nabla} f(x_{k+1}), x - x_{k+1}\rangle\big],\end{aligned}$$

and condition (2.3), together with (2.7), implies that:

$$\begin{aligned}&A_k[f(x_k) + \langle \nabla f(x_{k+1}), y_k - x_{k+1}\rangle] + \alpha_{k+1}\big[f(x_{k+1}) + \langle \tilde{\nabla} f(x_{k+1}), x - x_{k+1}\rangle\big]\\ &\geq A_{k+1}f(x_{k+1}) + \langle \nabla f(x_{k+1}), A_k y_k - A_k x_{k+1} + \alpha_{k+1}(x - x_{k+1})\rangle - \alpha_{k+1}\delta\\ &= A_{k+1}f(x_{k+1}) + \alpha_{k+1}\langle \nabla f(x_{k+1}), x - z_k\rangle - \alpha_{k+1}\delta.\end{aligned}$$

Because α_k satisfies (2.8), we have $\tau_k^2 \leq A_{k+1}^{-1}$ and can combine the last three inequalities to get

$$\begin{aligned}\psi_{k+1} &\geq A_{k+1}f(x_{k+1}) - A_k g(k,\delta) - \alpha_{k+1}\delta\\ &\quad + \min_{x\in Q}\left\{\frac{1}{2}L\|x-z_k\|^2 + \alpha_{k+1}\langle \nabla f(x_{k+1}), x - z_k\rangle\right\}\\ &\geq A_{k+1}\Bigg[f(x_{k+1}) - (1-\tau_k)g(k,\delta) - \tau_k\delta\\ &\quad + \min_{x\in Q}\left\{\frac{1}{2}L\tau_k^2\|x-z_k\|^2 + \tau_k\langle \nabla f(x_{k+1}), x - z_k\rangle\right\}\Bigg].\end{aligned} \tag{2.9}$$

Let us define $y \triangleq \tau_k x + (1-\tau_k)y_k$ so that $y - x_{k+1} = \tau_k(x - z_k)$, with

$$\begin{aligned}&\min_{x\in Q}\left\{\frac{1}{2}L\tau_k^2\|x-z_k\|^2 + \tau_k\langle \nabla f(x_{k+1}), x - z_k\rangle\right\}\\ &= \min_{\{y\in\tau_k Q + (1-\tau_k)y_k\}}\left\{\frac{1}{2}L\|y - x_{k+1}\|^2 + \langle \nabla f(x_{k+1}), y - x_{k+1}\rangle\right\}.\end{aligned}$$

Combining condition (2.3) with the fact that $y - x_{k+1} = \tau_k(x - z_k)$ for some $x, z_k \in Q$, we get

$$\begin{aligned}&\min_{\{y\in\tau_kQ+(1-\tau_k)y_k\}}\left\{\frac{1}{2}L\|y-x_{k+1}\|^2+\langle\nabla f(x_{k+1}),y-x_{k+1}\rangle\right\}\\&\geq\min_{\{y\in\tau_kQ+(1-\tau_k)y_k\}}\left\{\frac{1}{2}L\|y-x_{k+1}\|^2+\left\langle\tilde{\nabla}f(x_{k+1}),y-x_{k+1}\right\rangle\right\}-\tau_k\delta.\end{aligned}$$

Now, because Q is convex, we must have $\tau_kQ + (1-\tau_k)y_k \subset Q$ and

$$\begin{aligned}&\min_{\{y\in\tau_kQ+(1-\tau_k)y_k\}}\left\{\frac{1}{2}L\|y-x_{k+1}\|^2+\left\langle\tilde{\nabla}f(x_{k+1}),y-x_{k+1}\right\rangle\right\}-\tau_k\delta\\&\geq\min_{y\in Q}\left\{\frac{1}{2}L\|y-x_{k+1}\|^2+\left\langle\tilde{\nabla}f(x_{k+1}),y-x_{k+1}\right\rangle\right\}-\tau_k\delta.\end{aligned}$$

By the definition of $y_{k+1} = \tilde{T}_Q(x_{k+1})$ and using condition (2.3), we get

$$\begin{aligned}&\min_{y\in Q}\left\{\frac{1}{2}L\|y-x_{k+1}\|^2+\left\langle\tilde{\nabla}f(x_{k+1}),y-x_{k+1}\right\rangle\right\}-\tau_k\delta\\&=\frac{1}{2}L\|\tilde{T}_Q(x_{k+1})-x_{k+1}\|^2+\left\langle\tilde{\nabla}f(x_{k+1}),\tilde{T}_Q(x_{k+1})-x_{k+1}\right\rangle-\tau_k\delta\\&\geq\frac{1}{2}L\|y_{k+1}-x_{k+1}\|^2+\left\langle\nabla f(x_{k+1}),y_{k+1}-x_{k+1}\right\rangle-2\tau_k\delta,\end{aligned}$$

and inequality (2.2) gives

$$\begin{aligned}&\frac{1}{2}L\|y_{k+1}-x_{k+1}\|^2+\langle\nabla f(x_{k+1}),y_{k+1}-x_{k+1}\rangle-2\tau_k\delta\\&\geq f(y_{k+1})-f(x_{k+1})-2\tau_k\delta.\end{aligned}$$

Combining these inequalities with the inequality on ψ_{k+1} in (2.9), we finally get

$$\psi_{k+1}\geq A_{k+1}\left[f(y_{k+1})-(1-\tau_k)g(k,\delta)-3\tau_k\delta\right],$$

which is the desired result. □

We can use this result to study the convergence of the following algorithm given only approximate gradient information.

Smooth minimization with approximate gradient.

Starting from x_0, the prox center of the set Q, we iterate:

1. compute $\tilde{\nabla}f(x_k)$,
2. compute $y_k = \tilde{T}_Q(x_k)$,
3. compute $z_k = \operatorname{argmin}_{x\in Q}\{\frac{L}{\sigma}d(x) + \sum_{i=0}^{k}\alpha_i[f(x_i) + \langle\tilde{\nabla}f(x_i), x - x_i\rangle]\}$,
4. update x using $x_{k+1} = \tau_k z_k + (1-\tau_k)y_k$.

Again, because solving for y_k and z_k can often be done very efficiently, the dominant numerical step in this algorithm is the evaluation of $\tilde{\nabla}f(x_k)$. If the step size sequence α_k satisfies the conditions of Lemma 2.1, we can show the following convergence result.

THEOREM 2.2. *Suppose α_k satisfies* (2.8)*, with the iterates x_k and y_k defined in* (2.6) *and* (2.7)*, and then for any $k \geq 0$ we have:*

$$f(y_k) - f(x^\star) \leq \frac{Ld(x^\star)}{A_k \sigma} + 3\delta,$$

where $x^\star$ is an optimal solution to problem (2.1)*.*

Proof. If α_k satisfies the hypotheses of Lemma 2.1, we have

$$A_k f(y_k) \leq \psi_k + A_k g(k, \delta),$$

where $A_k = \sum_{i=0}^{k} \alpha_i$ and $g(k, \delta) \leq 3\delta$. Now, because $f(x)$ is convex, we also have

$$\psi_k \leq \frac{L}{\sigma} d(x^\star) + A_k f(x^\star) + A_k 3\delta,$$

which yields the desired result. □

When $d(x^\star) < +\infty$ (e.g., if Q is bounded), if we set the step sequence as $\alpha_k = (k+1)/2$ and δ to some fraction of the target precision ϵ (here $\epsilon/6$), A_k grows as $O(k^2)$, and Theorem 2.2 ensures that the algorithm will converge to an ϵ solution in less than

$$\sqrt{\frac{8Ld(x^\star)}{\sigma\epsilon}} \tag{2.10}$$

iterations. In practice, of course, $d(x^\star)$ needs to be bounded a priori, and L is often hard to evaluate. A notable exception is when $f(x)$ is a smooth approximation (as in [15, 16], for example), in which case L is known explicitly as a function of the precision. We have implicitly assumed, as in [13], that the set Q is simple enough so that the complexity of solving the two minimization subproblems in steps 2 and 3 of the algorithm is low relative to that of approximating the gradient. We also implicitly assumed that the set Q is simple enough so that condition (2.3) can be checked efficiently. In the numerical experiments of section 4, for example, steps 2 and 3 are Euclidean projections on the unit box, and condition (2.3) is a simple inequality on the leading eigenvalues of the current iterate.

3. Semidefinite optimization. Here, we describe in detail how the results of the previous section can be applied to semidefinite optimization. We consider the following maximum eigenvalue problem:

$$\begin{array}{ll} \text{minimize} & \lambda^{\max}\left(A^T y + c\right) - b^T y \\ \text{subject to} & y \in Q, \end{array} \tag{3.1}$$

in the variable $y \in \mathbf{R}^m$, with parameters $A \in \mathbf{R}^{m \times n^2}$, $b \in \mathbf{R}^m$, and $c \in \mathbf{R}^{n^2}$. Let us remark that when Q is equal to $\mathbf{R}^m$, the dual of this program is a semidefinite program with constant trace written:

$$\begin{array}{ll} \text{maximize} & c^T x \\ \text{subject to} & Ax = b, \\ & \mathbf{Tr}(x) = 1, \\ & x \succeq 0, \end{array} \tag{3.2}$$

in the variable $x \in \mathbf{R}^{n^2}$, where $\mathbf{Tr}(x) = 1$ means that the matrix obtained by reshaping the vector x has trace equal to one and $x \succeq 0$ means that this same matrix is symmetric, positive semidefinite.

3.1. Smoothing technique. As in [12], [16], [3], or [2], we can find a uniform ϵ-approximation to $\lambda^{\max}(X)$ with Lipschitz continuous gradient. Let $\mu > 0$ and $X \in \mathbf{S}_n$, and we define

$$f_\mu(X) = \mu \log \left(\sum_{i=1}^n e^{\lambda_i(X)/\mu} \right) = \mu \log \left(e^{\frac{\lambda^{\max}(X)}{\mu}} \left(1 + \sum_{i=2}^n e^{\frac{\lambda_i(X) - \lambda^{\max}(X)}{\mu}} \right) \right),$$

where $\lambda_i(X)$ is the i^{th} eigenvalue of X. This is also:

$$f_\mu(X) = \lambda^{\max}(X) + \mu \log \mathbf{Tr} \left(\exp \left(\frac{X - \lambda^{\max}(X)\mathbf{I}}{\mu} \right) \right), \tag{3.3}$$

which requires computing a matrix exponential at a numerical cost of $O(n^3)$. We then have

$$\lambda^{\max}(X) \leq f_\mu(X) \leq \lambda^{\max}(X) + \mu \log n,$$

so if we set $\mu = \epsilon / \log n$, $f_\mu(X)$ becomes a uniform ϵ-approximation of $\lambda^{\max}(X)$. In [16] it was shown that $f_\mu(X)$ has a Lipschitz continuous gradient with constant

$$L = \frac{1}{\mu} = \frac{\log n}{\epsilon}.$$

The gradient $\nabla f_\mu(X)$ can also be computed explicitly as

$$\frac{\exp \left(\frac{X - \lambda^{\max}(X)\mathbf{I}}{\mu} \right)}{\mathbf{Tr} \left(\exp \left(\frac{X - \lambda^{\max}(X)\mathbf{I}}{\mu} \right) \right)}, \tag{3.4}$$

using the same matrix exponential as in (3.3). Let $\|y\|$ be some norm on $\mathbf{R}^m$ and $d(x)$ a strongly convex prox-function with parameter $\sigma > 0$. As in [16], we define

$$\|A\| = \max_{\|h\|=1} \|A^T h\|_2,$$

where $\|A^T h\|_2 = \max_i |\lambda_i(A^T h)|$. The algorithm detailed in [15], where *exact* function values and gradients are computed, will find an ϵ solution to (3.1) after at most

$$\frac{2\|A\|}{\epsilon} \sqrt{\frac{\log n}{\sigma} d(y^\star)} \tag{3.5}$$

iterations, each iteration requiring a matrix exponential computation.

3.2. Spectrum and expected performance gains. The complexity estimate above is valid when the matrix exponential in (3.3) is computed exactly, at a cost of $O(n^3)$. As we will see below, only a few leading eigenvalues are sometimes required to satisfy conditions (2.3) and obtain a comparable complexity estimate at a much lower numerical cost. To illustrate the potential complexity gains, let us pick a matrix $X \in \mathbf{S}_n$ whose coefficients are centered independent normal variables with the second moment given by σ^2/n. From Wigner's semicircle law, $\lambda^{\max}(X) \sim 2\sigma$ as n goes to infinity, and the eigenvalues of X are asymptotically distributed according to the density

$$p(x) = \frac{1}{2\pi\sigma^2} \sqrt{4\sigma^2 - x^2},$$

which means that, in the limit, the proportion of eigenvalues required to reach a precision of γ in the exponential is given by

$$P_\lambda \triangleq P\left(e^{\frac{\lambda_i(X)-\lambda^{\max}(X)}{\mu}} \leq \gamma\right) = \int_{-2\sigma}^{2\sigma+\epsilon\frac{\log\gamma}{\log n}} \frac{1}{2\pi\sigma^2}\sqrt{4\sigma^2-x^2}dx.$$

Since the problems under consideration are relaxations of sparse principal component analysis (PCA), we can also consider the case where $X \in \mathbf{S}_n$ is sampled from the Wishart distribution. In that case, the eigenvalues are distributed according to the Marčenko–Pastur distribution (see [10]), and the above proportion becomes

$$P_\lambda = P\left(e^{\frac{\lambda_i(X)-\lambda^{\max}(X)}{\mu}} \leq \gamma\right) = \int_{-2\sigma}^{2\sigma+\epsilon\frac{\log\gamma}{\log n}} \frac{\sqrt{x(4\sigma-x)}}{2\pi x}dx.$$

With $n = 5000$, $\gamma = 10^{-6}$, and $\epsilon = 10^{-2}$, we get $nP_\lambda = 2.3$, so the approximations above would suggest that, in theory, it is only necessary to compute about three eigenvalues per iteration to get an approximation with precision $\gamma = 10^{-6}$. In practice, however, the results of section 4 show that these rough estimates should be significantly tempered.

3.3. Global complexity bound. Let us now focus on the following program:

$$\begin{array}{ll} \text{minimize} & \lambda^{\max}\left(A^Ty+c\right) \\ \text{subject to} & \|y\| \leq \beta, \end{array} \tag{3.6}$$

where the set Q is here explicitly given by

$$Q = \{y \in \mathbf{R}^p : \ \|y\| \leq \beta\}$$

for some $\beta > 0$ with $\|.\|$ the Euclidean norm here. We can pick $\|x\|^2/2$ as a prox-function for Q, which is strongly convex with convexity parameter 1. Let $\lambda(X) \in \mathbf{R}^n$ be the eigenvalues of the matrix $X = A^Ty + c$, in decreasing order, with $u_i(X) \in \mathbf{R}^n$ an orthonormal set of eigenvectors. The gradient matrix of $\exp(X/\mu)$ is written:

$$\nabla f_\mu(X) = \left(\sum_{i=1}^n e^{\frac{\lambda_i(X)}{\mu}}\right)^{-1} \sum_{i=1}^n e^{\frac{\lambda_i(X)}{\mu}} u_i(X)u_i(X)^T.$$

Suppose we compute only the first m eigenvalues and use them to approximate this gradient by

$$\tilde{\nabla} f_\mu(X) = \left(\sum_{i=1}^m e^{\frac{\lambda_i(X)}{\mu}}\right)^{-1} \sum_{i=1}^m e^{\frac{\lambda_i(X)}{\mu}} u_i(X)u_i(X)^T,$$

and we get the following bound on the error:

$$\|\nabla f_\mu(X) - \tilde{\nabla} f_\mu(X)\| \leq \frac{\sqrt{2}(n-m)e^{\frac{\lambda_m(X)-\lambda_1(X)}{\mu}}}{\left(\sum_{i=1}^m e^{\frac{\lambda_i(X)-\lambda_1(X)}{\mu}}\right)}.$$

In this case, with $X = A^Ty - c$ here, condition (2.3) means that we only need to compute m eigenvalues with m such that

$$\frac{\sqrt{2}(n-m)e^{\frac{\lambda_m(X)-\lambda_1(X)}{\mu}}}{\left(\sum_{i=1}^m e^{\frac{\lambda_i(X)-\lambda_1(X)}{\mu}}\right)} \leq \frac{\delta}{\sigma^{\max}(A)}, \tag{3.7}$$

where $\sigma^{\max}(A)$ is the largest singular value of the matrix A. Using the result in [16], if we define $\|A\| = \max_{\|h\|=1} \|Ah\|_2$ and set $\delta = \epsilon/6$, the algorithm in section 2 will then converge to an ϵ solution of problem (3.6) in at most

$$\frac{4\|A\|\beta}{\epsilon}\sqrt{\log n} \tag{3.8}$$

iterations. This bound on the number of iterations is independent of m in condition (3.7), i.e., the number of eigenvalues required at each iteration. The cost per iteration, however, varies with problem structure as each iteration requires computing m leading eigenvalues, which can be performed in $O(mn^2)$ operations. Note that partial eigenvalue decompositions only access the matrix through matrix-vector products (see [8]); hence, they can handle sparse problems very efficiently. The threshold δ can be adjusted empirically to trade off between the number of iterations and the numerical cost of each iteration. Unfortunately, we can't directly infer a bound on m from the structure of A, so in the next section we study the link between m and the matrix spectrum in numerical examples.

4. Examples and numerical performance. In this section, we illustrate the behavior of the approximate gradient algorithm on various semidefinite optimization problems. Overall, while there appears to be a direct link between problem structure and complexity (i.e., the number of eigenvalues required in the gradient approximation) in the first sparse PCA example discussed below, we will observe on random maximum eigenvalue minimization problems that predicting complexity based on overall problem structure remains an open numerical question in general.

4.1. Sparse principal component analysis. Based on the results in [3], the problem of finding a sparse leading eigenvector of a matrix $C \in \mathbf{S}_n$ can be written

$$\begin{array}{ll} \text{maximize} & x^T C x \\ \text{subject to} & \|x\|_2 = 1, \\ & \mathbf{Card}(x) \leq k, \end{array} \tag{4.1}$$

where $\mathbf{Card}(x)$ is the number of nonzero coefficients in x, and admits the following semidefinite relaxation:

$$\begin{array}{ll} \text{maximize} & \mathbf{Tr}(CX) - \rho \mathbf{1}^T |X| \mathbf{1} \\ \text{subject to} & \mathbf{Tr}(X) = 1, \\ & X \succeq 0, \end{array} \tag{4.2}$$

which is a semidefinite program in the variable $X \in \mathbf{S}^n$, where $\rho > 0$ is the penalty controlling the sparsity of the solution. Its dual is given by

$$\begin{array}{ll} \text{minimize} & \lambda^{\max}(C + U) \\ \text{subject to} & |U_{ij}| \leq \rho, \quad i, j = 1, \ldots, n, \end{array} \tag{4.3}$$

which is of the form (3.1) with

$$Q = \{U \in \mathbf{S}_n : \ |U_{ij}| \leq \rho, \ i, j = 1, \ldots, n\}.$$

The smooth algorithm detailed in section 2 is explicitly described for this problem in [3] and implemented in a numerical package called DSPCA which we have used in the examples here. To test its performance, we generate a matrix M with uniformly distributed coefficients in $[0, 1]$. We let $e \in \mathbf{R}^{250}$ be a sparse vector with

$$e = (1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, \ldots).$$

We then form a test matrix $C = M^T M + vee^T$, where v is a signal-to-noise ratio.

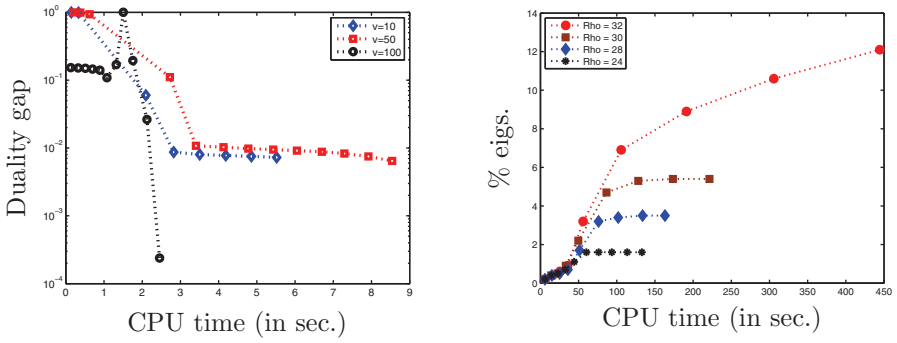


FIG. 4.1. *Left: Duality gap versus CPU time for various values of the signal-to-noise ratio v. Right: Percentage of eigenvalues required versus CPU time, for various values of the penalty parameter ρ controlling sparsity.*

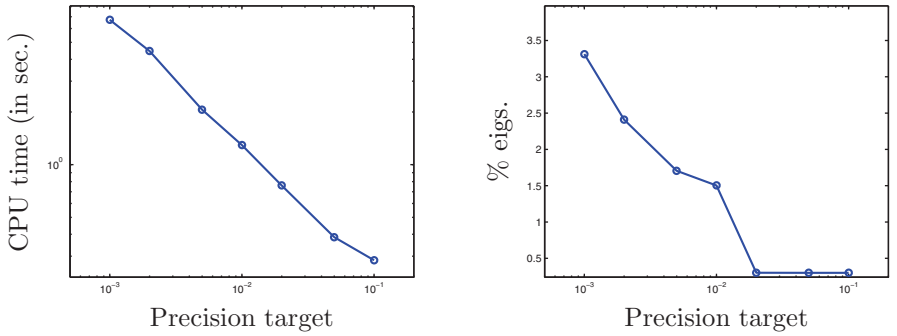


FIG. 4.2. *Left: CPU time (in seconds) versus target precision in loglog scale. Right: Average percentage of eigenvalues required at each iteration versus target precision, in semilog scale.*

In Figure 4.1 on the left, we plot duality gap versus CPU time used for values of the signal-to-noise ratio v ranging from 10 to 100. In Figure 4.1 on the right, we plot the number of eigenvalues required against computing time using a covariance matrix of dimension $n = 500$ sampled from the colon cancer data set in [1] and a noisy rank one matrix. Finally, we measure total computing time versus problem dimension n on this same data set, by solving problem (4.2) for increasingly large submatrices of the original covariance matrix. In each of these examples, we stop after the duality gap has been reduced by 10^{-2}, which is enough here to identify sparse principal components. In Figure 4.2 on the left, we plot computing time versus target precision in loglog scale, on a sparse PCA problem of size 200 extracted from the colon cancer data set. In the previous section, we have seen that precision impacts computing time both through the total number of iterations in (3.5) and through condition (3.7) on the number of eigenvalues required in the gradient approximation. In this example, we observe that CPU time increases a little bit slower than the upper bound of $O(1/\epsilon)$ given in (4.2). In Figure 4.2 on the right, we plot the average percentage of eigenvalues required at each iteration versus target precision, in semilog scale. We observe, on this example of dimension 200, that for low target precisions one eigenvalue is often enough to approximate the gradient but that this number quickly increases for higher precision targets. Note that in all cases, the precision targets are significantly lower than those achieved by interior point methods (usually at least 10^{-8}), but the cost

TABLE 4.1
CPU time (in seconds) versus problem dimension n for full and partial eigenvalue matrix exponential computations.

	$n = 100$	$n = 200$	$n = 500$
Rank one, full	3.2	8.0	14.7
Rank one, **partial**	0.4	0.75	1.6
Colon, full	2.6	18.1	274.3
Colon, **partial**	0.3	1.3	17.7

per iteration and storage requirements of the first-order algorithms detailed here are also significantly lower.

In Table 4.1, we then compare total CPU time using a full precision matrix exponential against CPU time using only a partial eigenvalue decomposition to approximate this exponential. Note that other classic methods for computing the matrix exponential, such as Padé approximations (see [11]), did not provide a significant performance benefit and are not included here. Both exact and approximate gradient codes are fully written in C, with partial eigenvalue decompositions computed using the FORTRAN package ARPACK (see [8]) with calls to vendor-optimized BLAS and LAPACK for matrix operations. To improve stability, the size of the Lanczos basis in ARPACK was set at four times the number of eigenvectors required. We observe that, on these problems, the partial eigenvalue decomposition method is about ten times faster.

4.2. Matrix structure and complexity: Open numerical issues. The previous section showed how the spectrum of the current iterate impacts the complexity of the algorithm detailed in this paper: a steeply decreasing spectrum allows fewer eigenvalues to be computed in the matrix exponential approximation, and a wider gap between eigenvalues improves the convergence rate of these eigenvalue computations. In this section, we study the number of eigenvalues required in randomly generated maximum eigenvalue minimization problems. Because of the measure concentration phenomenon, there is nothing really random about the spectrum of large-scale, naively generated semidefinite optimization problems, so we begin by detailing a simple method for generating random matrices with a given spectrum.

Generating random matrices with a given spectrum. Suppose $X \in \mathbf{S}_n$ is a matrix with normally distributed coefficients, $X_{ij} \sim \mathcal{N}(0,1)$, $i, j = 1, \ldots, n$. If we write its QR decomposition, $X = QR$ with $Q,\ R \in \mathbf{R}^{n\times n}$, then the orthogonal matrix Q is Haar distributed on the orthogonal group $\mathcal{O}_n$ (see [4], for example). This means that to generate a random matrix with given spectrum $\lambda \in \mathbf{R}^n$, we generate a normally distributed matrix X, compute its QR decomposition, and the matrix $Q\,\mathrm{diag}(\lambda)Q^T$ will be uniformly distributed on the set of symmetric matrices with spectrum λ.

Maximum eigenvalue minimization. We now form random maximum eigenvalue minimization problems and then study how the number of required eigenvalues in the gradient computation evolves as the solution approaches optimality. We solve the following problem:

$$\begin{array}{ll} \text{minimize} & \lambda^{\max}\left(A^T y + c\right) \\ \text{subject to} & \|y\| \leq \beta, \end{array}$$

in the variable $y \in \mathbf{R}^m$, where $c \in \mathbf{R}^{n^2}$, $A \in \mathbf{R}^{m\times n^2}$, and $\beta > 0$ is an upper bound on the norm of the solution. In Figure 4.3 we plot percentage of eigenvalues required in the gradient computation versus duality gap for randomly generated

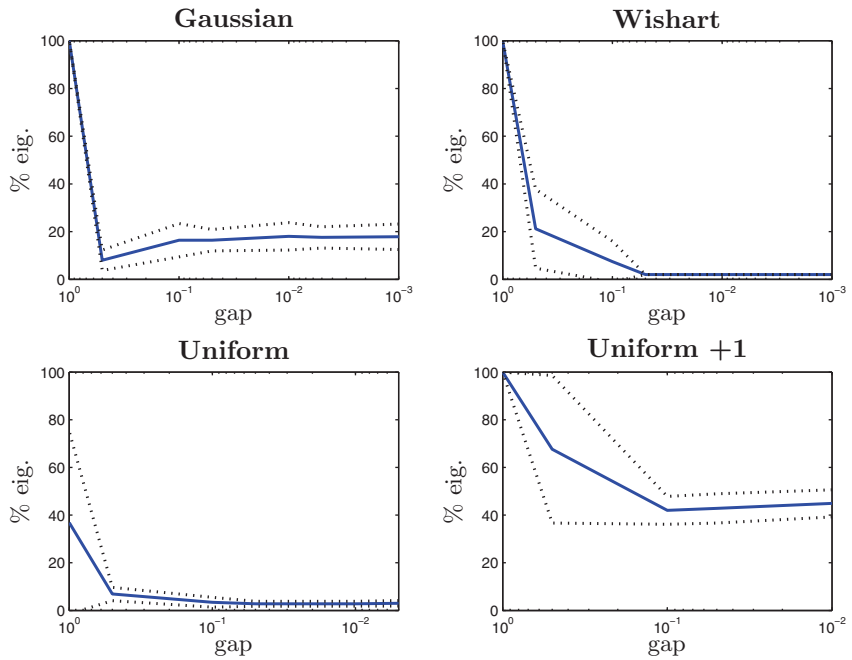


FIG. 4.3. *Average percentage of eigenvalues required (solid line) versus duality gap on randomly generated maximum eigenvalue minimization problems, for various problem matrix distributions. Dashed lines are at plus and minus one standard deviation.*

problem instances where $n = 50$ and $m = 25$. The first two plots use data matrices with Gaussian and Wishart distributions, whose spectrum are distributed according to Wigner's semicircle law and the Marčenko–Pastur distribution, respectively. The last two plots use the procedure described above to generate matrices with uniform spectrum on $[0, 1]$ and uniform spectrum on $[0, 1]$ with one eigenvalue set to 5. We observe that the number of eigenvalues required in the algorithm varies significantly with matrix spectrum.

Problem structure and effective complexity. The results on sparse PCA in section 4.1 and on the random problems of this section show that problem structure has a significant impact on performance. Predicting how many eigenvalues will be required at each iteration based on structural properties of the problem is an important but difficult question. In particular, the number of eigenvalues required in the Gaussian case is much higher than what the asymptotic analysis in section 3.2 predicted. Furthermore, in the sparse PCA example, complexity seems to vary with problem structure somewhat intuitively: a higher signal-to-noise ratio means lower complexity, and a higher sparsity target means higher complexity. However, this is not the case in the random problems studied here; two unstructured problems (uniform and Wishart) have low complexity, while one requires computing many more eigenvalues per iteration (Gaussian), and a more structured example (uniform plus rank one) also requires many eigenvalues. Overall then, predicting effective complexity (i.e., the number of eigenvalues required at each iteration) based on problem structure remains a difficult open question at this point.

Also, it is well known empirically (see [17], [9], and [18], among others) that the largest eigenvalues of $A^T y - c$ in (3.1) tend to coalesce near the optimum, thus

potentially increasing the number of eigenvalues required when computing $\tilde{\nabla} f(x)$ and the number of iterations required for computing leading eigenvalues (see [5], for example), but in these references, too, no a priori link between coalescence and problem structure is established. This coalescence phenomenon is never apparent in the numerical examples studied here, perhaps because it appears only at the much higher precision targets reached by interior point methods.

Acknowledgments. The author would like to thank Noureddine El Karoui for very useful comments.

REFERENCES

[1] A. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays*, Cell Biology, 96 (1999), pp. 6745–6750.

[2] A. Ben-Tal and A. Nemirovski, *Non-Euclidean restricted memory level method for large-scale convex optimization*, Math. Program., 102 (2005), pp. 407–456.

[3] A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. G. Lanckriet, *A direct formulation for sparse PCA using semidefinite programming*, SIAM Rev., 49 (2007), pp. 434–448.

[4] P. Diaconis, *Patterns in eigenvalues: The 70th Josiah Willard Gibbs lecture*, Bull. Amer. Math. Soc., 40 (2003), pp. 155–178.

[5] G. H. Golub and C. F. Van Loan, *Matrix computation*, North Oxford Academic, Oxford, 1990.

[6] C. Helmberg and F. Rendl, *A spectral bundle method for semidefinite programming*, SIAM J. Optim., 10 (2000), pp. 673–696.

[7] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms*, Springer, New York, 1993.

[8] R. B. Lehoucq, D. C. Sorensen, and C. Yang, *ARPACK: Solution of Large-scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, SIAM, Philadelphia, 1998.

[9] A. S. Lewis and M. L. Overton, *Eigenvalue optimization*, Acta Numer., 5 (1996), pp. 149–190.

[10] V. A. Marčenko and L. A. Pastur, *Distribution of eigenvalues for some sets of random matrices*, Math. Sb., 1 (1967), pp. 457–483.

[11] C. Moler and C. Van Loan, *Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later*, SIAM Rev., 45 (2003), pp. 3–49.

[12] A. Nemirovski, *Prox-method with rate of convergence $O(1/T)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems*, SIAM J. Optim., 15 (2004), pp. 229–251.

[13] Y. Nesterov, *A method of solving a convex programming problem with convergence rate $O(1/k^2)$*, Soviet Math. Doklady, 27 (1983), pp. 372–376.

[14] Y. Nesterov, *Introductory Lectures on Convex Optimization*, Springer, New York, 2003.

[15] Y. Nesterov, *Smooth minimization of non-smooth functions*, Math. Program., 103 (2005), pp. 127–152.

[16] Y. Nesterov, *Smoothing technique and its applications in semidefinite optimization*, Math. Program., 110 (2007), pp. 245–259.

[17] M. L. Overton, *Large scale optimization of eigenvalues*, SIAM J. Optim., 2 (1992), pp. 88–120.

[18] G. Pataki, *On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues*, Math. Oper. Res., 23 (1998), pp. 339–358.

[19] N. Z. Shor, *Minimization Methods for Non-differentiable Functions*, Springer, Berlin, 1985.

SIAM J. OPTIM.
Vol. 19, No. 3, pp. 1184–1210

A NEW UNBLOCKING TECHNIQUE TO WARMSTART INTERIOR POINT METHODS BASED ON SENSITIVITY ANALYSIS*

JACEK GONDZIO† AND ANDREAS GROTHEY†

Abstract. One of the main drawbacks associated with Interior Point Methods (IPMs) is the perceived lack of an efficient warmstarting scheme which would enable the use of information from a previous solution of a similar problem. Recently there has been renewed interest in the subject. A common problem with warmstarting for IPM is that an advanced starting point which is close to the boundary of the feasible region, as is typical, might lead to blocking of the search direction. Several techniques have been proposed to address this issue. Most of these aim to lead the iterate back into the interior of the feasible region—we classify them as either "modification steps" or "unblocking steps" depending on whether the modification is taking place before solving the modified problem to prevent future problems, or during the solution if and when problems become apparent. A new "unblocking" strategy is suggested which attempts to directly address the issue of blocking by performing sensitivity analysis on the Newton step with the aim of increasing the size of the step that can be taken. This analysis is used in a new technique to warmstart interior point methods: we identify components of the starting point that are responsible for blocking and aim to improve these by using our sensitivity analysis. The relative performance of a selection of different warmstarting techniques suggested in the literature and the new proposed unblocking by sensitivity analysis is evaluated on the warmstarting test set based on a selection of NETLIB problems proposed by [Benson and Shanno, *Comput. Optim. Appl.*, 38 (2007), pp. 371–399]. Warmstarting techniques are also applied in the context of solving nonlinear programming problems as a sequence of quadratic programs solved by interior point methods. We also apply the warmstarting technique to the problem of finding the complete efficient frontier in portfolio management problems (a problem with 192 million variables—to our knowledge the largest problem to date solved by a warmstarted IPM). We find that the resulting best combined warmstarting strategy manages to save between 50 and 60% of interior point iterations, consistently outperforming similar approaches reported in current optimization literature.

Key words. interior-point methods, warm-start, quadratic programming

AMS subject classifications. 90C51, 90C20, 65K05

DOI. 10.1137/060678129

1. Introduction.

Since their introduction, Interior Point Methods (IPMs) have been recognized as an invaluable tool to solve linear, quadratic, and nonlinear programming problems, in many cases outperforming traditional simplex and active set-based approaches. This is especially the case for large scale problems. One of the weaknesses of IPMs is, however, that unlike their active set-based competitors, they cannot easily exploit an advanced starting point obtained from the preceding solution process of a similar problem. Many optimization problems require the solution of a sequence of closely related problems, either as part of an algorithm (e.g., SQP, Branch & Bound) or as a direct application to a problem (e.g., finding the efficient frontier in portfolio optimization). Because of their weakness in warmstarting, IPMs have not made as big an impact in these areas.

Over the years there have been several attempts to improve the warmstarting capabilities of IPMs [5, 8, 15, 6, 1, 2, 10]. All of these, apart from [1, 2], involve

*Received by the editors December 19, 2006; accepted for publication (in revised form) April 23, 2008; published electronically November 19, 2008.
http://www.siam.org/journals/siopt/19-3/67812.html
†School of Mathematics, University of Edinburgh, Mayfield Road, Edinburgh EH9 3JZ United Kingdom (J.Gondzio@ed.ac.uk, http://www.maths.ed.ac.uk/~gondzio/, A.Grothey@ed.ac.uk, http://www.maths.ed.ac.uk/~agr/).

remembering a primal/dual iterate encountered during the solution of the original problem and using this (or some modification of it) as a starting point for the modified problem. All of these papers (apart from [2]) deal with the linear programming (LP) case, whereas we are equally interested in the quadratic programming (QP) case.

A typical way in which a 'bad' starting point manifests itself is *blocking*: The Newton direction from this point leads far outside the positive orthant, resulting in only a very small fraction of it to be taken. Consequently, the next iterate will be close to the previous one, and the search direction will likely block again. In our observation this blocking is usually due only to a small number of components of the Newton direction. We therefore suggest an unblocking strategy which attempts to modify these blocking components without disturbing the primal-dual direction too much. The unblocking strategy is based on performing sensitivity analysis of the primal-dual direction with respect to the components of the current primal/dual iterate.

As a separate thread to the paper, it is our feeling that a wealth of warmstarting heuristics have been proposed by various authors, each demonstrating improvements over a coldstarted IPM. However, there has been no attempt at comparing these in a unified environment, or indeed investigating how these might be combined. This paper will give an overview of some of the warmstarting techniques that have been suggested and explore what benefit can be obtained from combining them.

This will also set the scene for evaluating the new unblocking strategy derived in this paper, within a variety of different warmstarting settings.

We continue by stating the notation used in this paper. In section 3, we review traditionally used warmstart strategies. In section 4 we present the new unblocking techniques based on sensitivity analysis. Numerical comparisons as to the efficiency of the suggested techniques are reported in section 5. In section 6, we draw our conclusions.

2. Notation and background. The infeasible primal dual interior point methods applied to solve the quadratic programming problem

$$\begin{aligned} \min \quad & c^T x + \frac{1}{2} x^T Q x \\ \text{s.t.} \quad & Ax = b \\ & x \geq 0 \end{aligned} \tag{1}$$

can be motivated from the KKT conditions for (1)

$$c + Qx - A^T y - z = 0 \tag{2a}$$

$$Ax = b \tag{2b}$$

$$XZe = \mu e \tag{2c}$$

$$x, z \geq 0, \tag{2d}$$

where the zero right-hand side of the complementary products has been replaced by the centrality parameter $\mu > 0$. The set of solutions to (2) for different values of μ is known as the *central path*. It is beneficial in this context to consider two neighborhoods of the central path, the N_2 neighborhood

$$N_2(\theta) := \left\{ (x, y, z) : Ax = b, A^T y - Qx + z = c, \|XZe - \mu e\|_2 \leq \theta \right\}$$

and the wider $N_{-\infty}$ neighborhood

$$N_{-\infty}(\gamma) := \left\{ (x, y, z) : Ax = b, A^T y - Qx + z = c, x_i z_i \geq \gamma \mu \right\}.$$

Assume that at some stage during the algorithm the current iterate is (x, y, z). Our variant of the predictor-corrector algorithm [4, 7] will calculate a predictor direction $(\Delta x_p, \Delta y_p, \Delta z_p)$ as the Newton direction for system (2) and a small μ-target $(\mu^0 \approx 0.001\frac{x^T z}{n})$:

$$\begin{array}{lll} -Q\Delta x_p + A^T\Delta y_p + \Delta z_p = c + Qx - A^T y - z = \xi_c \\ A\Delta x_p \qquad\qquad\quad = b - Ax \qquad\quad = \xi_b \\ X\Delta z_p + Z\Delta x_p \qquad = \mu^0 e - XZe \qquad = r_{xz}, \end{array} \tag{3}$$

which can be further condensed by using the third equation to eliminate Δz_p

$$\begin{bmatrix} -Q - X^{-1}Z & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x_p \\ \Delta y_p \end{bmatrix} = \begin{bmatrix} r_x \\ r_y \end{bmatrix} := \begin{bmatrix} \xi_c - X^{-1} r_{xz} \\ \xi_b \end{bmatrix} \tag{4a}$$

$$\Delta z_p = X^{-1} r_{xz} - X^{-1} Z \Delta x_p. \tag{4b}$$

As in Mehrotra's predictor-corrector algorithm [13], we calculate maximal primal and dual stepsizes for the predictor direction

$$\bar{\alpha}_p = \max\{\alpha > 0 : x + \alpha\Delta x_p \geq 0\}, \quad \bar{\alpha}_d = \max\{\alpha > 0 : z + \alpha\Delta z_p \geq 0\}$$

and determine a target μ-value by

$$\mu = \frac{[(x + \bar{\alpha}_p\Delta x_p)^T (z + \bar{\alpha}_d \Delta z_p)]^3}{n(x^T z)^2}.$$

With these we compute the corrector direction $(\Delta x_c, \Delta y_c, \Delta z_c)$ by

$$\begin{array}{ll} A^T\Delta y_c + \Delta z_c &= 0 \\ A\Delta x_c &= 0 \\ X\Delta z_c + Z\Delta x_c &= (\mu - \mu^0)e - \Delta X_p \Delta Z_p e, \end{array} \tag{5}$$

and finally the new primal and dual stepsizes and the new iterate (x^+, z^+) as

$$\begin{aligned} \alpha_p &= 0.995 \max\{\alpha > 0 : x + \alpha(\Delta x_p + \Delta x_c) \geq 0\} \\ \alpha_d &= 0.995 \max\{\alpha > 0 : z + \alpha(\Delta z_p + \Delta z_c) \geq 0\} \\ x^+ &= x + \alpha_p(\Delta x_p + \Delta x_c), \quad z^+ = z + \alpha_d(\Delta z_p + \Delta z_c). \end{aligned}$$

Our main interest is generating a good starting point for the QP problem (1) — the *modified problem* — from the solution of a previously solved similar QP problem

$$\begin{array}{ll} \min & \tilde{c}^T x + \frac{1}{2} x^T \tilde{Q} x \\ \text{s.t.} & \tilde{A}x = \tilde{b} \\ & x \geq 0, \end{array} \tag{6}$$

the *original problem*. The difference between the two problems, i.e., the change from the original problem to the second problem, is denoted by

$$(\Delta A, \Delta Q, \Delta c, \Delta b) = \big(A - \tilde{A}, Q - \tilde{Q}, c - \tilde{c}, b - \tilde{b}\big).$$

3. Warmstart heuristics. Unlike the situation in the Simplex Method, for IPMs it is not a good strategy to use the optimal solution of a previously solved problem as the new starting point for a similar problem. This is because problems are often ill-conditioned; hence the final solution of the original problem might be far away from the central path of the modified problem. Furthermore, [9] demonstrates that the predictor direction tends to be parallel to nearby constraints, resulting in difficulties to drop misidentified nonbasic variables.

Over the years numerous contributions [11, 5, 8, 15, 6] have addressed this problem, with renewed interest in the subject from [1, 2, 10] over the last year. With the exception of [1, 2] which use an L_1-penalty reformulation of the problem that has better warmstarting capabilities, all remedies follow a common theme: They identify an advanced center [5], a point close to the central path of the original problem (usually a nonconverged iterate), and **modify** it in such a manner that the modified point is close to the central path of the new problem. Further, in the first few iterations of the reoptimization, additional techniques which address the issue of getting stuck at nearby constraints may be employed. In this paper these will be called **unblocking heuristics**. The generic IPM warmstarting algorithm is as follows:

Algorithm: Generic Interior Point Warmstart

1. Solve the original problem (6) by an Interior Point Algorithm. From it choose one of (or a selection of) the iterates $(\tilde{x}, \tilde{y}, \tilde{z}, \tilde{\mu})$ encountered during the solution process. We will assume that this iterate (or any one of these iterates) satisfies

$$\begin{aligned} \tilde{c} + \tilde{Q}\tilde{x} - \tilde{A}^T\tilde{y} - \tilde{z} &= 0 \\ \tilde{b} - \tilde{A}\tilde{x} &= 0 \\ \tilde{x}_i\tilde{z}_i \approx \tilde{\mu} \qquad & \forall i = 1, \ldots, n. \end{aligned}$$

2. **Modify** the chosen iterate to obtain a starting point (x, y, z, μ) for the modified problem.

3. Solve the modified problem by an Interior Point Algorithm using (x, y, z, μ) as the starting point. During the first few iterations of the IPM a special **unblocking step** might be taken.

The question arises as to what should guide the construction of modification and unblocking steps. It is well known that for a feasible method (i.e., $\xi_b = \xi_c = 0$), a well-centered point (i.e., in $N_2(\theta)$ or $N_{-\infty}(\gamma)$) and a small target decrease ($\mu \lessapprox \mu^0$), and the Newton step is feasible. Analysis by [15] and [6] identifies two factors that lead to the ability of IPMs to absorb infeasibilities ξ_b, ξ_c present at the starting point. Firstly, the larger the value of μ the more infeasibility can be absorbed in one step. Secondly, the centrality of the iterate: from a well-centered point the IPM can again absorb more infeasibilities. Using these general guidelines, a number of different warmstarting techniques have been suggested. We review some of them here:

Modification Steps:

(i) *Shift small components*: [11] shift $\tilde{x}, \tilde{z}$ by $h_x = \epsilon D^{-1}e, h_z = \epsilon De$, where $D = \text{diag}\{\|a_j\|_1\}$ and a_j is the jth column of A to ensure $x_i z_i \geq \gamma\mu$ for some small $\gamma > 0$, i.e., improve centrality by aiming for a point in $N_{-\infty}(\gamma)$.

(ii) References [15, 10] suggest a *Weighted Least Squares Step* (WLS) that finds the minimum step (with respect to a weighted 2-norm) from the starting point, to

a point that is both primal and dual feasible. The WLS step does not necessarily preserve positiveness of the iterate. To overcome this, [15] suggests keeping a selection of potential warmstart iterates and retracing to one corresponding to a large μ, which will guarantee that the WLS step is feasible. Since we do not want to remember several different points from the solution of the original problem, we will take a fraction of the WLS step should the full step be infeasible. Mehrotra's starting point [13] can be seen as a (damped) WLS step from the origin.

(iii) References [15, 10] further suggest a *Newton Modification Step*, i.e., an interior point step (3) correcting only for the primal and dual infeasibilities introduced by the change of problem, with no attempt to improve centrality: (3) is solved with $r_{xz} = 0$. Again only a fraction of this step might be taken.

Unblocking Heuristics

(i) *Splitting Directions*: Reference [6] advocates computing separate search directions aimed at achieving primal feasibility, dual feasibility, and centrality separately. These are combined into the complete step by taking the maximum of each step that can be taken without violating the positivity of the iterates. A possible interpretation of this strategy is to emulate a gradual change from the original problem to the modified problem where for each change the modification step is feasible.

(ii) *Higher Order Correctors*: The $\Delta X_p \Delta Z_p$ component in (5) is a correction for the linearization error in $XZe - \mu e = 0$. A corrector of this type can be repeated several times. Reference [5] employs this idea by additionally correcting only for small complementary products to avoid introducing additional blocking. This is used in [6] as an unblocking technique with the interpretation of choosing a target complementary vector $\bar{t} \approx \mu e$ in such a way that a large step in the resulting Newton direction is feasible, aiming to absorb as much of the primal/dual infeasibility as possible in the first step.

(iii) *Change Diagonal Scaling:* Reference [9] investigates changing elements in the scaling matrix $\Theta = XZ^{-1}$ to make nearby constraints repelling rather than attracting to the Newton step. However, we are not aware of any implementation of this technique in a warmstarting context.

A number of additional interesting techniques are listed here and described below:

(i) *Dual adjustment*: Adjust advanced starting point $\tilde{z}$ to compensate for changes to c, A, and Q in the dual feasibility constraint (2a).

(ii) *Additional centering* iterations before the advanced starting point is used.

(iii) Unblocking of the step direction by *sensitivity analysis*.

We will give a brief description of the first two of these strategies. The third (unblocking by sensitivity analysis) is the subject of section 4.

Dual adjustment

Using $(\tilde{x}, \tilde{y}, \tilde{z})$ as a starting point in problem (1) will result in the initial dual infeasibility

$$\xi_c = c + Q\tilde{x} - A^T\tilde{y} - \tilde{z} = \Delta c + \Delta Q\tilde{x} - \Delta A^T\tilde{y}.$$

Setting $z = \tilde{z} + \Delta z$, where $\Delta z = \Delta c + \Delta Q\tilde{x} - \Delta A^T\tilde{y}$, would result in a point satisfying the dual feasibility constraint (2a). However, the conditions $z \geq 0$ and $x_i z_i \approx \mu$ are likely violated by this, so instead we set

$$z_i = \max\{\tilde{z}_i + \Delta z_i, \min\{\sqrt{\mu}, \tilde{z}_i/2\}\};$$

i.e., we try to absorb as much of the dual infeasibility into z as possible without decreasing z either below $\sqrt{\mu}$ or half its value.

Adjusting the saved iterate $(\tilde{x}, \tilde{y}, \tilde{z})$ in a minimal way to absorb primal/dual infeasibilities is similar in spirit to the WLS modification step. Unlike this, however, direct adjustment of z is much cheaper to compute.

Additional centering iterations

The aim of improving the centrality of the saved iterate can also be achieved by performing an additional pure centering iteration, i.e., choose $\xi_c = \xi_b = 0, \mu^0 = x^T z/n$ in (3), in the original problem before saving the iterate as a starting point for the new problem. This pure centering iteration could be performed with respect to the original or the modified problem. In the latter case, this is similar in spirit to the Newton Modification Step of [15, 10] (whereas [15, 10] use $r_{xz} = 0$, we use $r_{xz} = \mu^0 e - \tilde{X}\tilde{Z}$ with $\mu^0 = \tilde{x}^T\tilde{z}/n$. In the case of a perfectly centered saved iterate—as we hope to achieve at least approximately by the previous centering in the original problem—these two are identical). We refer to these as centering iteration at the *beginning* of solving the modified problem or at the *end* of solving the original problem.

In the next section we will derive the unblocking strategy based on sensitivity analysis.

4. Unblocking by sensitivity analysis.

4.1. Sensitivity analysis. In this section we will lay the theoretical foundations for our proposed unblocking strategy. Much of it is based on the observation that the advanced starting information (x, y, z, μ) with which to start the solution of the modified problem is to some degree arbitrary. It is therefore possible to treat it as parameters to the solution process and to explore how certain properties of the solution process change as the starting point is changed. In particular we are interested in the primal and dual stepsizes that can be taken for the Newton direction computed from this point.

At some iterate (x, y, z) of the IPM, the primal-dual direction $(\Delta x, \Delta y, \Delta z)$ is obtained as the solution to the system (3) or (4) for some target value μ^0. If we think of (x, y, z) as the advanced starting point, the step $(\Delta x, \Delta y, \Delta z)$ can be obtained as a function of the current point (x, y, z). The aim of this section is to derive a procedure by which the sensitivity of $\Delta x(x, y, z)$, $\Delta y(x, y, z)$, $\Delta z(x, y, z)$, that is the first derivatives of these functions can be computed.

First note that the value of y has no influence on the new step $\Delta x, \Delta z$. This is because after substituting for ξ_b, ξ_c, r_{xz} in (4a)

$$\begin{bmatrix} -Q - X^{-1}Z & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = \begin{bmatrix} c + Qx - A^T y - \mu X^{-1} e \\ b - Ax \end{bmatrix}$$

we can rewrite this as

$$\begin{bmatrix} -Q - X^{-1}Z & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ y^{(k+1)} \end{bmatrix} = \begin{bmatrix} c + Qx - \mu X^{-1} e \\ b - Ax \end{bmatrix} \tag{7}$$

with $\Delta y = y^{(k+1)} - y$. In effect (7) solves for the new value of $y^{(k+1)} = y^{(k)} + \Delta y$ directly, whereas all influence of y onto $\Delta x, \Delta z$ has been removed. Notice also that only the step components in x, z variables can lead to a blocking of the step; therefore we are interested only in the functional relationship and sensitivity for the functions $\Delta x = \Delta x(x, z), \Delta z = \Delta z(x, z)$. To this end we start by differentiating with respect to

x_i in (3):

$$-Q\frac{\mathrm{d}\Delta x}{\mathrm{d}x_i} + A^T\frac{\mathrm{d}\Delta y}{\mathrm{d}x_i} + \frac{\mathrm{d}\Delta z}{\mathrm{d}x_i} = Qe_i, \tag{8a}$$

$$A\frac{\mathrm{d}\Delta x}{\mathrm{d}x_i} = -Ae_i, \tag{8b}$$

$$X\frac{\mathrm{d}\Delta z}{\mathrm{d}x_i} + Z\frac{\mathrm{d}\Delta x}{\mathrm{d}x_i} + \Delta Ze_i = -Ze_i. \tag{8c}$$

Note that this result is independent of the value of μ^0 that is used as a target. Similarly, differentiating with respect to y_i yields

$$-Q\frac{\mathrm{d}\Delta x}{\mathrm{d}y_i} + A^T\frac{\mathrm{d}\Delta y}{\mathrm{d}y_i} + \frac{\mathrm{d}\Delta z}{\mathrm{d}y_i} = -A^Te_i \tag{9a}$$

$$A\frac{\mathrm{d}\Delta x}{\mathrm{d}y_i} = 0 \tag{9b}$$

$$X\frac{\mathrm{d}\Delta z}{\mathrm{d}y_i} + Z\frac{\mathrm{d}\Delta x}{\mathrm{d}y_i} = 0, \tag{9c}$$

and finally differentiating with respect to z_i yields

$$-Q\frac{\mathrm{d}\Delta x}{\mathrm{d}z_i} + A^T\frac{\mathrm{d}\Delta y}{\mathrm{d}z_i} + \frac{\mathrm{d}\Delta z}{\mathrm{d}z_i} = -e_i \tag{10a}$$

$$A\frac{\mathrm{d}\Delta x}{\mathrm{d}z_i} = 0 \tag{10b}$$

$$X\frac{\mathrm{d}\Delta z}{\mathrm{d}z_i} + Z\frac{\mathrm{d}\Delta x}{\mathrm{d}z_i} + \Delta Xe_i = -Xe_i. \tag{10c}$$

Taking all three systems together we have

$$\begin{bmatrix} -Q & A^T & I \\ A & 0 & 0 \\ Z & 0 & X \end{bmatrix} \begin{bmatrix} \frac{\mathrm{d}\Delta x}{\mathrm{d}x} & \frac{\mathrm{d}\Delta x}{\mathrm{d}y} & \frac{\mathrm{d}\Delta x}{\mathrm{d}z} \\ \frac{\mathrm{d}\Delta y}{\mathrm{d}x} & \frac{\mathrm{d}\Delta y}{\mathrm{d}y} & \frac{\mathrm{d}\Delta y}{\mathrm{d}z} \\ \frac{\mathrm{d}\Delta z}{\mathrm{d}x} & \frac{\mathrm{d}\Delta z}{\mathrm{d}y} & \frac{\mathrm{d}\Delta z}{\mathrm{d}z} \end{bmatrix} = \begin{bmatrix} Q & -A^T & -I \\ -A & 0 & 0 \\ -Z-\Delta Z & 0 & -X-\Delta X \end{bmatrix}. \tag{11}$$

Under the assumption that A has full row rank, the system matrix is nonsingular, therefore

$$\begin{bmatrix} \frac{\mathrm{d}\Delta x}{\mathrm{d}x_i} \\ \frac{\mathrm{d}\Delta y}{\mathrm{d}x_i} \\ \frac{\mathrm{d}\Delta z}{\mathrm{d}x_i} \end{bmatrix} = \begin{bmatrix} -e_i \\ 0 \\ 0 \end{bmatrix} + \Delta z_i \begin{bmatrix} -Q & A^T & I \\ A & 0 & 0 \\ Z & 0 & X \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 0 \\ -e_i \end{bmatrix} \tag{12a}$$

$$\begin{bmatrix} \frac{\mathrm{d}\Delta x}{\mathrm{d}y_i} \\ \frac{\mathrm{d}\Delta y}{\mathrm{d}y_i} \\ \frac{\mathrm{d}\Delta z}{\mathrm{d}y_i} \end{bmatrix} = \begin{bmatrix} 0 \\ -e_i \\ 0 \end{bmatrix} \tag{12b}$$

$$\begin{bmatrix} \frac{\mathrm{d}\Delta x}{\mathrm{d}z_i} \\ \frac{\mathrm{d}\Delta y}{\mathrm{d}z_i} \\ \frac{\mathrm{d}\Delta z}{\mathrm{d}z_i} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -e_i \end{bmatrix} + \Delta x_i \begin{bmatrix} -Q & A^T & I \\ A & 0 & 0 \\ Z & 0 & X \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 0 \\ -e_i \end{bmatrix}, \tag{12c}$$

where the system common to (12a/12c)

$$\begin{bmatrix} -Q & A^T & I \\ A & 0 & 0 \\ Z & 0 & X \end{bmatrix} \begin{bmatrix} \widetilde{d\Delta x} \\ \widetilde{d\Delta y} \\ \widetilde{d\Delta z} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -e_i \end{bmatrix} \tag{13}$$

can be solved by using the third line to substitute for $\widetilde{d\Delta z}$ as

$$\begin{bmatrix} -Q - X^{-1}Z & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \widetilde{\mathrm{d}\Delta x} \\ \widetilde{\mathrm{d}\Delta y} \end{bmatrix} = \begin{bmatrix} X^{-1}e_i \\ 0 \end{bmatrix} \tag{14a}$$

$$\widetilde{\mathrm{d}\Delta z} = -X^{-1}Z\widetilde{\mathrm{d}\Delta x} - X^{-1}e_i. \tag{14b}$$

There are a few insights to be gained from these formulas. First, they confirm that the step $(\Delta x, \Delta z)$ does not depend on y.

Second, the *sensitivity* of the primal-dual step with respect to the current iterate (x, y, z)—unlike the step $(\Delta x, \Delta y, \Delta z)$ itself—does not depend on the target value μ^0 either. We will exploit this property when constructing a warmstart heuristic that uses the sensitivity information.

Finally we can get the complete sensitivity information with respect to (x_i, z_i) for a given component i by solving a single system of linear equations with the same augmented system matrix that has been used to obtain the step $(\Delta x, \Delta y, \Delta z)$ (and for which a factorization is available); the solution of n such systems will likewise retrieve the complete sensitivity information.

Although this system matrix is already factorized as part of the normal interior point algorithm, and backsolves are an order of magnitude cheaper than the factorization, obtaining the complete sensitivity information is prohibitively expensive. The aim of the following section is therefore to propose a warmstarting heuristic that uses the sensitivity information derived above, but requires only a few, rather than all n backsolves.

4.2. Unblocking the primal-dual direction using sensitivity information. Occasionally, despite all our attempts, a starting point might result in a Newton direction that leads to blocking: i.e., only a very small step can be taken along it. We do not want to abandon the advanced starting information at this point, but rather try to *unblock* the search direction. To this end we will make use of the sensitivity analysis presented in section 4.1. The following Lemma 1 gives conditions under which a step (d_x, d_z) can be expected to unblock based on the sensitivity analysis.

LEMMA 1. *A necessary and sufficient condition for a step (d_x, d_z) to unblock to first order to a given level ρl, i.e.,*

$$x + d_x + \Delta x + \frac{\mathrm{d}\Delta x}{\mathrm{d}x} d_x + \frac{\mathrm{d}\Delta x}{\mathrm{d}z} d_z \geq \rho l, \tag{15a}$$

$$z + d_z + \Delta z + \frac{\mathrm{d}\Delta z}{\mathrm{d}x} d_x + \frac{\mathrm{d}\Delta z}{\mathrm{d}z} d_z \geq \rho l, \tag{15b}$$

is that there exists vectors d_x, d_z, t_x, t_y, t_z of appropriate dimensions that satisfy the system of equations

$$At_x = 0 \tag{16a}$$

$$-Qt_x + A^T t_y + t_z = 0 \tag{16b}$$

$$Zt_x + Xt_z = -\Delta Z d_x - \Delta X d_z \tag{16c}$$

$$t_x \geq -x - \Delta x + \rho l \tag{16d}$$

$$t_z \geq -z - \Delta z + \rho l \tag{16e}$$

Proof. Note that the relations of (11),(12) can be more concisely written as

$$\begin{bmatrix} \frac{\mathrm{d}\Delta x}{\mathrm{d}x} + I & \frac{\mathrm{d}\Delta x}{\mathrm{d}y} & \frac{\mathrm{d}\Delta z}{\mathrm{d}z} \\ \frac{\mathrm{d}\Delta y}{\mathrm{d}x} & \frac{\mathrm{d}\Delta y}{\mathrm{d}y} + I & \frac{\mathrm{d}\Delta y}{\mathrm{d}z} \\ \frac{\mathrm{d}\Delta z}{\mathrm{d}x} & \frac{\mathrm{d}\Delta z}{\mathrm{d}y} & \frac{\mathrm{d}\Delta z}{\mathrm{d}z} + I \end{bmatrix} = - \begin{bmatrix} -Q & A^T & I \\ A & 0 & 0 \\ Z & 0 & X \end{bmatrix}^{-1} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \Delta Z & 0 & \Delta X \end{bmatrix}. \tag{17}$$

Conditions (15) are equivalent to the existence of (d_x, d_y, d_z) such that

$$\begin{bmatrix} \frac{\mathrm{d}\Delta x}{\mathrm{d}x} + I & \frac{\mathrm{d}\Delta x}{\mathrm{d}y} & \frac{\mathrm{d}\Delta z}{\mathrm{d}z} \\ \frac{\mathrm{d}\Delta y}{\mathrm{d}x} & \frac{\mathrm{d}\Delta y}{\mathrm{d}y} + I & \frac{\mathrm{d}\Delta y}{\mathrm{d}z} \\ \frac{\mathrm{d}\Delta z}{\mathrm{d}x} & \frac{\mathrm{d}\Delta z}{\mathrm{d}y} & \frac{\mathrm{d}\Delta z}{\mathrm{d}z} + I \end{bmatrix} \begin{bmatrix} d_x \\ d_y \\ d_z \end{bmatrix} \geq \begin{bmatrix} -x - \Delta x + \rho l \\ -\infty \\ -z - \Delta z + \rho l \end{bmatrix}, \tag{18}$$

where d_y is an arbitrary vector (note that $\frac{\mathrm{d}\Delta x}{\mathrm{d}y} = \frac{\mathrm{d}\Delta z}{\mathrm{d}y} = 0$). This, on the other hand, is satisfied, if and only if there exists (t_x, t_y, t_z) such that

$$\begin{bmatrix} \frac{\mathrm{d}\Delta x}{\mathrm{d}x} + I & \frac{\mathrm{d}\Delta x}{\mathrm{d}y} & \frac{\mathrm{d}\Delta z}{\mathrm{d}z} \\ \frac{\mathrm{d}\Delta y}{\mathrm{d}x} & \frac{\mathrm{d}\Delta y}{\mathrm{d}y} + I & \frac{\mathrm{d}\Delta y}{\mathrm{d}z} \\ \frac{\mathrm{d}\Delta z}{\mathrm{d}x} & \frac{\mathrm{d}\Delta z}{\mathrm{d}y} & \frac{\mathrm{d}\Delta z}{\mathrm{d}z} + I \end{bmatrix} \begin{bmatrix} d_x \\ d_y \\ d_z \end{bmatrix} = \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} \geq \begin{bmatrix} -x - \Delta x + \rho l \\ -\infty \\ -z - \Delta z + \rho l \end{bmatrix}. \tag{19}$$

Now using (17) to substitute for the matrix of derivatives, multiplying both sides of the equality with the augmented system matrix and multiplying out we see that (19) is equivalent to

$$\begin{bmatrix} 0 \\ 0 \\ -\Delta X d_z - \Delta Z d_x \end{bmatrix} = \begin{bmatrix} -Q & A^T & I \\ A & 0 & 0 \\ Z & 0 & X \end{bmatrix} \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix}, \quad \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} \geq \begin{bmatrix} -x - \Delta x + \rho l \\ -\infty \\ -z - \Delta z + \rho l \end{bmatrix}$$

that is to (16). ∎

The sensitivity analysis thus gives us conditions that an unblocking direction needs to satisfy. However it is unclear if a direction (d_x, d_z) and the corresponding (t_x, t_y, t_z) to satisfy the conditions of Lemma 1 exist. We can however prove existence of such a direction by assuming that we know the analytic center $\hat{p} = (\hat{x}, \hat{y}, \hat{z})$ of the problem (or indeed any strictly primal-dual feasible point) and denote by $\underline{\hat{p}}, \hat{\bar{p}}$ its largest and smallest component:

$$0 < \underline{\hat{p}} \leq \hat{x}_i, \hat{z}_i \leq \hat{\bar{p}}$$

LEMMA 2. *For all $l : 0 < l < \min\{\underline{\hat{p}}/4, 1\}$, $\rho < 1$ and fixed μ and γ there exists a $c = c(\gamma, \mu)$ such that for all starting points (x, y, z) and corresponding blocking step*

$(\Delta x, \Delta y, \Delta z)$ *obtained from* (3) *with* $\mu^0 = \mu^+$ *satisfying*

$$x^T z/n = \mu, \quad x_i z_i \geq \gamma\mu, \quad x_i \leq b_u, z_i \leq b_u, \quad \mu^+ \leq \frac{1}{2}\gamma\mu,$$
$$z + \Delta z \geq -le, \quad x + \Delta x \geq -le,$$

there exists a step $(d_x, d_z) : \|d_x, d_z\|_\infty \leq c(1+\rho)l$ *that unblocks to first order to level* ρL*, i.e., that satisfies conditions* (15).

Proof. With $\alpha = 2(1+\rho)l/\underline{\hat{p}}$ set

$$t_x = \alpha(\hat{x} - (x + \Delta x)),$$
$$t_y = \alpha(\hat{y} - (y + \Delta y)),$$
$$t_z = \alpha(\hat{z} - (z + \Delta z)).$$

We will show that (t_x, t_y, t_z) satisfies (16a/b/d/e), that we can construct a corresponding (d_x, d_z) satisfying (16c), and finally that $(t_x, t_y, t_z), (d_x, d_z) = \mathcal{O}((1+\rho)l)$.

First we notice that both $\hat{p} = (\hat{x}, \hat{y}, \hat{z})$ and $(x + \Delta x, y + \Delta y, z + \Delta z)$ are primal and dual feasible (although in the latter case, of course, not positive). Hence, their difference (and therefore (t_x, t_y, t_z)) satisfies (16a/b).

To proof (16d) we need to distinguish the two cases: $x_i + \Delta x_i < \rho l$ and $x_i + \Delta x_i \geq \rho l$. In the first case $x_i + \Delta x_i < \rho l$ we have

$$\hat{x}_i - (x_i + \Delta x_i) \geq \underline{\hat{p}} - \rho l \geq \frac{1}{2}\underline{\hat{p}},$$

where the last inequality is due to $\rho \leq 1$ and $l \leq \underline{\hat{p}}/4$. Then

$$t_{x,i} = \frac{2(1+\rho)l}{\underline{\hat{p}}}(\hat{x}_i - (x_i + \Delta x_i)) \geq \frac{2(1+\rho)l}{\underline{\hat{p}}}\frac{1}{2}\underline{\hat{p}} = (1+\rho)l \geq -x_i - \Delta x_i + \rho l.$$

In the second case $x_i + \Delta x_i \geq \rho l$, we note that $\hat{x}_i \geq \underline{\hat{p}} \geq 4l \geq \rho l$. Since $2(1+\rho) \leq 4$ we have $0 < \alpha \leq 1$, and hence

$$x_i + \Delta x_i + \alpha(\hat{x}_i - (x_i + \Delta x_i)) = (1-\alpha)(x_i + \Delta x_i) + \alpha\hat{x}_i \geq \rho l,$$

which proves (16d). (16e) is proven in the same manner.

Next we establish a bound for $\|t_x\|, \|t_z\|$. Since $\mu^+ < \gamma\mu/2$ we have from the last equation of (3):

$$x_i\Delta z_i + z_i\Delta x_i = \mu^+ - x_i z_i \leq \frac{1}{2}\gamma\mu - \gamma\mu = -\frac{1}{2}\gamma\mu < 0, \tag{20}$$

and hence at least one of $\Delta x_i, \Delta z_i$ must be negative. Assume w.l.o.g. that $\Delta x_i < 0$. Then $x_i + \Delta x_i \geq -l$ implies

$$|\Delta x_i| = -\Delta x_i \leq l + x \leq l + b_u.$$

We can make no further assumptions on the sign of Δz_i. If $\Delta z_i < 0$, then $|\Delta z_i|$ is bounded in the same way as Δx_i. If, on the other hand, $\Delta z_i \geq 0$, then (20) together with $x_i \geq \gamma/z_i > \gamma/b_u$ implies

$$\Delta z_i < -z_i\Delta x_i/x_i < b_u(l + b_u)b_u/\gamma = b_u^2(l + b_u)/\gamma.$$

Since we can reasonably assume that $b_u^2/\gamma > 1$, we have

$$\|\Delta x\|, \|\Delta z\| \le b_u^2(1+b_u)/\gamma.$$

From this we get

$$\|t_x\| = \alpha\|\hat{x} - (x+\Delta x)\| \le \frac{2(1+\rho)l}{\underline{\hat{p}}}(\hat{\bar{p}} + b_u + b_u^2(1+b_u)/\gamma) = c_1(1+\rho)l,$$

where $c_1 = c_1(\gamma) = 2(\hat{\bar{p}} + b_u + b_u^2(1+b_u)/\gamma)/\underline{\hat{p}}$. $\|t_z\| \le c_1(1+\rho)l$ follows in the same manner.

Finally we know from (20) that for all i at least one of $x_i\Delta z_i, z_i\Delta x_i$ must be less than $-\gamma\mu/4$. Assume w.l.o.g. $z_i\Delta x_i < -\gamma\mu/4$, and then we get

$$\Delta x_i < -\frac{\gamma\mu}{4z_i} < 0. \tag{21}$$

Therefore we can set

$$d_{z,i} = -\frac{z_i t_{x,i} + x_i t_{z,i}}{\Delta x_i}, \quad d_{x,i} = 0 \tag{22}$$

(and vice versa if $x_i\Delta z_i < -\gamma\mu/4$) to construct a direction (d_x, d_z) that satisfies (16c). It remains to be shown that $(d_x, d_z) = \mathcal{O}((1+\rho)l)$:

From (21) we know

$$|\Delta x_i| = -\Delta x_i > \frac{\gamma\mu}{4z_i} > \frac{\gamma\mu}{4b_u};$$

hence (22) gives

$$|d_{z,i}| \le (b_u c_1(1+\rho)l + b_u c_1(1+\rho)l)/\frac{\gamma\mu}{4b_u} = \frac{8b_u^2 c_1}{\gamma\mu}(1+\rho)l := c(1+\rho)l$$

with $c = c(\gamma,\mu) = (8b_u^2 c_1)/(\gamma\mu)$. ∎

We can now proof the main result of this section.

THEOREM 1. *There exists $L > 0$ such that for all $l : 0 < l < L$ and all starting points (x,y,z) and their corresponding blocking step $(\Delta x, \Delta y, \Delta z)$ obtained from* (3) *with $\mu^0 = \mu^+$ that satisfy*

$$x^T z/n = \mu, \quad x_i z_i \ge \gamma\mu, \quad x_i, z_i \le b_u, \quad \mu^+ < \frac{1}{2}\gamma\mu, \quad x + \Delta x \ge -le, \quad z + \Delta z \ge -le,$$

there is a step (d_x, d_z) that unblocks, i.e.,

$$x + d_x + \Delta x(x+d_x, z+d_z) \ge 0,$$
$$z + d_z + \Delta z(x+d_x, z+d_z) \ge 0.$$

Proof. Set $\epsilon = \frac{1}{10c}$. From the differentiability of $\Delta x(x,z), \Delta z(x,z)$ there exists a δ such that for all $(d_x, d_z) : \|(d_x, d_z)\|_\infty \le \delta$:

$$\begin{aligned} \left\| \Delta x(x+d_x, z+d_z) - \Delta x - \frac{\mathrm{d}\Delta x}{\mathrm{d}x} d_x - \frac{\mathrm{d}\Delta x}{\mathrm{d}z} d_z \right\| &\le \epsilon\|(d_x, d_z)\|, \\ \left\| \Delta z(x+d_x, z+d_z) - \Delta z - \frac{\mathrm{d}\Delta z}{\mathrm{d}x} d_x - \frac{\mathrm{d}\Delta z}{\mathrm{d}z} d_z \right\| &\le \epsilon\|(d_x, d_z)\|. \end{aligned} \tag{23}$$

Now set $\rho = \frac{1}{4}$ and $L = \min\{\frac{\hat{p}}{4}, \frac{4}{5}\frac{\delta}{c}\}$. Then from Lemma 2 there exists (d_x, d_z) : $\|(d_x, d_z)\|_\infty \le c\frac{5}{4}l \le \delta$ such that

$$x + d_x + \Delta x + \frac{\mathrm{d}\Delta x}{\mathrm{d}x}d_x + \frac{\mathrm{d}\Delta x}{\mathrm{d}z}d_z \ge \rho l e = \frac{1}{4}le,$$
$$z + d_z + \Delta z + \frac{\mathrm{d}\Delta z}{\mathrm{d}x}d_x + \frac{\mathrm{d}\Delta z}{\mathrm{d}z}d_z \ge \rho l e = \frac{1}{4}le,$$

and therefore

$$\begin{aligned} & x_i + d_{x,i} + \Delta x_i(x + d_x, z + d_z) \\ & \quad = x_i + d_{x,i} + \Delta x_i + \frac{\mathrm{d}\Delta x_i}{\mathrm{d}x}d_x + \frac{\mathrm{d}\Delta x_i}{\mathrm{d}z}d_z \\ & \qquad - \left(\Delta x_i + \frac{\mathrm{d}\Delta x_i}{\mathrm{d}x}d_x + \frac{\mathrm{d}\Delta x_i}{\mathrm{d}z}d_z - \Delta x_i(x + d_x, z + d_z)\right) \\ & \quad \ge \frac{1}{4}l - \epsilon\|(d_x, d_z)\| = \frac{1}{4}l - \frac{1}{10c}\|(d_x, d_z)\| \\ & \quad \ge \frac{1}{4}l - \frac{1}{10c}c\frac{5}{4}l \ge \frac{1}{8}l > 0, \end{aligned}$$

and the same for the z components. □

The insight gained from this theorem is that our proposed unblocking strategy is sound in principle: If the negative components of the prospective next iterate $(x+\Delta x, z+\Delta z)$ are bounded in size by L, then there exists an unblocking perturbation (d_x, d_z) of the current iterate. The size of this perturbation is $\mathcal{O}(L)$. Unfortunately the construction of (d_x, d_z) relies on the knowledge of the analytic center $\hat{p}$ of the problem (or at least any other strictly primal/dual feasible point). Therefore the construction used in the proof cannot be implemented in practice. In the following section we will derive an implementable heuristic.

4.3. Implementation. There is a principle difficulty with finding a solution to the *unblocking equations* (16). Theorem 1 guarantees that a solution (of bounded size) exists. The system (15):

$$x + d_x + \Delta x + \frac{\mathrm{d}\Delta x}{\mathrm{d}x}d_x + \frac{\mathrm{d}\Delta x}{\mathrm{d}z}d_z \ge \rho L,$$
$$z + d_z + \Delta z + \frac{\mathrm{d}\Delta z}{\mathrm{d}x}d_x + \frac{\mathrm{d}\Delta z}{\mathrm{d}z}d_z \ge \rho L$$

seems to imply that we could gather the complete sensitivity information ($\frac{\mathrm{d}\Delta x}{\mathrm{d}x}$, $\frac{\mathrm{d}\Delta x}{\mathrm{d}z}$, $\frac{\mathrm{d}\Delta z}{\mathrm{d}x}$, $\frac{\mathrm{d}\Delta z}{\mathrm{d}z}$), requiring n backsolves to do so, and find d_x, d_z to satisfy

$$\begin{bmatrix} \frac{\mathrm{d}\Delta x}{\mathrm{d}x} + I & \frac{\mathrm{d}\Delta x}{\mathrm{d}z} \\ \frac{\mathrm{d}\Delta z}{\mathrm{d}x} & \frac{\mathrm{d}\Delta z}{\mathrm{d}z} + I \end{bmatrix} \begin{bmatrix} d_x \\ d_z \end{bmatrix} \ge \begin{bmatrix} -x - \Delta x + \rho L \\ -z - \Delta z + \rho L \end{bmatrix}. \tag{24}$$

However, the system matrix in (24) is singular (actually of rank n) as can be seen from (17); hence it is unclear if a solution (d_x, d_z) exists at all.

In the results of Theorem 1 we get around this difficulty by assuming the knowledge of the analytic center, something that does not hold in practice. The only

solution we can suggest is to use the sensitivity information in a heuristic targeted at unblocking the search direction.

The idea is based on the observation that typically only a few components of the Newton step $(\Delta x, \Delta z)$ are blocking seriously and that these can be effectively influenced by changing the corresponding components of (x, z) only. One potential danger of aiming solely at unblocking the step direction is that we might have to accept a significant worsening of centrality or feasibility of the new iterate, which is clearly not in our interest. The proposed strategy attempts to avoid this as well by minimizing the perturbation (d_x, d_z) to the current point.

The heuristic that we are proposing is based on the assumption that a change in the ith component x_i, z_i will have a strong influence on the ith component of the step $\Delta x_i, \Delta z_i$, so changing only x_i, z_i components corresponding to blocking components of the step might be sufficient. Indeed our strategy will identify a (small) index set $\mathcal{I}$ of most blocking components, obtain the sensitivity information with respect to these components, and attempt to unblock each $(\Delta x_i, \Delta z_i)$ by changes to component i of (x, z) *only*. Since usually only Δx_i or Δz_i but not both are blocking, allowing perturbations in both x_i or z_i leaves one degree of freedom, which will be used to minimize the size of the required unblocking step.

The assumption made above can be justified as follows: according to (12), the sensitivity $\mathrm{d}(\Delta x, \Delta z)/\mathrm{d}x_i$ (and similarly $\mathrm{d}/\mathrm{d}z_i$) is made up of two components: the ith unit vector e_i and the solution to (13), which according to (14) is the weighted projection of the ith unit vector onto the null space of A.

Our implemented unblocking strategy is thus as follows:

Algorithm: Unblocking Strategy

---

1) Choose the size of the unblocking set $|\mathcal{I}|$, a target unblocking level $t > 1$, and bounds $0 < \underline{\gamma} < 1 < \bar{\gamma}$ on the acceptable change to a component.

2) find the set $\mathcal{I}$ of most blocking components (in x or z)

   `for all` i in 10% most blocking components `do`

3) find sensitivity of $(\Delta x, \Delta z)$ with respect to (x_i, z_i)

4) find the change $(d_{x,i}, d_{z,i})$ needed in x_i or z_i to unblock component i

5) change either x_i or z_i depending on where the change would be *more effective*.

   `next` i

6) update $x = x + d_x$ and $z = z + d_z$ and recompute the affine scaling direction

---

Steps 4) and 5) of the above algorithm need further clarification: For each blocking component x_i (or z_i) we have $x_i + \alpha_x \Delta x_i < 0$ for small positive values of α_x, or $\Delta x_i / x_i \ll -1$. From the sensitivity analysis we know $\frac{\mathrm{d}\Delta x_i}{\mathrm{d}x_i}$, the rate of change of Δx_i when x_i changes. We are interested in the necessary change $d_{x,i}$ to x_i such that the search direction is unblocked, that is to say

$$\frac{\Delta x_i + \frac{\mathrm{d}\Delta x_i}{\mathrm{d}x_i} d_{x,i}}{x_i + d_{x,i}} \geq -t, \quad (t \approx 5);$$

in other words a step of $\alpha_p \geq 1/t$ $(1/t \approx 0.2)$ will be possible. From this requirement

we get the provisional change

$$\widetilde{d_{x,i}} = -\frac{tx_i + \Delta x_i}{t + \frac{\mathrm{d}\Delta x_i}{\mathrm{d}x_i}}.$$

We need to distinguish several cases:

(i) $\dfrac{\mathrm{d}\Delta x_i}{\mathrm{d}x_i} \le \dfrac{\Delta x_i}{x_i}$:

A step in positive direction would lead to even more blocking. A negative step will unblock. However, we are not prepared to let $x_i + d_{x,i}$ approach zero; hence we choose

$$\overline{d_{x,i}} = \max\{\widetilde{d_{x,i}}, (\underline{\gamma} - 1)x_i\}.$$

(ii) $\dfrac{\mathrm{d}\Delta x_i}{\mathrm{d}x_i} > \dfrac{\Delta x_i}{x_i}$:

A positive step would weaken the blocking. However, if $\frac{\mathrm{d}\Delta x_i}{\mathrm{d}x_i} < -t$, the target unblocking level $-t$ can never be reached (and the provisional $\widetilde{d_{x,i}}$ is negative). In this case (and also if the provisional $\widetilde{d_{x,i}}$ is very large) we choose the maximal step that we are prepared to take:

$$\overline{d_{x,i}} = \begin{cases} d_{max} & \text{if } \widetilde{d_{x,i}} < 0, \\ \min\{\widetilde{d_{x,i}}, d_{max}\} & \text{otherwise} \end{cases}$$

with $d_{max} = (\bar{\gamma} - 1)x_i$.

Alternatively we can unblock a blocking Δx_i by changing z_i. The required provisional change $\widetilde{d_{z,i}}$ can be obtained from

$$\frac{\Delta x_i + \frac{\mathrm{d}\Delta x_i}{\mathrm{d}z_i} d_{z,i}}{x_i} \ge -t$$

as

$$\widetilde{d_{z,i}} = -\frac{tx_i + \Delta x_i}{\frac{\mathrm{d}\Delta x_i}{\mathrm{d}z_i}}.$$

In this case $\widetilde{d_{z,i}}$ indicates the correct sign of the change, but for $\frac{\mathrm{d}\Delta x_i}{\mathrm{d}z_i}$ close to zero the provisional step might be very large. We apply the same safeguards as for the step in x to obtain

$$\overline{d_{z,i}} = \begin{cases} \max\{\widetilde{d_{z,i}}, (\underline{\gamma} - 1)z_i\} & \widetilde{d_{z,i}} < 0, \\ \min\{\widetilde{d_{z,i}}, d_{max}\} & \widetilde{d_{z,i}} \ge 0, \end{cases}$$

where $d_{max} = (\bar{\gamma} - 1)z_i$. Since our aim was to reduce the blocking level from $-\Delta x_i / x_i$ to t, we can evaluate the effectiveness of the suggested changes $\overline{d_{x,i}}, \overline{d_{z,i}}$ by

$$p_x = \frac{\text{(old blocking level)} - \text{(new blocking level)}}{\text{(old blocking level)} - \text{(target blocking level)}} = \frac{-\frac{\Delta x_i}{x_i} + \frac{\Delta x_i + \frac{\mathrm{d}\Delta x_i}{\mathrm{d}x_i}\overline{d_{x,i}}}{x_i + \overline{d_{x,i}}}}{-\frac{\Delta x_i}{x_i} + t}$$

and

$$p_z = \frac{-\frac{\Delta x_i}{x_i} + \frac{\Delta x_i + \frac{\mathrm{d}\Delta x_i}{\mathrm{d}z_i}\overline{d_{z,i}}}{x_i}}{-\frac{\Delta x_i}{x_i} + t}.$$

Given these quantities we use $p_x/|\overline{d_{x,i}}|, p_z/|\overline{d_{z,i}}|$ as measures of the *relative effectiveness* of changing the x_i, z_i component. Our strategy is to first change the component for which this ratio is larger, and, should the corresponding p_x, p_z be less than 1, add a proportional change in the other component, i.e., if $p_x/|\overline{d_{x,i}}| > p_z/|\overline{d_{z,i}}|$:

$$d_{x,i} = \overline{d_{x,i}},$$
$$d_{z,i} = \min\{(1-p_x)/p_z, 1\}\overline{d_{z,i}}.$$

An analogous derivation can be performed to unblock the z-component Δz_i of the search direction.

The analysis in the previous section was aimed at unblocking the primal-dual direction corresponding to a fixed target value μ^0. We are, however, interested in using this analysis in the context of a predictor-corrector method. This seems to complicate the situation, since the predictor-corrector direction is now the result of a two-step procedure. As pointed out earlier, however, while the primal-dual direction and subsequently the length of the step that can be taken along it does depend on the target μ^0 value, the sensitivity of this step does not depend on μ^0. This leads us to the following strategy: We obtain the sensitivity with respect to the most blocking components after the predictor step and use these to unblock the combined predictor-corrector (and higher order corrector steps) separately following the above heuristic.

5. Numerical results. In order to evaluate the relative merit of the suggested warmstarting schemes, we have run a selection of numerical tests. In the first instance we have used a warmstarting setup based on the NETLIB LP test set as described in [1, 10] to evaluate a selection of the described heuristics.

In a second set of tests we have used the best warmstart settings from the first set and used these to warmstart the NETLIB LP test set, a selection of QP problems from [12] as well as some large scale QP problems arising from the problem of finding the efficient frontier in portfolio optimization and solving a nonlinear capacitated Multi-Commodity Network Flow problem (MCNF).

All warmstarting strategies have been implemented in our interior point solver OOPS [7]. For all tests we save the first iterate in the original problem solution process for which the relative duality gap satisfies

$$\frac{(c^Tx + 0.5x^TQx) - (b^Ty - 0.5x^TQx)}{(c^Tx + 0.5x^TQx) + 1} = \frac{x^Tz}{(c^Tx + 0.5x^TQx) + 1} \leq 0.01$$

for use as a warmstarting point. We do not attempt to find an "optimal" value for $\bar{\mu}$: our motivation is primarily to evaluate unblocking techniques in order to recover from "bad" warmstarting situations; furthermore it is likely that the optimal $\bar{\mu}$ is highly problem (and perturbation) dependent. On the contrary, we assume that a 2-digit approximate optimal solution of the original problem should be a good starting point for the perturbed problem.

5.1. The NETLIB warmstarting test set. In order to compare our results more easily to other contributions, we use the NETLIB warmstarting testbed suggested by [1]. This uses the smaller problems from the NETLIB LP test set as the original problems and considers changes to the right-hand side b, the objective vector c, the system matrix A, and different perturbation sizes δ. The perturbed problem instances are randomly generated as follows:

For perturbations to b and c we first generate a uniform-[0,1] distributed random number for every vector component. Should this number be less than $\min\{0.1, 20/n\}$

TABLE 1
Higher order correctors as unblocking device.

	b			c			A			
	0.1	0.01	0.001	0.1	0.01	0.001	0.1	0.01	0.001	total
base	6.4	5.6	6.1	14.5	8.5	6.4	10.2	7.0	7.2	8.1
hoc	6.0	5.4	5.6	11.3	7.6	6.3	8.6	6.5	6.8	7.2

(n being the dimension of the vector), this component is marked for modification. That is, we modify on average 10% (but at most 20) of the components. For all marked components we will generate a second uniform-$[-1, 1]$ distributed random number r. The new component $\tilde{b_i}$ is generated from the old one b_i as

$$\tilde{b_i} = \begin{cases} \delta r & |b_i| \leq 10^{-6}, \\ (1 + \delta r) b_i & \text{otherwise.} \end{cases}$$

For perturbations to A we proceed in the same manner, perturbing the vector of nonzero elements in A as before. For the results presented in this paper we have solved each problem for each warmstart strategy for 10 random perturbations of each type (b, c, and A). We will use these to evaluate the merit of each of the considered modifications and unblocking heuristics. A list of the considered NETLIB problems can be obtained from Tables 6–9.

In the numerical test performed we were guided by two objectives: first to evaluate if and how the various warmstarting strategies presented in section 3 can be combined, and second to evaluate the merit of the proposed unblocking strategy. In order to save on the total amount of computation, we will use the following strategy: Every warmstarting heuristic is tested against a *base* warmstarting code and against the best combination found so far. If a heuristic is found to be advantageous, it will be added to the *best* benchmark strategy for the future tests.

5.1.1. Higher order correctors. We investigate the use of higher-order correctors as an unblocking device. The interior point code OOPS applied for these calculations uses higher-order correctors by default if the Mehrotra corrector step (5) has been successful (i.e., it leads to larger stepsizes α_P, α_D than the predictor step). When using higher order correctors as an unblocking device, we will attempt them even if the Mehrotra corrector has been rejected. Table 1 gives results with and without forcing higher order correctors (`hoc` and `base`, respectively). The numbers reported are the average number of iterations of the warmstarted problem over all problems in the test set and all 10 random perturbations. Problem instances which are infeasible or unbounded after the perturbation have been discarded. Clearly the use of higher order correctors is advantageous. We therefore recommend the use of higher order correctors in all circumstances in the context of warmstarting. All following tests are performed with the use of higher order correctors.

5.1.2. Centering steps. We explore the benefit of using centering steps as a technique to facilitate warmstarting. These are performed either at the end of the solution process for the original problem before the advanced center is returned (`end`) or at the beginning of the modified problem solution, before any reduction of the barrier μ is applied (`beg`). As pointed out earlier the latter corresponds to the Newton corrector step of [15]. We have tested several settings of `end` and `beg` corresponding to the number of steps of this type being taken. The additional centering iterations are included in the numbers reported. Results are summarized in Table 2.

TABLE 2
Additional centering iterations.

	b			c			A			
	0.1	0.01	0.001	0.1	0.01	0.001	0.1	0.01	0.001	total
base										
beg=0, end=0	6.4	5.6	6.1	14.5	8.5	6.4	10.2	7.0	7.2	8.1
beg=0, end=1	6.3	5.3	5.2	15.4	8.5	6.3	11.6	6.9	6.9	8.2
beg=1, end=0	6.1	5.4	5.9	13.9	7.9	6.3	9.7	6.7	7.1	7.8
beg=1, end=1	6.1	5.0	5.2	14.7	8.4	6.2	10.8	7.0	6.9	8.0
beg=1, end=2	6.1	5.0	5.0	14.9	8.7	6.2	11.5	7.0	6.6	8.0
best										
beg=0, end=0	6.0	5.4	5.6	11.3	7.6	6.3	8.6	6.5	6.8	7.2
beg=1, end=0	6.0	5.3	5.5	10.9	7.4	6.1	8.4	6.6	7.0	7.1
beg=0, end=1	6.0	4.9	5.1	11.9	7.6	5.9	9.2	6.4	6.5	7.2
beg=1, end=1	5.7	5.0	5.1	11.8	7.4	5.9	9.2	6.6	6.5	7.2
beg=1, end=2	5.7	4.7	5.2	11.6	7.1	5.8	9.4	6.4	6.5	7.0

TABLE 3
z-adjustment as modification step.

		b			c			A			
		0.1	0.01	0.001	0.1	0.01	0.001	0.1	0.01	0.001	total
base	no-adj	6.4	5.6	6.1	14.5	8.5	6.4	10.2	7.0	7.2	8.1
	z-adj	6.3	5.5	5.8	12.5	7.7	6.3	9.2	7.1	7.1	7.6
	WLS-0.01	6.3	5.5	6.1	14.0	8.3	6.4	9.9	7.0	7.1	8.0
	WLS-0.1	7.0	6.6	6.9	12.7	9.1	7.4	8.1	7.1	8.5	8.1
best	no-adj	5.7	4.7	5.2	11.6	7.1	5.8	9.4	6.4	6.5	7.0
	z-adj	5.7	4.8	5.1	10.5	6.8	5.7	8.8	6.3	6.4	6.8
	WLS-0.01	5.7	4.8	5.2	11.6	7.0	5.9	9.3	6.4	6.5	7.0
	WLS-0.1	6.3	5.9	5.9	10.0	7.9	6.8	7.4	6.7	7.7	7.2

Compared with the base, strategy $(1, 0)$ is the best, whereas compared to the best (which just includes higher-order correctors at this point), strategy $(1, 2)$ is preferable. Due to the theoretical benefits of working with a well-centered point, we will use centering strategy $(1, 2)$ in the *best* benchmark strategy for the following tests.

5.1.3. z-adjustment/WLS-step. We have evaluated the benefit of attempting to absorb dual infeasibilities into the z value of the warmstart vector, together with the related WLS heuristic (which attempts to find a least squares correction to the saved iterate such that the resulting point is primal/dual feasible). The results are summarized in Table 3. Surprisingly there is a clear advantage of the simple z-adjustment heuristic, whereas the (computationally more expensive and more sophisticated) WLS step (`WLS-0.01`) hardly improves on the base strategy. Our only explanation for this behavior is that for our fairly low saved μ-value (2-digit approximate optimal solution to the original problem) the full WLS direction is usually infeasible, so only a fractional step in it can be taken. The z-adjustment, on the other hand, has a more sophisticated fallback strategy which considers adjustment for each component separately, so it is not quite as easily affected by blocking in the modification direction. Reference [15] suggests employing the WLS step together with a backtracking strategy, which saves several iterates from the original problem for different μ and chooses one for which the WLS step does not block. We have emulated this by trying the WLS step for a larger μ (`WLS-0.1`). Any gain of a larger portion of the WLS step being taken, however, is offset by the starting point now being further away from optimality, resulting in an increase of the number of iterations. We have added the z-adjustment heuristic to our *best* benchmark strategy.

TABLE 4
Splitting directions.

		b			c			A			
		0.1	0.01	0.001	0.1	0.01	0.001	0.1	0.01	0.001	total
base	it=0	6.4	5.6	6.1	14.5	8.5	6.4	10.2	7.0	7.2	8.1
	it=1	6.3	5.5	6.1	14.4	8.6	6.5	10.1	6.9	7.2	8.1
	it=2	6.3	5.5	6.1	14.3	8.6	6.5	10.1	6.9	7.2	8.1
best	it=0	5.7	4.8	5.1	10.5	6.8	5.7	8.8	6.3	6.4	6.8
	it=1	5.7	4.8	5.1	10.5	6.8	5.8	8.7	6.3	6.4	6.8
	it=2	5.8	4.8	5.1	10.4	6.7	5.7	8.7	6.4	6.4	6.8

TABLE 5
Sensitivity based unblocking heuristic.

	b			c			A			
	0.1	0.01	0.001	0.1	0.01	0.001	0.1	0.01	0.001	total
base										
unblk=0	6.4	5.6	6.1	14.5	8.5	6.4	10.2	7.0	7.2	8.1
unblk=1	6.1	5.5	6.0	13.2	8.2	6.4	9.7	7.0	7.1	7.8
unblk=2	6.1	5.3	5.9	12.1	8.1	6.1	9.2	6.8	6.9	7.5
unblk=3	6.0	5.6	6.1	11.4	8.0	6.2	9.0	7.4	7.1	7.5
best: hoc, beg=1, end=2, z-adj										
unblk=0	5.7	4.8	5.1	10.5	6.8	5.7	8.8	6.3	6.4	6.8
unblk=1	5.6	4.8	5.1	9.8	6.5	5.7	8.3	6.4	6.4	6.6
unblk=2	5.7	5.1	5.5	9.4	6.8	5.9	8.2	6.4	6.1	6.7
unblk=3	5.6	5.1	5.7	9.5	6.8	5.8	8.2	6.2	6.5	6.7
beg=0, end=0, z-adj										
unblk=0	6.1	5.0	5.0	14.9	8.7	6.2	11.5	7.0	6.6	8.0
unblk=1	5.9	4.9	5.0	13.2	7.9	6.0	10.4	6.8	6.7	7.6
unblk=2	5.8	5.0	5.0	11.9	8.0	6.1	9.7	6.7	6.9	7.4
unblk=3	5.8	5.2	5.1	11.5	7.7	5.8	9.7	6.8	6.8	7.3
hoc, beg=0, end=0, z-adj										
unblk=0	5.7	4.7	5.2	11.6	7.1	5.8	9.4	6.4	6.5	7.0
unblk=1	5.5	4.8	5.3	10.7	6.9	5.6	9.1	6.5	6.4	6.8
unblk=2	5.8	4.9	5.1	9.8	7.4	5.7	8.7	6.7	5.4	6.7
unblk=3	5.6	5.0	5.5	9.4	6.7	5.7	9.0	6.3	5.5	6.6

5.1.4. Splitting directions. This analyzes the effectiveness of using the computations of separate primal, dual, and centrality correcting directions as in [6] as an unblocking strategy. The results given in Table 4 correspond to different numbers of initial iterations in the solution process of the modified problem using this technique.

As can be seen there is no demonstrable benefit from using this unblocking technique, we have therefore left it out of all subsequent tests.

5.1.5. Unblocking by sensitivity. Finally we have tested the effectiveness of our unblocking scheme based on using sensitivity information. We are considering employing this heuristic for up to the first three iterations. The parameters we have used are $|\mathcal{I}| \leq 0.1n$ (i.e., the worst 10% of components are unblocked), $t = 5$, $\bar{\gamma} = 10$, and $\underline{\gamma} = 0.1$. Results are summarized in Table 5. Unlike the other tests, we have not only tested the unblocking strategy against the base and the best but also against two further setups to evaluate the effectiveness of the strategy to recover from blocking in different environments.

As can be seen there is a clear benefit in employing this heuristic in all tests. The results are less pronounced when comparing against the *best* strategy, but even here there is a clear advantage of performing one iteration of the unblocking strategy.

TABLE 6
Results (best warmstart)—perturbations in b.

Problem	0.1			0.01			0.001		
	cold	warm	red	cold	warm	red	cold	warm	red
ADLITTLE	10.0	6.0	40.0	10.0	5.0	50.0	11.4	6.0	47.3
AFIRO	10.1	4.2	58.4	10.1	4.3	57.4	10.1	4.3	57.4
AGG2	16.1	4.6	71.4	16.2	4.0	75.3	16.1	4.0	75.1
AGG3	15.7	5.6	64.3	15.5	5.0	67.7	16.0	5.0	68.7
BANDM	13.8	8.2	40.5	14.0	4.1	70.7	13.5	4.0	70.3
BEACONFD	-	-	-	-	-	-	-	-	-
BLEND	9.0	4.0	55.5	9.0	4.3	52.2	9.0	4.2	53.3
BOEING1	19.3	7.2	62.6	21.5	8.3	61.3	19.1	5.1	73.2
BORE3D	-	-	-	-	-	-	-	-	-
BRANDY	-	-	-	-	-	-	-	-	-
DEGEN2	-	-	-	-	-	-	-	-	-
E226	16.0	12.8	20.0	15.8	5.0	68.3	15.0	4.8	68.0
GROW15	13.0	4.0	69.2	13.0	4.0	69.2	13.0	4.0	69.2
GROW7	12.0	4.0	66.6	12.0	4.0	66.6	12.0	4.0	66.6
ISRAEL	21.0	6.9	67.1	20.5	4.0	80.4	19.9	4.0	79.8
KB2	17.7	5.0	71.7	17.4	5.0	71.2	17.2	5.0	70.9
LOTFI	19.3	6.8	64.7	20.0	5.7	71.5	20.0	5.8	71.0
RECIPELP	14.0	7.0	50.0	14.0	7.0	50.0	14.5	10.8	25.5
SC105	12.0	5.0	58.3	12.0	5.1	57.5	12.0	5.0	58.3
SC205	12.0	5.2	56.6	12.0	5.0	58.3	12.0	5.0	58.3
SC50A	11.0	4.0	63.6	11.0	4.0	63.6	11.0	4.0	63.6
SC50B	10.0	4.2	58.0	10.0	4.0	60.0	12.1	14.2	−17.3
SCAGR25	12.0	4.8	60.0	11.9	4.1	65.5	12.7	4.0	68.5
SCAGR7	10.1	4.1	59.4	9.9	4.0	59.5	9.8	4.0	59.1
SCFXM1	14.6	5.0	65.7	15.2	5.8	61.8	14.1	4.1	70.9
SCSD1	9.9	9.5	4.0	10.3	5.9	42.7	10.2	5.1	50.0
SCTAP1	14.7	6.0	59.1	14.9	5.0	66.4	15.6	5.3	66.0
SHARE1B	21.5	5.8	73.0	20.8	5.4	74.0	21.3	5.0	76.5
SHARE2B	9.3	5.2	44.0	9.2	5.1	44.5	9.1	5.1	43.9
STOCFOR1	13.5	5.4	60.0	13.0	5.1	60.7	15.4	5.3	65.5
Average	13.8	5.8	56.3	13.8	4.9	62.6	13.9	5.3	60.0

5.2. Results for best warmstart strategy. After these tests we have combined the best setting for all of the considered warmstart heuristics and give more detailed results on the NETLIB test set as well as for a selection of large scale quadratic programming problems.

Tables 6–9 compare the best combined warmstarting strategy for all test problems with a cold start. We give in each case the average number of iterations over 10 random perturbations. Column `red` gives the average percentage iteration reduction achieved by employing the warmstart. An entry "`-`" denotes that all corresponding perturbations of the problem were either infeasible or unbounded. As can be seen we are able to save between 50% and 60% of iterations on all considered problems.

5.3. Comparison with LOQO results. To judge the competitiveness of our best combined warmstarting strategy, we have compared the results on the NETLIB test set with those reported by [1] which use a different warmstarting methodology. Figure 1 gives a summary of this comparison. The four lines on the left graph give the number of iterations needed for each of the 30 NETLIB problems reported in Tables 6–9 averaged over all perturbations for OOPS and LOQO [1], using a warmstart and a coldstart. As can be seen the default version of OOPS (solid line) needs fewer iterations than LOQO (dotted line). The warmstarted versions of each code

TABLE 7
Results (best warmstart)—perturbations in c.

Problem	0.1			0.01			0.001		
	cold	warm	red	cold	warm	red	cold	warm	red
ADLITTLE	10.3	7.3	29.1	10.1	5.2	48.5	10.4	5.0	51.9
AFIRO	10.3	5.3	48.5	10.3	4.8	53.3	10.7	4.8	55.1
AGG2	16.7	6.6	60.4	16.4	4.8	70.7	16.0	4.1	74.3
AGG3	16.0	6.9	56.8	16.0	5.3	66.8	15.9	4.9	69.1
BANDM	13.7	14.2	−3.6	13.9	5.2	62.5	13.6	4.0	70.5
BEACONFD	10.1	4.7	53.4	10.0	4.0	60.0	11.0	4.8	56.3
BLEND	9.4	7.3	22.3	9.0	4.6	48.8	9.0	4.3	52.2
BOEING1	19.6	24.2	−23.4	19.6	8.6	56.1	19.1	5.8	69.6
BORE3D	12.9	6.1	52.7	13.2	4.4	66.6	13.2	4.2	68.1
BRANDY	15.2	8.7	42.7	15.5	4.3	72.2	15.3	4.0	73.8
DEGEN2	9.8	4.5	54.0	10.0	4.8	52.0	10.0	5.0	50.0
E226	15.6	15.0	3.8	15.2	9.0	40.7	15.1	4.5	70.1
GROW15	22.9	13.7	40.1	22.9	9.2	59.8	17.7	11.0	37.8
GROW7	18.9	14.3	24.3	19.9	12.4	37.6	23.6	17.5	25.8
ISRAEL	20.4	7.7	62.2	21.0	4.2	80.0	21.1	4.3	79.6
KB2	17.8	6.8	61.7	17.9	5.0	72.0	18.0	5.0	72.2
LOTFI	19.0	30.7	−61.5	23.0	20.9	9.1	22.4	12.7	43.3
RECIPELP	-	-	-	-	-	-	-	-	-
SC105	11.4	15.4	−35.0	11.8	5.9	50.0	11.5	5.0	56.5
SC205	12.7	20.9	−64.5	13.1	18.2	−38.9	12.1	6.7	44.6
SC50A	11.2	6.8	39.2	11.0	4.1	62.7	11.0	4.0	63.6
SC50B	10.3	7.2	30.0	10.0	4.4	56.0	10.0	4.0	60.0
SCAGR25	12.0	4.7	60.8	12.4	4.4	64.5	13.0	4.0	69.2
SCAGR7	10.1	4.8	52.4	9.9	4.1	58.5	10.0	4.0	60.0
SCFXM1	14.4	7.4	48.6	14.0	4.0	71.4	14.0	4.0	71.4
SCSD1	9.5	5.2	45.2	9.2	5.0	45.6	9.0	5.0	44.4
SCTAP1	16.2	6.6	59.2	16.1	5.8	63.9	15.8	6.0	62.0
SHARE1B	22.6	8.9	60.6	21.9	6.0	72.6	20.9	5.5	73.6
SHARE2B	9.2	7.2	21.7	9.0	5.0	44.4	9.1	5.0	45.0
STOCFOR1	12.8	5.0	60.9	13.0	5.0	61.5	14.4	5.0	65.2
Average	14.2	9.8	31.1	14.3	6.5	54.1	14.2	5.7	59.8

(solid and dotted lines with markers, respectively) need significantly fewer iterations on average than their coldstarted siblings, with warmstarted OOPS being the most effective strategy over all. This plot indicates only the best combination of interior point code and warmstarting strategy without giving any insight into the relative effectiveness of the warmstarting approaches themselves. In order to measure the efficiency of the warmstart approaches, the second plot in Figure 1 compares the number of iterations *saved* by each warmstarting strategy as compared with its respective coldstarted variant. As can be seen our suggested warmstart implemented in OOPS is able to save around 50-60% of iterations, outperforming the LOQO warmstart which averages around 30% saved iterations.

5.4. Medium scale QP problems. We realize that the NETLIB testbed proposed in [1] includes only small LP problems. While this makes it ideal for the extensive testing that we have reported in the previous section, there is some doubt over whether the achieved warmstarting performance can be maintained for quadratic and (more realistic) large scale problems. In order to counter such criticism we have conducted warmstarting tests on two selections of small to medium scale QP problems as well as two sources of large scale quadratic programming. For the small and medium scale tests we have used the quadratic programming collection of Maros

TABLE 8
Results (best warmstart)—perturbations in A.

Problem	0.1			0.01			0.001		
	cold	warm	red	cold	warm	red	cold	warm	red
ADLITTLE	10.8	9.4	12.9	10.5	5.0	52.3	10.4	5.0	51.9
AFIRO	10.1	5.0	50.4	10.0	4.1	59.0	10.0	4.0	60.0
AGG2	15.9	5.3	66.6	16.0	4.2	73.7	16.2	4.0	75.3
AGG3	15.2	6.3	58.5	15.7	5.2	66.8	16.1	5.0	68.9
BANDM	13.8	7.9	42.7	13.8	4.4	68.1	13.4	4.1	69.4
BEACONFD	10.1	4.8	52.4	10.0	4.0	60.0	10.0	4.0	60.0
BLEND	9.0	9.5	−5.5	9.2	5.3	42.3	9.0	4.4	51.1
BOEING1	19.3	5.2	73.0	19.6	5.0	74.4	19.8	5.0	74.7
BORE3D	15.0	4.0	73.3	13.9	4.0	71.2	13.6	4.0	70.5
BRANDY	14.2	14.1	0.7	17.8	15.4	13.4	28.1	18.8	33.0
DEGEN2	11.1	13.4	−20.7	29.2	30.5	−4.4	93.0	86.0	7.5
E226	15.5	10.2	34.1	15.1	4.9	67.5	15.0	4.1	72.6
GROW15	20.2	12.9	36.1	15.3	11.3	26.1	13.4	5.0	62.6
GROW7	24.0	16.1	32.9	17.1	8.8	48.5	13.5	6.4	52.5
ISRAEL	19.8	5.4	72.7	20.0	4.0	80.0	19.9	4.0	79.8
KB2	18.2	15.3	15.9	18.2	5.1	71.9	17.8	5.0	71.9
LOTFI	20.0	7.1	64.5	25.8	12.3	52.3	50.1	36.2	27.7
RECIPELP	13.9	7.1	48.9	13.9	6.6	52.5	14.0	6.0	57.1
SC105	11.8	7.1	39.8	11.5	5.0	56.5	12.0	5.0	58.3
SC205	12.6	7.7	38.8	12.0	5.0	58.3	12.0	5.0	58.3
SC50A	11.1	7.1	36.0	11.0	4.0	63.6	11.0	4.0	63.6
SC50B	10.0	5.1	49.0	10.0	4.0	60.0	10.0	4.0	60.0
SCAGR25	11.7	9.4	19.6	11.8	4.3	63.5	12.5	4.3	65.6
SCAGR7	10.1	6.5	35.6	10.0	4.0	60.0	9.7	4.0	58.7
SCFXM1	15.2	8.0	47.3	14.9	4.6	69.1	14.4	5.0	65.2
SCSD1	9.1	6.3	30.7	9.3	5.2	44.0	9.2	4.8	47.8
SCTAP1	14.2	9.5	33.0	15.6	6.2	60.2	15.1	5.2	65.5
SHARE1B	21.0	9.4	55.2	21.2	7.0	66.9	22.1	5.6	74.6
SHARE2B	9.6	9.9	−3.1	9.2	5.7	38.0	9.0	5.0	44.4
STOCFOR1	11.5	5.8	49.5	12.3	5.2	57.7	12.1	5.1	57.8
Average	14.1	8.4	38.0	14.7	6.7	55.8	17.7	8.9	58.9

and Meszaros [12]. This includes QP problems from the CUTE test set as well as quadratic modifications of the NETLIB LP test set used in the previous comparisons. We have excluded problems that either have free variables (since OOPS currently has no facility to deal with free variables effectively), or where random perturbations of the problem data yield the problem primal or dual infeasible. The same methodology in perturbing the problems as for the NETLIB LP test set has been used, apart that perturbations in the objective function will now perturb random elements of c and Q. The results are displayed in Table 10. As for the LP case we list for each problem and perturbation the average number of iterations needed by OOPS when coldstarted and when warmstarted with the best strategy found in section 5.1 over the 10 random runs and 3 perturbation sizes. We also state the percentage of iterations saved by the warmstart. A blank entry indicates that all 30 random perturbations lead to primal or dual infeasible problems. The results demonstrate a similar performance of our best combined warmstarting strategy as obtained earlier for the LP problems.

5.5. Large scale QP problems. Finally we have evaluated our warmstart strategy in the context of two sources of large scale quadratic problems. In the first

TABLE 9
Results (best warmstart)—all perturbations.

Problem	b			c			A		
	cold	warm	red	cold	warm	red	cold	warm	red
ADLITTLE	10.4	5.6	46.1	10.2	5.8	43.1	10.5	6.4	39.0
AFIRO	10.1	4.2	58.4	10.4	4.9	52.8	10.0	4.3	57.0
AGG2	16.1	4.2	73.9	16.3	5.1	68.7	16.0	4.5	71.8
AGG3	15.7	5.2	66.8	15.9	5.7	64.1	15.6	5.5	64.7
BANDM	13.7	5.4	60.5	13.7	7.8	43.0	13.6	5.4	60.2
BEACONFD	-	-	-	10.3	4.5	56.3	10.0	4.2	58.0
BLEND	9.0	4.1	54.4	9.1	5.4	40.6	9.0	6.4	28.8
BOEING1	19.9	6.8	65.8	19.4	12.8	34.0	19.5	5.0	74.3
BORE3D	-	-	-	13.1	4.9	62.5	14.1	4.0	71.6
BRANDY	-	-	-	15.3	5.6	63.3	20.0	16.1	19.5
DEGEN2	-	-	-	9.9	4.7	52.5	44.4	43.3	2.4
E226	15.6	7.5	51.9	15.3	9.5	37.9	15.2	6.4	57.8
GROW15	13.0	4.0	69.2	21.1	11.3	46.4	16.3	9.7	40.4
GROW7	12.0	4.0	66.6	20.8	14.7	29.3	18.2	10.4	42.8
ISRAEL	20.4	4.9	75.9	20.8	5.4	74.0	19.9	4.4	77.8
KB2	17.4	5.0	71.2	17.9	5.6	68.7	18.0	8.4	53.3
LOTFI	19.7	6.1	69.0	21.4	21.4	0.0	31.9	18.5	42.0
RECIPELP	14.1	8.2	41.8	-	-	-	13.9	6.5	53.2
SC105	12.0	5.0	58.3	11.5	8.7	24.3	11.7	5.7	51.2
SC205	12.0	5.0	58.3	12.6	15.2	-20.6	12.2	5.9	51.6
SC50A	11.0	4.0	63.6	11.0	4.9	55.4	11.0	5.0	54.5
SC50B	10.7	7.4	30.8	10.1	5.2	48.5	10.0	4.3	57.0
SCAGR25	12.2	4.3	64.7	12.4	4.3	65.3	12.0	6.0	50.0
SCAGR7	9.9	4.0	59.5	10.0	4.3	57.0	9.9	4.8	51.5
SCFXM1	14.6	4.9	66.4	14.1	5.1	63.8	14.8	5.8	60.8
SCSD1	10.1	6.8	32.6	9.2	5.0	45.6	9.2	5.4	41.3
SCTAP1	15.0	5.4	64.0	16.0	6.1	61.8	14.9	6.9	53.6
SHARE1B	21.2	5.4	74.5	21.8	6.8	68.8	21.4	7.3	65.8
SHARE2B	9.2	5.1	44.5	9.1	5.7	37.3	9.2	6.8	26.0
STOCFOR1	13.9	5.2	62.5	13.4	5.0	62.6	11.9	5.3	55.4
Average	13.8	5.3	59.6	14.2	7.3	48.4	15.5	8.0	50.9

instance we have solved the capacitated MCNF problem

$$
\begin{aligned}
\min \quad & \sum_{(i,j)\in\mathcal{E}} \frac{x_{ij}}{K_{ij}-x_{ij}}, \\
\text{s.t.} \quad & \sum_{k\in\mathcal{D}} x_{ij}^{(k)} \le K_{ij}, \qquad \forall (i,j)\in\mathcal{E}, \\
& Nx^{(k)} = d^{(k)}, \qquad \forall k\in\mathcal{D}, \\
& x^{(k)} \ge 0, \qquad \forall k\in\mathcal{D},
\end{aligned} \tag{25}
$$

where N is the node-arc incidence matrix of the network, $d^{(k)}, k\in\mathcal{D}$ are the demand points, K_{ij} is the capacity of each arc (i,j), and x_{ij} is the flow along the arc. This is a nonlinear problem formulation. We have solved it by SQP using the interior point code OOPS as the QP solver and employing our best combined warmstart strategy between QP solutions. We have tested this on nine different MCNF models using from 4–300 nodes, up to 600 arcs, and up to 7021 commodities. The largest problem in the selection has $353,400$ variables. All solutions have required more than 10 SQP iterations. Table 11 gives the average number of IPM iterations for each SQP iteration both for cold- and warmstarting the IPM.

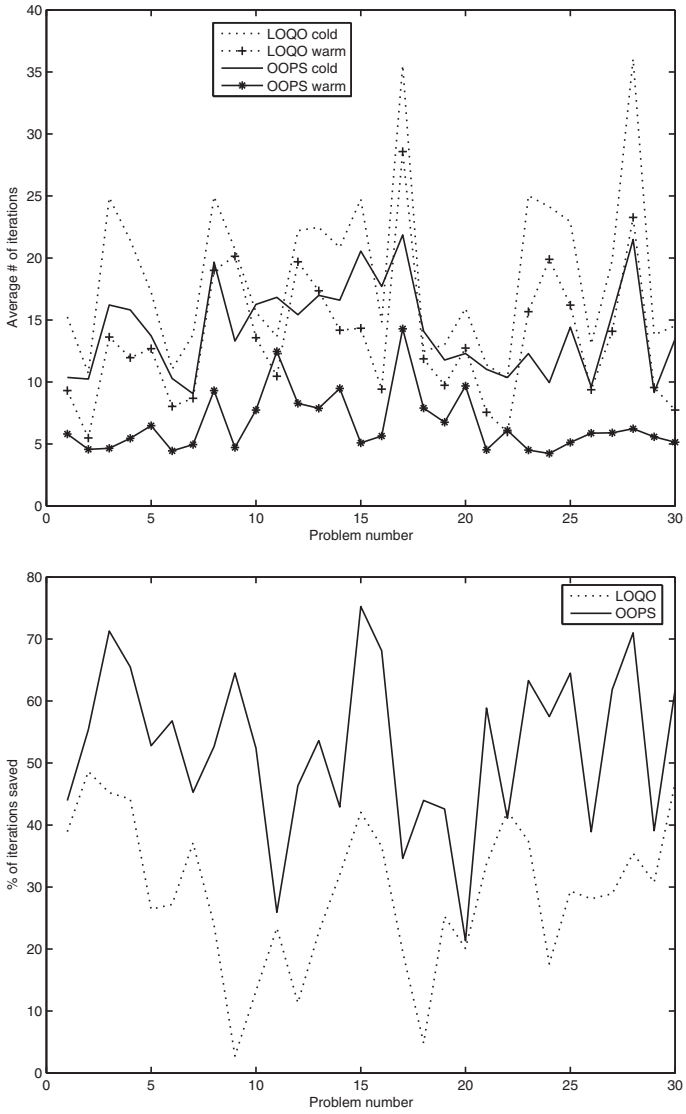


FIG. 1. *Results of LOQO and OOPS on warmstarting NETLIB problems.*

As before we achieve between 50 and 60% reduction in the number of interior point iterations.

Our last test example consists of calculating the complete efficient frontier in a Markowitz Portfolio Selection problem (see [14]). A Portfolio Selection problem aims to find the optimal investment strategy in a selection of assets over time. If the value of the portfolio at the end of the time horizon is denoted by the random variable X, the Markowitz formulation of the portfolio selection problem requires one to maximize the final expected wealth $\mathbb{E}(X)$ and minimize the associated risk, measured as the variance $\mathrm{Var}(X)$ which are combined into a single objective:

$$\min -\mathbb{E}(X) + \rho \mathrm{Var}(X) \tag{26}$$

TABLE 10
Results QP (best warmstart)—all perturbations.

Problem	b			c and Q			A		
	cold	warm	red	cold	warm	red	cold	warm	red
AUG2DCQP	10.0	5.2	48.1	10.0	5.1	49.1	10.9	6.8	37.7
AUG2DQP	10.0	4.1	59.0	10.0	4.1	59.1	10.7	7.0	34.6
AUG3DCQP	8.0	4.0	50.0	8.0	3.8	53.0	8.0	3.7	53.9
AUG3DQP	11.0	5.1	53.4	10.2	5.2	48.9	10.9	4.8	56.3
CVXQP1_S	11.2	5.3	52.9	11.0	5.1	53.8	11.2	5.4	51.7
CVXQP2_M	15.1	4.8	68.1	15.0	5.3	64.5	15.1	5.0	67.1
CVXQP2_S	11.9	6.1	49.2	12.0	6.1	49.0	12.0	6.1	49.4
CVXQP3_M	-	-	-	-	-	-	-	-	-
CVXQP3_S	10.5	6.6	37.1	10.0	5.0	49.9	10.2	6.4	37.2
DUAL1	10.0	4.9	51.2	10.0	5.1	48.6	9.9	5.2	47.6
DUAL2	10.0	5.3	47.1	10.0	4.9	51.1	9.7	4.8	50.7
DUAL3	11.0	6.1	44.1	10.8	5.6	48.0	10.7	5.7	46.5
DUAL4	9.0	4.8	46.9	9.0	5.1	43.1	9.0	5.2	42.2
DUALC1	21.9	3.8	82.6	22.0	4.2	81.1	22.6	4.0	82.4
DUALC2	22.0	3.8	82.6	21.9	3.9	82.0	21.5	3.9	82.0
DUALC5	12.0	3.9	67.9	12.0	4.1	65.4	12.2	3.8	68.4
DUALC8	14.6	3.9	73.2	15.0	4.2	72.3	15.2	3.9	74.5
GOULDQP2	6.0	4.9	18.3	8.0	7.8	2.5	6.0	5.1	14.3
GOULDQP3	9.0	5.0	44.4	9.0	5.1	43.3	9.0	4.9	45.3
HS118	9.0	3.7	59.1	9.0	3.6	59.8	9.0	4.2	53.1
HS21	17.0	6.9	59.5	16.8	6.7	60.3	17.0	7.0	58.7
HS35MOD	9.9	5.8	41.2	9.8	6.0	39.0	9.8	5.9	40.3
HS35	7.0	4.4	37.8	7.1	4.1	42.3	7.0	3.9	44.5
HS53	6.0	5.3	11.9	6.0	5.2	13.7	6.1	4.9	18.6
HS76	7.0	4.1	40.8	7.0	4.1	41.2	7.0	4.0	43.6
HUES-MOD	14.8	5.2	64.6	15.0	4.9	67.5	17.4	12.1	30.4
LOTSCHD	7.9	5.6	29.4	8.0	5.6	29.9	7.6	5.5	28.0
MOSARQP1	7.0	3.9	44.5	7.0	4.1	41.2	7.0	4.4	37.8
MOSARQP2	8.0	4.1	49.2	8.1	4.5	44.2	8.0	3.8	52.0
QPCBOEI1	35.3	23.0	34.8	34.0	22.9	32.4	35.2	23.1	34.3
QPCBOEI2	-	-	-	20.8	7.2	65.4	28.1	12.3	56.0
STCQP1	-	-	-	15.0	5.7	61.9	-	-	-
STCQP2	-	-	-	15.0	7.1	52.7	15.0	6.9	54.1
TAME	6.0	2.2	63.9	6.0	1.9	67.6	6.0	1.9	68.7
ZECEVIC2	7.0	4.8	32.3	7.1	5.2	26.8	7.0	5.3	24.9
25FV47	38.0	7.5	80.3	38.3	7.3	80.8	38.0	8.2	78.4
ADLITTLE	10.6	5.6	47.4	10.2	5.9	42.6	10.2	5.8	43.4
AFIRO	15.1	4.7	69.2	15.0	6.9	54.0	15.0	4.2	72.0
BEACONFD	-	-	-	10.0	4.2	58.2	10.0	4.2	57.9
BORE3D	-	-	-	15.3	4.7	69.0	15.5	4.1	73.5
BRANDY	-	-	-	12.8	5.5	57.3	14.3	14.3	0.1
E226	14.6	7.3	50.11	14.0	7.9	43.7	13.8	5.0	63.5
ETAMACRO	-	-	-	31.9	10.8	66.0	39.5	18.9	52.2
FFFFF800	63.1	8.6	86.4	61.3	7.0	88.7	56.9	6.6	88.3
GROW15	14.0	5.4	61.6	18.4	13.6	26.1	19.6	10.9	44.5
GROW22	15.9	4.8	69.7	15.6	6.3	59.5	17.0	7.4	56.6
GROW7	15.0	4.9	67.3	21.7	11.5	47.1	17.5	6.5	62.6
ISRAEL	18.9	5.1	73.2	20.1	5.5	72.4	18.4	5.2	71.6
SC205	16.8	7.1	57.6	18.9	19.3	-2.5	17.2	6.9	59.6
SCAGR25J	11.0	4.0	64.1	11.0	3.8	65.3	11.3	4.8	57.2
SCAGR25	11.1	5.0	55.3	11.2	5.2	53.8	11.6	5.3	54.5
SCAGR7	12.4	5.1	59.3	11.8	5.1	56.7	12.0	5.0	58.3
SCFXM1	21.0	7.4	64.6	20.8	7.1	66.0	21.5	7.3	66.1
SCFXM2	23.9	10.7	55.1	23.6	10.5	55.4	24.1	10.8	55.2
SCFXM3	26.6	13.5	49.2	24.5	13.2	46.2	27.8	13.6	51.1
SCORPION	-	-	-	12.1	3.1	74.3	-	-	-

TABLE 10
Continued.

Problem	b			c and Q			A		
	cold	warm	red	cold	warm	red	cold	warm	red
SCRS8	19.6	8.6	56.1	19.6	5.4	72.5	20.0	8.1	59.4
SCSD1	10.8	6.8	36.4	10.2	5.3	48.2	10.2	5.3	47.6
SCSD6	10.9	7.1	34.6	11.4	4.8	58.3	11.1	7.0	37.3
SCSD8	9.0	4.0	55.2	11.3	6.0	46.7	11.4	11.5	-0.2
SCTAP1	13.9	5.5	60.3	15.7	6.5	58.7	15.0	7.5	49.6
SCTAP2	16.0	4.4	72.2	17.9	6.0	66.1	15.1	4.7	68.9
SCTAP3	16.9	5.2	69.5	17.9	6.6	63.1	16.5	5.6	66.2
SEBA	53.3	24.3	54.3	53.7	22.8	57.6	53.5	23.7	55.6
SHARE1B	20.1	6.2	69.2	20.6	6.5	68.6	19.2	6.6	65.8
SHARE2B	24.9	15.1	39.2	24.3	14.9	39.0	25.9	16.8	35.2
SHELL	20.0	7.5	62.3	20.1	6.9	65.8	20.5	9.6	53.4
SHIP04L	-	-	-	11.9	3.7	68.7	11.6	12.0	-3.7
SHIP04S	-	-	-	12.0	4.0	66.8	11.6	11.1	4.3
SHIP08L	-	-	-	11.0	5.0	54.1	11.1	13.3	-19.7
SHIP08S	-	-	-	11.0	4.1	62.4	11.1	9.3	16.4
SHIP12L	-	-	-	16.0	5.1	67.8	14.7	11.4	22.3
SHIP12S	-	-	-	14.4	6.1	57.7	14.2	15.0	-5.1
SIERRAJG	-	-	-	37.4	5.5	85.3	-	-	-
SIERRA	-	-	-	38.3	5.1	86.7	-	-	-
STANDATA	23.0	17.8	22.7	17.4	5.2	70.0	17.9	4.9	72.8
Average	11.5	5.3	53.9	11.8	5.6	52.5	11.9	5.9	50.4

TABLE 11
Capacitated MCNF solved by warmstarted IPM-SQP.

iter	1	2	3	4	5	6	7	8	9	10
cold	12.7	11.9	13.7	15.8	16.2	15.6	14.9	14.6	14.5	15.0
warm	12.7	7.0	6.0	5.8	6.4	7.0	7.0	6.7	6.2	6.0
red	0.0	41.2	56.2	63.3	60.5	55.1	53.0	54.1	57.2	60.0

which leads to a QP problem. We use the multistage stochastic programming version of this model (described in [7]). This formulation leads to very large problem sizes.

The parameter ρ in (26) is known as the *Risk Aversion Parameter* and captures the investor's attitude to risk. A low value of ρ will lead to a riskier strategy with a higher value for the final expected wealth, but a higher risk associated with it.

Often the investor's attitude to risk is difficult to capture a priori in a single parameter. A better decision tool is the efficient frontier, a plot of $\mathbb{E}(X)$ against the corresponding $\mathrm{Var}(X)$ values for different settings of ρ. Computing the efficient frontier requires the solution of a series of problems for different values of ρ. Apart from this all the problems in the sequence are identical, which makes them prime candidates for a warmstarting strategy (although see [3] for a different approach). Table 12 gives results for four different problem sizes with up to 192 million variables and 70 million constraints. For each problem the top line gives the number of iterations a coldstarted IPM needs to solve the problem for a given value of ρ, whereas the middle line gives the number of iterations when warmstarting each problem from the one with the next lowest setting of ρ. The last line gives the percentage saving in IPM iterations. Again we are able to save in the range of 50 and 60% of IPM iterations. As far as we are aware these are the largest problems to date for which an interior point warmstart has been employed.

TABLE 12
Computation of efficient frontier with IPM warmstarts.

variables (n) constraints (m)	$\rho =$	1e-3	5e-3	0.01	0.05	0.1	0.5	1	5	10
$n = 223.321$ $m = 76.881$	cold	14	14	14	14	14	13	17	16	17
	warm	14	5	5	5	4	5	5	8	8
	red	0.0	64.2	64.2	64.2	71.4	61.5	70.5	50.0	52.9
$n = 533.725$ $m = 198.525$	cold	14	14	14	14	14	15	18	18	17
	warm	14	5	5	5	6	5	5	9	10
	red	0.0	64.3	64.3	64.3	57.1	66.7	72.2	50.0	41.2
$n = 16.316.191$ $m = 5.982.604$	cold	24	23	24	23	25	22	24	23	24
	warm	24	8	11	13	11	13	12	12	14
	red	0.0	65.2	54.2	43.5	56.0	40.9	50.0	47.8	41.7
$n = 192.478.111$ $m = 70.575.308$	cold	52	53	45	43	44	42	44	46	46
	warm	52	13	13	15	15	16	16	23	25
	red	0.0	75.5	71.1	65.1	65.9	61.9	63.6	50.0	45.6

6. Conclusions. In this paper we have compared the effectiveness of various interior point warmstarting schemes on the NETLIB base test set suggested by [1]. We have categorized warmstarting strategies into *modification* strategies and *unblocking* strategies. *Modification* strategies are aimed at modifying an advanced iterate from a previous solution of a nearby problem before it is used to warmstart an IPM, whereas unblocking strategies aim to directly address the negative effect known as blocking which typically affects a "bad" warmstart in the first few iterations. We suggest a new unblocking strategy based on sensitivity analysis of the step direction with respect to the current point. In our numerical tests we obtain an optimal combination of modification and unblocking strategies (including the new strategy based on sensitivity analysis) and are subsequently able to save an average of 50 and 60% of interior point iterations on a range of LP and QP problems varying from the small scale NETLIB test set to problems with over 192 million variables.

Acknowledgment. We would like to thank the anonymous referees for their constructive comments that helped to improve this paper.

REFERENCES

[1] H. Y. Benson and D. F. Shanno, *An exact primal-dual penalty method approach to warmstarting interior-point methods for linear programming*, Comput. Optim. Appl., 38 (2007), pp. 371–399.
[2] H. Y. Benson and D. F. Shanno, *Interior-point methods for nonconvex nonlinear programming: Regularization and warmstarts*, Comput. Optim. Appl., 40 (2008), pp. 143–189.
[3] J. Fliege, *An efficient interior-point method for convex multicriteria optimization problems*, Math. Oper. Res., 31 (2006), pp. 825–845.
[4] J. Gondzio, *HOPDM (version 2.12)—a fast LP solver based on a primal-dual interior point method*, European J. Oper. Res., 85 (1995), pp. 221–225.
[5] J. Gondzio, *Warm start of the primal-dual method applied in the cutting plane scheme*, Math. Program., 83 (1998), pp. 125–143.
[6] J. Gondzio and A. Grothey, *Reoptimization with the primal-dual interior point method*, SIAM J. Optim., 13 (2003), pp. 842–864.
[7] J. Gondzio and A. Grothey, *Parallel interior point solver for structured quadratic programs: Application to financial planning problems*, Ann. Oper. Res., 152 (2007), pp. 319–339.
[8] J. Gondzio and J.-P. Vial, *Warm start and ε-subgradients in cutting plane scheme for block-angular linear programs*, Comput. Optim. Appl., 14 (1999), pp. 17–36.
[9] A. L. Hipolito, *A weighted least squares study of robustness in interior point linear programming*, Comput. Optim. Appl., 2 (1993), pp. 29–46.

[10] E. John and E. A. Yildirim, *Implementation of warm-start strategies in interior-point methods for linear programming in fixed dimension*, Comput. Optim. Appl., 41 (2008), pp. 151–183.

[11] I. J. Lustig, R. E. Marsten, and D. F. Shanno, *Computational experience with a globally convergent primal-dual predictor-corrector algorithm for linear programming*, Math. Program., 66 (1994), pp. 123–135.

[12] I. Maros and C. Mészáros, *A repository of convex quadratic programming problems*, Technical report DOC 97/6, Department of Computing, Imperial College, London, U.K., 1997.

[13] S. Mehrotra, *On the implementation of a primal–dual interior point method*, SIAM J. Optim., 2 (1992), pp. 575–601.

[14] M. Steinbach, *Markowitz revisited: Mean-variance models in financial portfolio analysis*, SIAM Rev., 43 (2001), pp. 31–85.

[15] E. A. Yildirim and S. J. Wright, *Warm-start strategies in interior-point methods for linear programming*, SIAM J. Optim., 12 (2002), pp. 782–810.

SIAM J. OPTIM.
Vol. 19, No. 3, pp. 1211–1230

THE EXACT FEASIBILITY OF RANDOMIZED SOLUTIONS OF UNCERTAIN CONVEX PROGRAMS*

M. C. CAMPI† AND S. GARATTI‡

Abstract. Many optimization problems are naturally delivered in an uncertain framework, and one would like to exercise prudence against the uncertainty elements present in the problem. In previous contributions, it has been shown that solutions to uncertain convex programs that bear a high probability to satisfy uncertain constraints can be obtained at low computational cost through constraint randomization. In this paper, we establish new feasibility results for randomized algorithms. Specifically, the *exact* feasibility for the class of the so-called *fully-supported* problems is obtained. It turns out that all fully-supported problems share the same feasibility properties, revealing a deep kinship among problems of this class. It is further proven that the feasibility of the randomized solutions for all other convex programs can be bounded based on the feasibility for the prototype class of fully-supported problems. The feasibility result of this paper outperforms previous bounds and is not improvable because it is exact for fully-supported problems.

Key words. uncertain optimization, randomized methods, convex optimization, semi-infinite programming, robust optimization, chance-constrained

AMS subject classifications. 90C25, 90C15, 90C34, 68W20

DOI. 10.1137/07069821X

1. Introduction.

Uncertain convex optimization [21, 24, 25] deals with convex optimization in which the constraints are imprecisely known:

$$\begin{aligned} \text{UP}: \ & \min_{x \in \mathcal{X} \subseteq \mathbb{R}^d} c^T x \\ & \text{subject to: } x \in \mathcal{X}_\delta, \quad \delta \in \Delta, \end{aligned} \tag{1}$$

where UP stands for uncertain program, $\delta \in \Delta$ is an uncertain parameter, and $\mathcal{X}$ and $\mathcal{X}_\delta$ are convex and closed sets. Oftentimes, Δ is a set of infinite cardinality. The fact that the optimization objective is linear and does not carry any dependence on δ, that is, it is certain, is without loss of generality.

UP encompasses as special cases uncertain linear programs (LP), uncertain quadratic programs (QP), uncertain second-order cone programs (SOCP), and uncertain semi-definite programs (SDP) and plays a central role in many design endeavors such as [1, 15, 17, 14, 9, 24, 11, 6].

Dealing with uncertainty can be done along two distinct approaches. The first one consists in enforcing satisfaction of *all* constraints; that is, one optimizes the cost $c^T x$ over the set $\bigcap_{\delta \in \Delta} \mathcal{X}_\delta$ (see [2, 16, 3, 4]). Alternatively, one may want to satisfy the constraints with "high probability." Along this second approach one sees the uncertainty parameter δ as a random element with a probability $\mathbb{P}$ and seeks a solution that violates at most a fraction of constraints having small $\mathbb{P}$-probability (chance-constrained solution). Depending on the optimization problem at hand, $\mathbb{P}$

*Received by the editors July 24, 2007; accepted for publication (in revised form) May 13, 2008; published electronically November 19, 2008. This paper was supported by the MIUR national project "Identification and adaptive control of industrial systems."
http://www.siam.org/journals/siopt/19-3/69821.html

†Dipartimento di Elettronica per l'Automazione, Università di Brescia, via Branze 38, 25123 Brescia, Italia (marco.campi@ing.unibs.it, http://bsing.ing.unibs.it/∼campi/).

‡Dipartimento di Elettronica ed Informazione, Politecnico di Milano, piazza L. da Vinci 32, 20133 Milano, Italia (sgaratti@elet.polimi.it, http://home.dei.polimi.it/sgaratti/).

can have different interpretations. Sometimes, it is the actual probability with which the uncertainty parameter δ takes on value in Δ. Other times, it simply describes the relative importance attributed to different instances of δ. The use of a probabilistic description of uncertainty has a long history in optimization theory and dates back to the work [10] of Charnes and Cooper in the 1950s that in effect gave birth to the chance-constrained approach. See also [21, 22, 12, 25] for more information and [5] for a more in-depth comparison between deterministic and probabilistic uncertain optimization.

It is a fact that finding a solution carrying a high probability of constraint satisfaction is in general a very difficult task to achieve [21]. To circumvent this computational issue, recently, methodologies relying on the randomization over the set of constraints have been introduced [11, 5, 20, 6, 13]. Specifically, in [5, 6], the following randomized program RP_N is introduced, where N constraints $\delta^{(1)}, \ldots, \delta^{(N)}$ randomly extracted according to $\mathbb{P}$ in an independent fashion are simultaneously enforced:

$$\mathrm{RP}_N : \min_{x \in \mathcal{X} \subseteq \mathbb{R}^d} c^T x$$
$$\text{subject to: } x \in \bigcap_{i \in \{1,\ldots,N\}} \mathcal{X}_{\delta^{(i)}}.$$

RP_N is also known as "scenario program."

The distinctive feature of RP_N is that it is a program with a *finite* number of constraints, and, as such, it can be solved at low computational cost provided that N is not too large;[1] it is indeed a fact that RP_N has opened up new resolution avenues in uncertain optimization. On the other hand, the obvious question to ask is how feasible the solution of RP_N is; that is, how large the fraction of original constraints in Δ that are possibly violated by the solution x_N^* of RP_N is. Papers [5, 6] have pioneered a feasibility theory showing that x_N^* is feasible for the vast majority of the other unseen constraints—those that have not been used when optimizing according to RP_N—and this result holds in full generality, independently of the structure of the set of constraints Δ and the probability $\mathbb{P}$. So the vast majority of constraints take care of themselves, without explicitly accounting for them.

To allow for a sharper comparison with the results presented in this paper, we feel advisable to first recall in some detail the results in [5, 6]. The following notion of violation probability from [5] is central.

DEFINITION 1 (violation probability). *The* violation probability *of a given $x \in \mathcal{X}$ is defined as $V(x) = \mathbb{P}\{\delta \in \Delta : x \notin \mathcal{X}_\delta\}$.*

The problem addressed in [5, 6] is to evaluate if and when the violation probability of x_N^*, namely, $V(x_N^*)$, is below a satisfying level ϵ. To state the result precisely, note that $V(x_N^*)$ is a random variable since the solution x_N^* of RP_N is, due to the fact that it depends on the random extractions $\delta^{(1)}, \ldots, \delta^{(N)}$. Thus, $V(x_N^*) \leq \epsilon$ may hold for certain extractions $\delta^{(1)}, \ldots, \delta^{(N)}$, while $V(x_N^*) > \epsilon$ may be true for others. The following quantification of the "bad" extractions where $V(x_N^*) > \epsilon$ is the key result of [6]:

$$\mathbb{P}^N\{V(x_N^*) > \epsilon\} \leq \binom{N}{d}(1-\epsilon)^{N-d}. \tag{2}$$

[1] Depending on Δ and $\mathbb{P}$, the generation of N randomly extracted scenarios $\delta^{(1)}, \ldots, \delta^{(N)}$ from Δ can in itself be a nontrivial problem, and the reader is referred to [27, 8, 7] for further discussion on this issue.

Moving a fundamental step forward with respect to [6], in this paper we establish the validity of relation

$$\mathbb{P}^N\{V(x_N^*) > \epsilon\} = \sum_{i=0}^{d-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i} \tag{3}$$

(note that (3) holds with "="; that is, it is not a bound) for the prototype class of *fully-supported* problems according to Definition 3 in section 2. This result sheds new light on a deep kinship among all fully-supported problems, proving that their randomized solutions share the same violation properties, and writes a final word on the violation assessment for this type of problems.

It is further proven in this paper that the right-hand side of (3) is an upper bound for all convex problems; that is,

$$\mathbb{P}^N\{V(x_N^*) > \epsilon\} \leq \sum_{i=0}^{d-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i} \tag{4}$$

holds for all convex problems. Thus, in a real optimization problem one does not need to verify whether the problem is fully-supported or not, and $\sum_{i=0}^{d-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}$ can always be used as an upper bound for $\mathbb{P}^N\{V(x_N^*) > \epsilon\}$. This result (4) (i) cannot be improved (being tight for the prototype class of fully-supported problems) and (ii) outperforms the previous bound from [6] at times by a huge extent (note that when $\epsilon \to 0$, the previous bound (2) tends to $\binom{N}{d}$, while the new bound (4) goes naturally to 1!).

2. Main result. The technical result of this paper is precisely stated in this section, followed by a discussion on the significance of the result.

For a fixed integer m and fixed given constraints $\delta^{(1)}, \ldots, \delta^{(m)}$, program

$$\begin{aligned} \mathrm{P}_m : & \min_{x \in \mathcal{X} \subseteq \mathbb{R}^d} c^T x \\ & \text{subject to: } x \in \bigcap_{i \in \{1,\ldots,m\}} \mathcal{X}_{\delta^{(i)}} \end{aligned} \tag{5}$$

is called a *finite instance with m constraints* of the uncertain optimization program UP in (1). For the time being, we make the following assumption.

Assumption 1. Every P_m is feasible, and its feasibility domain has a nonempty interior. Moreover, the solution x_m^* of P_m exists and is unique.

The existence and uniqueness of x_m^* are here assumed to streamline the presentation. The reader is referred to point 5 in the discussion in section 2.1 to see how these assumptions can be removed.

We recall the following fundamental definition and proposition. Definition 2 was introduced in [5], while Proposition 1 was originally stated in a slightly different but equivalent way in [18].

DEFINITION 2 (support constraint). *Constraint $\delta^{(r)}$, $r \in \{1, \ldots, m\}$, is a* support constraint *for P_m if its removal changes the solution of P_m.*

PROPOSITION 1. *The number of support constraints for P_m is at most d, the size of x.*

Suppose now that Δ is endowed with a σ-algebra $\mathcal{D}$ and that a probability $\mathbb{P}$ over $\mathcal{D}$ is assigned. Further assume that m constraints $\delta^{(1)}, \ldots, \delta^{(m)}$ are randomly extracted from Δ according to $\mathbb{P}$ in an independent fashion. Differently stated, the

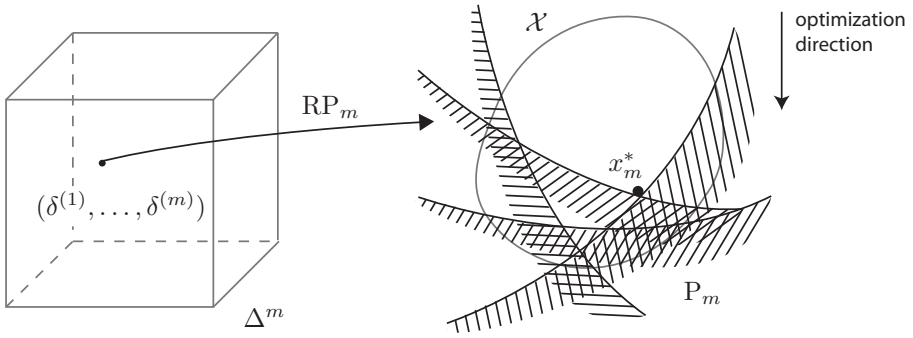


FIG. 1. *RP_m, a map from constraint multiextractions to finite instances P_m of the optimization problem.*

multiextraction $(\delta^{(1)}, \ldots, \delta^{(m)})$ is a random element from the probability space Δ^m equipped with the product probability $\mathbb{P}^m$. Each multiextraction $(\delta^{(1)}, \ldots, \delta^{(m)})$ generates a program P_m, and the map from Δ^m to P_m programs is a *randomized program* RP_m; see Figure 1. Note that this is the same as RP_N in section 1 with the only difference being that we have used here m to indicate the number of constraints, a choice justified by the fact that in this section m plays the role of a generic running argument taking on any integer value, while N represents in section 1 the fixed number of constraints picked by the user for the implementation of the randomized scheme.

We are now ready to introduce the notion of a *fully-supported* problem.

DEFINITION 3 (fully-supported problem). *A finite instance P_m, with $m \geq d$, is fully-supported if the number of support constraints of P_m is* exactly *d. Problem UP is fully-supported if, for any $m \geq d$, RP_m is fully-supported with probability* 1.

The main result of this paper is now stated in the following theorem.

THEOREM 1. *Under Assumption* 1,[2] *it holds that*

$$\mathbb{P}^N\{V(x_N^*) > \epsilon\} \leq \sum_{i=0}^{d-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}; \tag{6}$$

moreover, the bound is tight for all fully-supported *uncertain optimization problems; that is,*

$$\mathbb{P}^N\{V(x_N^*) > \epsilon\} = \sum_{i=0}^{d-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}. \tag{7}$$

The proof is given in section 3. The measurability of $\{V(x_N^*) > \epsilon\}$, as well as the measurability of other sets, is assumed in this paper.

One interpretation of Theorem 1 is that the randomized solution is, with high probability, a feasible solution for a chance-constrained problem; see [21].

2.1. Discussion. The following comments are in order.

1. Equation (7) delivers the *exact* probability distribution of the violation $V(x_N^*)$ for all fully-supported problems. Since (7) holds independently of the nature and characteristics of the fully-supported problem, it establishes a fundamental kinship among problems of this prototype class.

[2]See point 5 in section 2.1 for relaxations of this assumption.

TABLE 1
β vs. β_{old} for different values of N ($\epsilon = 0.05$, $d = 10$).

N	150	300	450	600	750
β	0.78	0.06	$8.8\ 10^{-4}$	$4.8\ 10^{-6}$	$1.5\ 10^{-8}$
β_{old}	$8.8\ 10^{11}$	$4.8\ 10^{11}$	$1.3\ 10^{10}$	$1.1\ 10^{8}$	$4.8\ 10^{5}$

N	900	1050	1200	1350	1500
β	$3.5\ 10^{-11}$	$6.2\ 10^{-14}$	$9.2\ 10^{-17}$	$1.2\ 10^{-19}$	$1.4\ 10^{-22}$
β_{old}	$1.3\ 10^{3}$	2.9	$5.1\ 10^{-3}$	$7.5\ 10^{-6}$	$9.9\ 10^{-9}$

TABLE 2
N vs. N_{old} for different values of ϵ ($\beta = 10^{-5}$, $d = 10$).

ϵ	0.1	0.05	0.025	0.01	0.005	0.0025	0.001
N	285	581	1171	2942	5895	11749	29513
N_{old}	579	1344	3035	8675	18943	41008	112686

Bound (6) further asserts that all possible sources of non-fully-supportedness can only improve the feasibility properties of the problem.

2. The quantity $\beta := \sum_{i=0}^{d-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}$ in the right-hand side of (6) and (7) is the tail of a binomial distribution and goes rapidly (exponentially) to zero as N increases. Letting $\beta_{\text{old}} := \binom{N}{d}(1-\epsilon)^{N-d}$ (bound in (2) from [6]), Table 1 provides a comparison between β and β_{old}.

3. A typical use of Theorem 1 consists in selecting ϵ (violation parameter) and β (confidence parameter) in $(0, 1)$ and then computing the smallest number N of constraints to be extracted in order to guarantee that $\mathbb{P}^N\{V(x_N^*) > \epsilon\} \leq \beta$ by solving equation $\beta = \sum_{i=0}^{d-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}$ for N. In Table 2, the values of N and of N_{old} obtained by using the bound in (2) are displayed for different values of ϵ, $\beta = 10^{-5}$ and $d = 10$.

4. A simple example illustrates Theorem 1.

$N = 1650$ points are independently extracted in $\mathbb{R}^2$ according to an unknown probability density $\mathbb{P}$, and the strip of smaller vertical width that contains all of the points is constructed; see Figure 2.

In mathematical terms—letting the points be $(u^{(i)}, y^{(i)})$, $i = 1, \ldots, N$, where u is the horizontal coordinate and y is the vertical coordinate—this amounts to solving the following program:

$$\begin{aligned} \mathrm{P}_N : \min_{x_1, x_2, x_3 \in \mathbb{R}^3} \ & x_1 \\ \text{subject to: } & \left| y^{(i)} - [x_2 u^{(i)} + x_3] \right| \leq x_1, \quad i = 1, \ldots, N, \end{aligned}$$

where $[x_2 u^{(i)} + x_3]$ is the median line of the strip and x_1 is the semiwidth of the strip.

What guarantee do we have that the strip contains at least 99% of the probability mass of $\mathbb{P}$?

One can easily recognize that this question is the same as asking for a guarantee, or a probability, that the violation is less than $\epsilon = 0.01$, and the answer can be found in Theorem 1: this probability is no less than $1 - \sum_{i=0}^{2} \binom{1650}{i} 0.01^i (1-0.01)^{1650-i} \approx 1 - 10^{-5}$. As a matter of fact, this probability is exact since, as it can be verified, this problem is fully-supported.

We can further ask for a different geometrical construction and look for the disk of smaller radius that contains all points; see Figure 3. Again, we are facing a finite

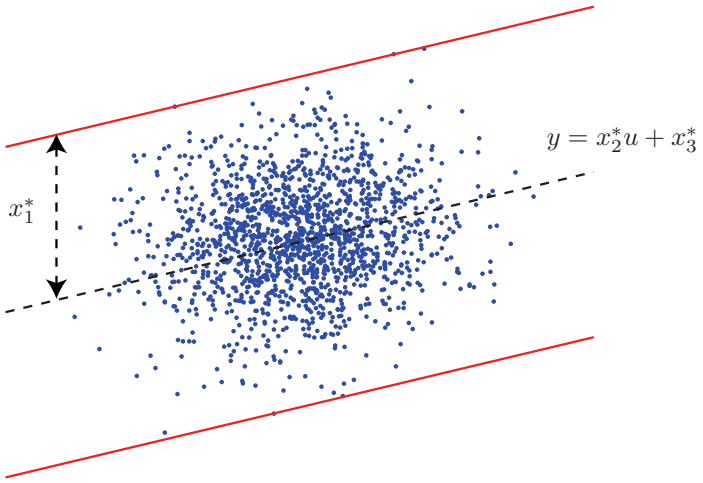


FIG. 2. *Strip of smaller vertical width.*

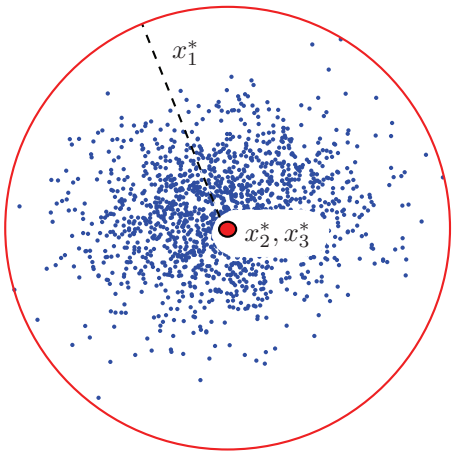


FIG. 3. *Disk of smaller radius.*

convex program

$$\mathrm{P}_N : \min_{x_1,x_2,x_3\in\mathbb{R}^3} x_1$$
$$\text{subject to: } \sqrt{(u^{(i)}-x_2)^2+(y^{(i)}-x_3)^2} \le x_1, \quad i=1,\dots,N,$$

where (x_2, x_3) is the center of the disk and x_1 is its radius, and again we can claim with confidence $1-10^{-5}$ that the constructed disk will contain at least 99% of the probability mass. In this disk case, the figure $1-10^{-5}$ is a lower bound since the problem is not fully-supported, as it can be easily recognized by noting that a configuration with two points away from each other and all of the other points concentrated near the midposition of the first two points generates a disk where the segment joining the first two points is a diameter and only these two points are of support.

Finally, let us compare the probability $1-10^{-5}$ with the probability that would have been obtained by applying the previous bound (2) from [6]. Applying the latter,

we would find that this probability is no smaller than $1 - 48.4 = -47.4$, a negative number clearly devoid of any meaning and that does not allow us to draw any conclusion as far as the confidence is concerned.

5. Here we discuss the assumption of the existence and uniqueness of the solution of P_m. Suppose first that the solution exists but it may be nonunique. Then, the tie can be broken by selecting among the optimal solutions the one with the minimum Euclidian norm, and one can prove that Theorem 1 holds unchanged.

If we further relax the assumption that the solution exists (note that the solution may not exist even if P_m is feasible since the solution can drift away to infinity), extending Theorem 1 we can show that

$$\mathbb{P}^N\{x_N^* \text{ exists, and } V(x_N^*) > \epsilon\} \leq \sum_{i=0}^{d-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i},$$

where x_N^* is unique after applying the tie-break rule as above. In words, this result says that, when a solution is found, its violation exceeds ϵ with small probability only. In normal problems the nonexistence of the solution is a rare event whose probability exponentially vanishes with N.

3. Proof of Theorem 1. We first prove that $\mathbb{P}^N\{V(x_N^*) > \epsilon\} = \sum_{i=0}^{d-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}$ for *fully-supported* problems and then that $\mathbb{P}^N\{V(x_N^*) > \epsilon\} \leq \sum_{i=0}^{d-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}$ for every problem.

PART 1: $\mathbb{P}^N\{V(x_N^*) > \epsilon\} = \sum_{i=0}^{d-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}$ FOR FULLY-SUPPORTED PROBLEMS. Consider the solution x_d^* of RP_d (recall that d is the size of x), and let

$$F(\alpha) := \mathbb{P}^d\{V(x_d^*) \leq \alpha\} \tag{8}$$

be the probability distribution of the violation of x_d^*. It is a remarkable fact that this distribution is

$$F(\alpha) = \alpha^d, \tag{9}$$

independent of the problem type.

To prove (9), we have to consider multiextractions of m elements, where m is a generic integer greater than or equal to d. To each multiextraction $(\delta^{(1)}, \ldots, \delta^{(m)}) \in \Delta^m$, associate the indexes of the corresponding d support constraints (this is always possible except for a probability 0 set because the problem is *fully-supported*).[3] Further, group all multiextractions having the same indexes. In this way, $\binom{m}{d}$ sets $S_{\mathcal{I}}$ are constructed forming a partition (up to a probability 0 set) of Δ^m, where $\mathcal{I} \subset \{1, \ldots, m\}$ is a set of cardinality d containing the indexes of the support constraints. We claim that the probability of each of these sets is

$$\mathbb{P}^m\{S_{\mathcal{I}}\} = \int_0^1 (1-\alpha)^{m-d} F(\mathrm{d}\alpha), \tag{10}$$

where $F(\alpha)$ is defined in (8); using (10), later on in the proof, we shall show that $F(\alpha)$ must have the expression in (9).

[3]The fact that a fully-supported problem is one where the RP_m are fully-supported with probability 1, as opposed to *always* fully-supported, is a source of a bit of complication in the proof. On the other hand, requiring always fully-supportedness is too limitative since, e.g., extracting the same constraint m times results in a non-fully-supported P_m.

To establish (10) in a more concrete way, consider one of the sets $S_{\mathcal{I}}$, e.g., the set $S_{\bar{\mathcal{I}}}$ where the support constraints indexes are $1, \ldots, d$. Also let $\widetilde{S}_{\bar{\mathcal{I}}}$ be the set where $\delta^{(d+1)}, \ldots, \delta^{(m)}$ are not violated by the solution generated by $\delta^{(1)}, \ldots, \delta^{(d)}$. It is an intuitive fact that $S_{\bar{\mathcal{I}}}$ and $\widetilde{S}_{\bar{\mathcal{I}}}$ are the same up to a probability 0 set. To streamline the presentation, we accept in the following this fact for granted; however, the interested reader can find full details at the end of this Part 1 of the proof.

We next compute $\mathbb{P}^m\{\widetilde{S}_{\bar{\mathcal{I}}}\}$, which is the same as $\mathbb{P}^m\{S_{\bar{\mathcal{I}}}\}$.

Select fixed values for $\bar{\delta}^{(1)}, \ldots, \bar{\delta}^{(d)}$, and let α be the violation of the solution with these d constraints only. Then, the probability that $\delta^{(d+1)}, \ldots, \delta^{(m)}$ fall in the nonviolated set, that is, $(\bar{\delta}^{(1)}, \ldots, \bar{\delta}^{(d)}, \delta^{(d+1)}, \ldots, \delta^{(m)}) \in \widetilde{S}_{\bar{\mathcal{I}}}$, is $(1-\alpha)^{m-d}$.

Integrating over the domain Δ^d for $(\bar{\delta}^{(1)}, \ldots, \bar{\delta}^{(d)})$, we then have

$$\begin{aligned}
&\mathbb{P}^m\left\{\widetilde{S}_{\bar{\mathcal{I}}}\right\} \\
&= [\text{letting } x^*_{\bar{\mathcal{I}}} \text{ be the solution with constraints } \bar{\delta}^{(1)}, \ldots, \bar{\delta}^{(d)} \text{ only}] \\
&= \int_{\Delta^d} (1-\alpha(x^*_{\bar{\mathcal{I}}}))^{m-d} \mathbb{P}^d(\mathrm{d}\bar{\delta}^{(1)}, \ldots, \mathrm{d}\bar{\delta}^{(d)}) \\
&= \int_0^1 (1-\alpha)^{m-d} F(\mathrm{d}\alpha),
\end{aligned}$$

where the third equality is a change of variables from the domain $(\bar{\delta}^{(1)}, \ldots, \bar{\delta}^{(d)})$ to that of the violation of the corresponding solution.

Since $\mathbb{P}^m\{S_{\bar{\mathcal{I}}}\} = \mathbb{P}^m\{\widetilde{S}_{\bar{\mathcal{I}}}\}$ and this probability is the same for any other set $S_{\mathcal{I}}$, (10) remains proven.

Now turn back to (9). Recalling that the sets $S_{\mathcal{I}}$ form a partition of Δ^m up to a probability 0 set and that $\mathbb{P}^m\{\Delta^m\} = 1$, (10) yields

$$\binom{m}{d} \int_0^1 (1-\alpha)^{m-d} F(\mathrm{d}\alpha) = 1 \quad \forall m \geq d. \tag{11}$$

Expression $F(\alpha) = \alpha^d$ in (9) is indeed a solution of (11) (integration by parts); on the other hand, no other solutions exist since determining an F satisfying (11) is a moment problem for a distribution with finite support, and its solution is unique; see, e.g., Chapter II, section 12.9, Corollary 1 of [26]. Thus, it remains proven that $F(\alpha)$ must have the expression (9).

To conclude the proof of Part 1, consider now the problem with N constraints and partition set $\{(\delta^{(1)}, \ldots, \delta^{(N)}) : \quad V(x^*_N) > \epsilon\}$ by intersecting it with the $\binom{N}{d}$ sets $S_{\mathcal{I}}$ grouping multiextractions such that the d support constraints have the same indexes. We then have

$$\begin{aligned}
&\mathbb{P}^N\{V(x^*_N) > \epsilon\} \\
&= \mathbb{P}^N\big\{ \cup_{\mathcal{I}} \{V(x^*_N) > \epsilon \text{ and } x^*_N \text{ is supported by the constraints} \\
&\qquad \text{with indexes in } \mathcal{I}\}\big\} \\
&= [\mathbb{I}_A \text{ is the indicator function of set } A; \text{ i.e., } \mathbb{I}_A = 1 \text{ over } A, \text{ and } \mathbb{I}_A = 0 \text{ otherwise}] \\
&= \binom{N}{d} \int_{\Delta^d} (1-\alpha(x^*_{\bar{\mathcal{I}}}))^{N-d} \mathbb{I}_{\{\alpha(x^*_{\bar{\mathcal{I}}}) > \epsilon\}} \mathbb{P}^d(\mathrm{d}\bar{\delta}^{(1)}, \ldots, \mathrm{d}\bar{\delta}^{(d)}) \\
&= \binom{N}{d} \int_\epsilon^1 (1-\alpha)^{N-d} \; F(\mathrm{d}\alpha)
\end{aligned}$$

$$
\begin{aligned}
&= [\text{since } F(\mathrm{d}\alpha) = d\alpha^{d-1}\,\mathrm{d}\alpha] \\
&= \binom{N}{d} \int_\epsilon^1 \left[(1-\alpha)^{N-d} d\alpha^{d-1}\right]\,\mathrm{d}\alpha \\
&= [\text{integrating by parts}] \\
&= \binom{N}{d}\left[-\frac{(1-\alpha)^{N-d+1}}{N-d+1} d\alpha^{d-1}\Big|_\epsilon^1 + \int_\epsilon^1 \frac{(1-\alpha)^{N-d+1}}{N-d+1} d(d-1)\alpha^{d-2}\,\mathrm{d}\alpha\right] \\
&= \binom{N}{d-1}\epsilon^{d-1}(1-\epsilon)^{N-d+1} + \binom{N}{d-1}\int_\epsilon^1 (1-\alpha)^{N-d+1}(d-1)\alpha^{d-2}\,\mathrm{d}\alpha \\
&= \cdots \\
&= \binom{N}{d-1}\epsilon^{d-1}(1-\epsilon)^{N-d+1} + \ldots + \binom{N}{1}\epsilon(1-\epsilon)^{N-1} + \binom{N}{1}\int_\epsilon^1 (1-\alpha)^{N-1}\,\mathrm{d}\alpha \\
&= \sum_{i=0}^{d-1}\binom{N}{i}\epsilon^i(1-\epsilon)^{N-i}.
\end{aligned}
$$

Proof of the fact that $S_{\bar{\mathcal{I}}} = \widetilde{S}_{\bar{\mathcal{I}}}$ up to a probability zero set.

$\boldsymbol{S_{\bar{\mathcal{I}}} \subseteq \widetilde{S}_{\bar{\mathcal{I}}}}$: Take a $(\delta^{(1)}, \ldots, \delta^{(m)}) \in S_{\bar{\mathcal{I}}}$ and eliminate a constraint among $\delta^{(d+1)}$, $\ldots, \delta^{(m)}$. Since this constraint is not of support, the solution remains unchanged; moreover, it is easy to see that the first d constraints are still the support constraints for the problem with $m-1$ constraints. If we now remove another constraint among those which are not of support, the conclusion is similarly drawn that the solution remains unchanged and that the first d constraints are still the support constraints for the problem with $m-2$ constraints. Proceeding this way until all constraints but the first d are removed, we obtain that the solution with the sole d support constraints $\delta^{(1)}, \ldots, \delta^{(d)}$ in place is the same as the solution with all m constraints. Since no constraint among $\delta^{(d+1)}, \ldots, \delta^{(m)}$ can be violated by the solution with all m constraints and such a solution is the same as the one with only the first d constraints, it follows that $(\delta^{(1)}, \ldots, \delta^{(m)}) \in \widetilde{S}_{\bar{\mathcal{I}}}$.

$\boldsymbol{\widetilde{S}_{\bar{\mathcal{I}}} \subseteq S_{\bar{\mathcal{I}}}}$ **up to a probability 0 set**: Suppose now that $\delta^{(d+1)}, \ldots, \delta^{(m)}$ are not violated by the solution generated by $\delta^{(1)}, \ldots, \delta^{(d)}$, i.e., $(\delta^{(1)}, \ldots, \delta^{(m)}) \in \widetilde{S}_{\bar{\mathcal{I}}}$. A simple reasoning reveals that $(\delta^{(1)}, \ldots, \delta^{(m)})$ does not belong to any one of sets $S_{\mathcal{I}}$, $\mathcal{I} \neq \bar{\mathcal{I}}$. In fact, adding nonviolated constraints to $\delta^{(1)}, \ldots, \delta^{(d)}$ does not change the solution, and each of the added constraints can be removed back without altering the solution. Therefore, none of the constraints $\delta^{(d+1)}, \ldots, \delta^{(m)}$ can be of support, and hence the multiextraction is not in $S_{\mathcal{I}}$, $\mathcal{I} \neq \bar{\mathcal{I}}$. It follows that $\widetilde{S}_{\bar{\mathcal{I}}}$ is a subset of the complement of $\cup_{\mathcal{I}, \mathcal{I} \neq \bar{\mathcal{I}}} S_{\mathcal{I}}$, which is $S_{\bar{\mathcal{I}}}$ up to a probability 0 set.

PART 2: $\boldsymbol{\mathbb{P}^N\{V(x_N^*) > \epsilon\} \leq \sum_{i=0}^{d-1}\binom{N}{i}\epsilon^i(1-\epsilon)^{N-i}$ FOR EVERY PROBLEM. A non-fully-supported problem admits with nonzero probability randomized instances where the number of support constraints is less than d. A support constraint has to be an active constraint, and the typical reason for a lack of support constraints is that at the optimum the active constraints are less than d; see Figure 4. To carry on a proof along lines akin to those for the fully-supported case, we are well-advised to generalize the notion of solution to that of ball-solution; a ball-solution has always at least d active constraints. For simplicity, we henceforth assume that constraints are not trivial, i.e., $\mathcal{X}_\delta \neq \mathbb{R}^d\ \forall \delta \in \Delta$.

Roughly speaking, given an optimization problem whose solution is x_m^*, its *ball-solution* is a ball centered in x_m^* and whose radius has been enlarged until the ball

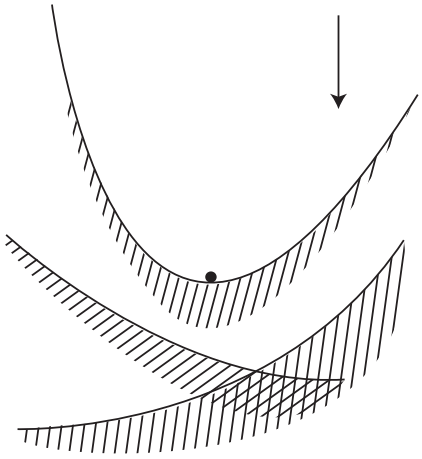


FIG. 4. *A two-dimensional problem with only one active constraint which is of support.*

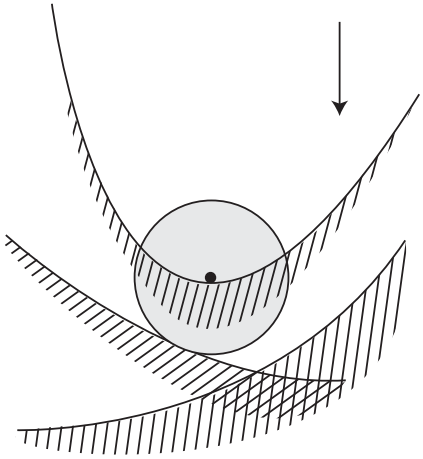


FIG. 5. *Ball-solution.*

touches the frontier of d constraints. See Figure 5 for an example of a ball-solution. The mathematical definition of a ball-solution is as follows.

DEFINITION 4 (ball-solution). *Consider a finite instance P_m of UP with $m \geq d$, and let x_m^* be its solution. The* ball-solution $\mathcal{B}(x_m^*, r_m^*)$ *of P_m is the largest closed ball centered in x_m^* fully contained in the feasibility domain of all constraints, with the exception of at most $d-1$ of them; i.e., $\mathcal{X}_{\delta^{(i)}} \cap \mathcal{B}(x_m^*, r_m^*) = \mathcal{B}(x_m^*, r_m^*)$ for all i's, except at most $d-1$ of them.*

Note also that, when active constraints are d or more, $r_m^* = 0$ and $\mathcal{B}(x_m^*, r_m^*)$ reduces to the standard solution x_m^*. Moreover, a ball-solution $\mathcal{B}(x_m^*, r_m^*)$ need not be contained in $\mathcal{X}$, although its center x_m^* does.

The notion of active constraint can be generalized to balls by saying that a constraint is active for a ball if the ball touches the frontier of the constraint. If in addition the ball is fully contained in the constraint, then the constraint is said to be strictly active. See Figure 6 for a graphical illustration of active and strictly active constraints for a ball, while the precise definition is as follows.

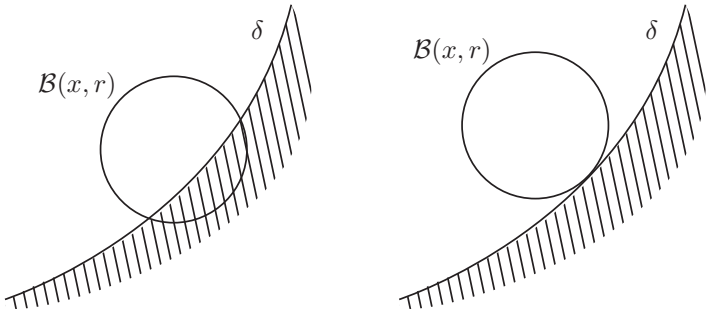


FIG. 6. *Active and strictly active constraints for a ball.*

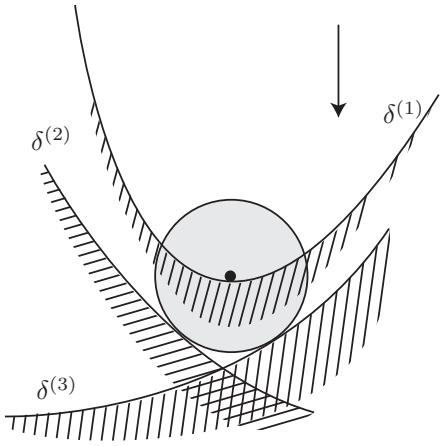


FIG. 7. *Only $\delta^{(1)}$ is a ball-support constraint.*

DEFINITION 5 (active constraint for a ball). *A constraint δ is* active *for a ball $\mathcal{B}(x,r)$ if $\mathcal{X}_\delta \cap \mathcal{B}(x,r) \neq \emptyset$ and $\mathcal{X}_\delta \cap \mathcal{B}(x,r+h) \neq \mathcal{B}(x,r+h)$ $\forall h > 0$. If in addition $\mathcal{X}_\delta \cap \mathcal{B}(x,r) = \mathcal{B}(x,r)$, $\mathcal{X}_\delta$ is said to be* strictly active.

If the ball is a single point, active and strictly active are the same and reduce to the standard notion of active.

By construction, a ball-solution has at least d active constraints. To go back to the track of the proof in Part 1, however, we need d support constraints, not just active constraints. The following definition naturally extends the notion of support constraint to the case of ball-solutions.

DEFINITION 6 (ball-support constraint). *Constraint $\delta^{(r)}$, $r \in \{1,\dots,m\}$, is a* ball-support constraint *for P_m if its removal changes the ball-solution of P_m.*

An active constraint is not necessarily a ball-support constraint, nor does a P_m necessarily have to have d ball-support constraints (see Figure 7, where $\delta^{(2)}$ and $\delta^{(3)}$ are not of support). It is clear that the number of ball-support constraints is less than or equal to d. The case with less than d ball-support constraints is regarded as *degenerate* and needs to be treated separately. We thus split the remaining part of the proof in two sections: Part 2a (fully-ball-supported problems) and Part 2b (degenerate problems). Before proceeding, we are well-advised to give a formal definition of fully-ball-supported problems.

DEFINITION 7 (fully-ball-supported problem). *A finite instance P_m, with $m \geq d$, is fully-ball-supported if the number of ball-support constraints of P_m is d. Problem UP is fully-ball-supported if, for any $m \geq d$, RP_m is fully-ball-supported with probability* 1.

PART 2a: FULLY-BALL-SUPPORTED PROBLEMS. We start by introducing the notion of a constraint violated by a ball: a constraint δ is violated by $\mathcal{B}(x,r)$ if $\mathcal{X}_\delta \cap \mathcal{B}(x,r) \neq \mathcal{B}(x,r)$. The definition of probability of violation then generalizes naturally to the ball case.

DEFINITION 8 (violation probability of a ball). *The* violation probability *of a ball $\mathcal{B}(x,r)$, $x \in \mathcal{X}$, is defined as $V_{\mathcal{B}}(x,r) = \mathbb{P}\{\delta \in \Delta : \mathcal{X}_\delta \cap \mathcal{B}(x,r) \neq \mathcal{B}(x,r)\}$.*

Clearly, for any x, $V_{\mathcal{B}}(x,r) \geq V(x)$. Hence, if $\mathcal{B}(x^*_N, r^*_N)$ is the ball-solution of RP_N, we have

$$\mathbb{P}^N\{V(x^*_N) > \epsilon\} \leq \mathbb{P}^N\{V_{\mathcal{B}}(x^*_N, r^*_N) > \epsilon\}. \tag{12}$$

Below, we show that a result similar to (7) holds for fully-ball-supported problems, namely,

$$\mathbb{P}^N\{V_{\mathcal{B}}(x^*_N, r^*_N) > \epsilon\} = \sum_{i=0}^{d-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}, \tag{13}$$

and this result together with (12) leads to the thesis

$$\mathbb{P}^N\{V(x^*_N) > \epsilon\} \leq \sum_{i=0}^{d-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}.$$

The proof of (13) is verbatim the same as the proof of Part 1 provided that one substitutes

- *solution* with *ball-solution*,
- *support constraint* with *ball-support constraint*,
- *violation probability V* with *violation probability of a ball $V_{\mathcal{B}}$*,

with only one exception: the part where we proved that $\boldsymbol{S}_{\bar{\boldsymbol{\mathcal{I}}}} \subseteq \widetilde{\boldsymbol{S}}_{\bar{\boldsymbol{\mathcal{I}}}}$ has to be modified in a way that we spell out in the following.

The first rationale to conclude that "the solution with only the d support constraints $\delta^{(1)}, \ldots, \delta^{(d)}$ in place is the same as the solution with all m constraints" is still valid and leads in our present context to the fact that the ball-solution with only the d ball-support constraints $\delta^{(1)}, \ldots, \delta^{(d)}$ in place is the same as the ball-solution with all m constraints. Instead, the last argument with which we concluded that $\boldsymbol{S}_{\bar{\boldsymbol{\mathcal{I}}}} \subseteq \widetilde{\boldsymbol{S}}_{\bar{\boldsymbol{\mathcal{I}}}}$ is no longer valid since ball-solutions can violate constraints.

To amend it, suppose for the purpose of contradiction that a constraint among $\delta^{(d+1)}, \ldots, \delta^{(m)}$, say, $\delta^{(d+1)}$, is violated by the ball-solution with d constraints. Two cases can occur: (i) the ball-solution has only one strictly active constraint among $\delta^{(1)}, \ldots, \delta^{(d)}$; or (ii) it has more than one. In case (i), $d-1$ constraints among $\delta^{(1)}, \ldots, \delta^{(d)}$ are violated by the ball-solution so that, with the extra $\delta^{(d+1)}$ violated constraint, the number of violated constraints of the ball-solution with m constraints would add up to at least d, and this contradicts the definition of ball-solution. If instead (ii) is true, a simple thought reveals that, with one more constraint $\delta^{(d+1)}$ violated by the ball-solution, the strictly active constraints (which, in this case, are more than one) cannot be of ball-support for the problem with m constraints, and this contradicts the fact that $(\delta^{(1)}, \ldots, \delta^{(m)}) \in S_{\bar{\mathcal{I}}}$.

PART 2b: DEGENERATE PROBLEMS. For not being fully-ball-supported, a finite problem P_m needs to have more than one strictly active constraint, a circumstance which requires that constraints are not "generically" distributed. This observation is at the basis of the rather technical proof of Part 2b, which proceeds along the following steps:

Step 1. A constraint "heating" is introduced; heating scatters constraints around, and the resulting heated problem is shown to be fully-ball-supported; by resorting to the result in Part 2a, conclusions are derived about the violation properties of the heated problem.

Step 2. It is shown that the solution of the original problem is recovered by cooling the heated problem down.

Step 3. The violation properties of the original (nonheated) problem are determined from the violation properties of the heated problem by a limiting process.

Step 1 (*heating*). Let $\Delta' := \Delta \times \mathcal{B}_\rho$, where $\rho > 0$ is the *heating parameter* and $\mathcal{B}_\rho \subset \mathbb{R}^d$ is the closed ball centered in the origin with radius ρ, and let $\mathbb{P}' := \mathbb{P} \times \mathbb{U}$ be the probability in Δ' obtained as the product probability between $\mathbb{P}$ and the uniform probability $\mathbb{U}$ in $\mathcal{B}_\rho$. Each $z \in \mathcal{B}_\rho$ represents a constraint translation, and the heated uncertain program (HUP) is defined as

$$\begin{aligned} \mathrm{HUP}: \min_{x \in \mathcal{X} \subseteq \mathbb{R}^d} & c^T x \\ \text{subject to: } & x \in [\mathcal{X}_\delta + z], \quad (\delta, z) \in \Delta', \end{aligned}$$

where $[\mathcal{X}_\delta + z]$ is set $\mathcal{X}_\delta$ translated by z, and the new uncertain parameter (δ, z) allows for different selections of $\mathcal{X}_\delta$ constraints as well as for any translation z in $\mathcal{B}_\rho$. We show that HUP is fully-ball-supported.

To start with, consider a given deterministic ball $\mathcal{B}(x, r)$. We first prove that the strictly active constraints $\delta' \in \Delta'$ for $\mathcal{B}(x, r)$ form a set of zero-probability $\mathbb{P}'$, and later on from this we shall conclude that HUP is fully-ball-supported.

Let $\delta' = (\delta, z)$, and $\mathbb{I}_A$ indicate the indicator function of set A, and write

$$\begin{aligned} &\mathbb{P}'\{\delta' \text{ is strictly active for } \mathcal{B}(x,r)\} \\ &= \int_{\Delta'} \mathbb{I}_{\{\delta' \text{ is strictly active for } \mathcal{B}(x,r)\}} \mathbb{P}'(\mathrm{d}\delta') \\ &= [\text{by Fubini's theorem [23]}] \\ &= \int_{\Delta} \left[\int_{\mathcal{B}_\rho} \mathbb{I}_{\{(\delta,z) \text{ is strictly active for } \mathcal{B}(x,r)\}} \frac{\mathrm{d}z}{\mathrm{Vol}(\mathcal{B}_\rho)} \right] \mathbb{P}(\mathrm{d}\delta). \end{aligned} \tag{14}$$

The result that

$$\mathbb{P}'\{\delta' \text{ is strictly active for } \mathcal{B}(x,r)\} = 0 \tag{15}$$

is established by showing that the term within square brackets in (14) is null for all δ's.

Fix a δ, and let $C = \{z \in \mathcal{B}_\rho : \mathcal{B}(x,r) \subseteq [\mathcal{X}_\delta + z]\}$ be the set of translations not violating $\mathcal{B}(x,r)$. We show that C is convex and that the set $\{z \in \mathcal{B}_\rho : (\delta, z) \text{ is strictly active for } \mathcal{B}(x,r)\}$ belongs to ∂C, the boundary of C. Since the

boundary of a convex set has zero Lebesgue measure,[4] the desired result that the term within square brackets in (14) is null follows, viz.

$$\int_{\mathcal{B}_\rho} \mathbb{I}_{\{(\delta,z) \text{ is strictly active for } \mathcal{B}(x,r)\}} \frac{\mathrm{d}z}{\mathrm{Vol}(\mathcal{B}_\rho)} = 0. \tag{16}$$

The convexity of C is immediate: let $z_1, z_2 \in C$, that is, $\mathcal{B}(x,r) \subseteq [\mathcal{X}_\delta + z_1]$ and $\mathcal{B}(x,r) \subseteq [\mathcal{X}_\delta + z_2]$, or, equivalently, $\mathcal{B}(x,r) - z_1 \subseteq \mathcal{X}_\delta$ and $\mathcal{B}(x,r) - z_2 \subseteq \mathcal{X}_\delta$. From the convexity of $\mathcal{X}_\delta$, it follows that $\mathcal{B}(x,r) - \alpha z_1 - (1-\alpha) z_2 \subseteq \mathcal{X}_\delta \ \forall \alpha \in [0,1]$; that is, $\alpha z_1 + (1-\alpha) z_2 \in C$ and C is convex.

Consider now an interior point z of C (if any); i.e., there exists a ball centered in z all contained in C. This means that $[\mathcal{X}_\delta + z]$ can be moved around in all directions by a small quantity, and $\mathcal{B}(x,r)$ remains contained in it. It easily follows that (δ, z) cannot be strictly active, and, thus, $\{z \in \mathcal{B}_\rho : (\delta, z) \text{ is strictly active for } \mathcal{B}(x,r)\}$ has to belong to ∂C.

Wrapping up, (16) is established and, substituting in (14), (15) is obtained.

We next prove that (15) entails the fact that HUP is fully-ball-supported.

Consider a finite instance HP_m of HUP with $m \geq d$. One by one, eliminate $m-d$ constraints choosing at any time a constraint among those nonviolated by the ball-solution in such a way that the ball-solution does not change. This is certainly possible because the ball-support constraints are at most d. In the end, we are left with d constraints, say, the first d $\delta'^{(1)}, \ldots, \delta'^{(d)}$. A simple thought reveals that these d constraints are actually of ball-support for HP_m, provided that none of the other $m-d$ constraints that have been removed were strictly active.

Repeat the same above procedure for every m-ple of constraints (that is, for every HP_m generated by HUP), and group together all of the m-ples for which the procedure returns in the end the first d constraints $\delta'^{(1)}, \ldots, \delta'^{(d)}$. Call this group of m-ples G. We shall show that the probability of the m-ples in G such that HP_m is not fully-ball-supported is zero, and from this—by the observation that only a finite number $\binom{m}{d}$ of groups of m-ples can be similarly constructed—the final conclusion that HUP is fully-ball-supported will be secured.

Select fixed values $\bar{\delta}'^{(1)}, \ldots, \bar{\delta}'^{(d)}$ for the first d constraints, and consider the ball-solution $\mathcal{B}$ that these constraints generate. Let the other $m-d$ constraints vary in such a way that the m-ple $\bar{\delta}'^{(1)}, \ldots, \bar{\delta}'^{(d)}, \delta'^{(d+1)}, \ldots, \delta'^{(m)}$ belongs to G. For one such m-ple to correspond to a *non*-fully-ball-supported HP_m, at least one among the $m-d$ constraints $\delta'^{(d+1)}, \ldots, \delta'^{(m)}$ must be strictly active for $\mathcal{B}$, but we have proven in (15) that this happens with probability zero. Integrating over all possible values $\bar{\delta}'^{(1)}, \ldots, \bar{\delta}'^{(d)}$ for the first d constraints, the conclusion is drawn that the *non*-fully-ball-supported HP_m in G have zero probability.

Hence, by the above observation that there are only a finite number $\binom{m}{d}$ of groups and by the fact that $\binom{m}{d}$ times zero is zero, we obtain that HUP is fully-ball-supported.

To conclude Step 1, note that if we suppose to extract N constraints $\delta'^{(1)}, \ldots, \delta'^{(N)}$ from Δ', according to probability $\mathbb{P}'$ and in an independent fashion, and we denote by x'^*_N the corresponding solution, the result of Part 2a can be invoked to establish

[4] This simple fact follows from the observation that a convex set C in $\mathbb{R}^d$ either belongs to a flat of dimension $d-1$—and therefore C has zero $\mathbb{R}^d$ Lebesgue measure—or admits an interior point $\bar{z}$, and every half-line from $\bar{z}$ crosses the boundary of C in only one point (see, e.g., Propositions 1.1.13 and 1.1.14 in [19]).

that

$$(\mathbb{P}')^N\{V'(x'^*_N) > \epsilon\} \leq \sum_{i=0}^{d-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}, \tag{17}$$

where $V'(x)$ is the probability of violation for the heated problem (i.e., $V'(x) = \mathbb{P}'\{(\delta, z) \in \Delta' : \ x \notin [\mathcal{X}_\delta + z]\}$). Equation (17) is the final result to which we wanted to arrive in this heating Step 1.

Step 2 (*cooling*). Fix a multiextraction $(\bar{\delta}^{(1)}, \ldots, \bar{\delta}^{(N)}) \in \Delta^N$, and consider x^*_N, the solution of the original optimization problem P_N with such constraints. We remark that in all of Step 2 the multiextraction $(\bar{\delta}^{(1)}, \ldots, \bar{\delta}^{(N)})$ is kept fixed and never changed throughout. Consider a closed ball $\mathcal{B}(x_f, r_f)$, $r_f > 0$, in the feasibility domain of P_N, which exists because the feasibility domain of P_N has a nonempty interior. Further, let $\rho_k \downarrow 0$ be a sequence of heating parameters monotonically decreasing to zero (*cooling of the heating parameter*) and such that $\rho_1 < \frac{r_f}{2}$. For all ρ_k, consider the heated versions of $(\bar{\delta}^{(1)}, \ldots, \bar{\delta}^{(N)})$, namely, $((\bar{\delta}^{(1)}, z_k^{(1)}), \ldots, (\bar{\delta}^{(N)}, z_k^{(N)}))$ where $z_k^{(1)}, \ldots, z_k^{(N)} \in \mathcal{B}_{\rho_k}$, and let $x'^*_N(z_k^{(1)}, \ldots, z_k^{(N)})$ be the solution of the heated optimization problem HP_N with heated constraints $(\bar{\delta}^{(1)}, z_k^{(1)}), \ldots, (\bar{\delta}^{(N)}, z_k^{(N)})$. The goal of Step 2 is to prove that

$$\sup_{z_k^{(1)}, \ldots, z_k^{(N)} \in \mathcal{B}_{\rho_k}} \left\| x'^*_N \left(z_k^{(1)}, \ldots, z_k^{(N)} \right) - x^*_N \right\| \longrightarrow 0 \quad \text{as } k \to \infty; \tag{18}$$

that is, the solution of the original problem is recovered by cooling the heated problem down.[5]

For brevity, from now on we omit the arguments $z_k^{(1)}, \ldots, z_k^{(N)}$ and write x'^*_N for $x'^*_N(z_k^{(1)}, \ldots, z_k^{(N)})$.

We first show that

$$\limsup_{k \to \infty} \sup_{z_k^{(1)}, \ldots, z_k^{(N)} \in \mathcal{B}_{\rho_k}} c^T x'^*_N \leq c^T x^*_N. \tag{19}$$

Following Figure 8, consider the convex hull $\mathrm{co}[\mathcal{B}(x_f, r_f) \cup x^*_N]$ generated by the feasibility ball $\mathcal{B}(x_f, r_f)$ and the solution x^*_N of the original problem with constraints $\bar{\delta}^{(1)}, \ldots, \bar{\delta}^{(N)}$. By convexity, $\mathrm{co}[\mathcal{B}(x_f, r_f) \cup x^*_N]$ is feasible for the original problem P_N. Construct the closed ball $\mathcal{B}(x_k, \rho_k) \subset \mathrm{co}[\mathcal{B}(x_f, r_f) \cup x^*_N]$ with radius ρ_k, whose center x_k is as close as possible to x^*_N and lies on the line segment connecting x_f with x^*_N (this ball exists since $\rho_1 < r_f$; the assumed stricter condition that $\rho_1 < \frac{r_f}{2}$ is required in the next construction). Clearly, $x_k \to x^*_N$ as $k \to \infty$. Since x_k is in the feasibility domain of P_N at a distance at least ρ_k from where $\bar{\delta}^{(1)}, \ldots, \bar{\delta}^{(N)}$ are violated, x_k is also in the feasibility domain of every heated problem HP_N with heating parameter ρ_k. Thus,

$$\limsup_{k \to \infty} \sup_{z_k^{(1)}, \ldots, z_k^{(N)} \in \mathcal{B}_{\rho_k}} c^T x'^*_N \leq \limsup_{k \to \infty} c^T x_k = c^T x^*_N;$$

that is, (19) holds.

[5] Although result (18) has an intuitive appeal, its proof is rather technical. The reader not interested in these technical details can jump to Step 3 from here without loss of continuity.

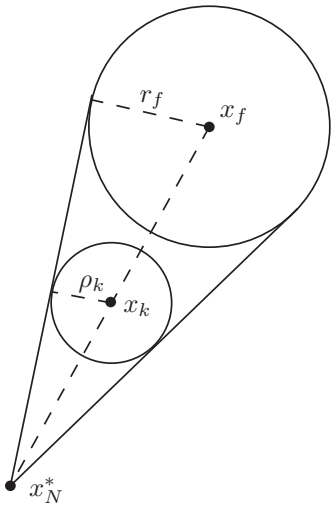


FIG. 8. *Convex hull of* $\mathcal{B}(x_f, r_f)$ *and* x_N^**, and construction of* $\mathcal{B}(x_k, \rho_k)$.

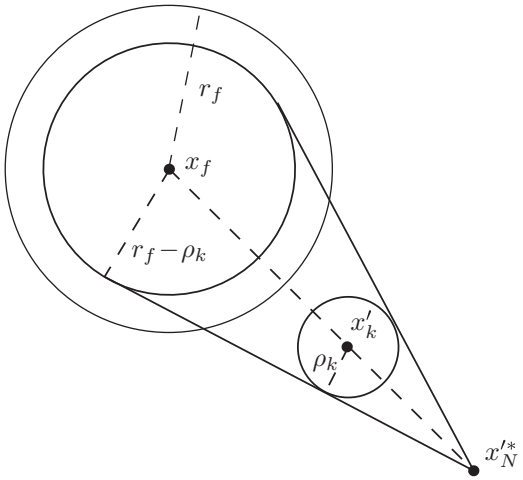


FIG. 9. *Convex hull of* $\mathcal{B}(x_f, r_f - \rho_k)$ *and* x'^*_N*, and construction of* $\mathcal{B}(x'_k, \rho_k)$.

Next, we construct a new convex hull which will allow us to reformulate goal (18) in a different, handier, way. Based on this reformulation, (18) will then be established in light of (19).

The new convex hull is $\mathrm{co}[\mathcal{B}(x_f, r_f - \rho_k) \cup x'^*_N]$; see Figure 9. Note that, for a given k, $\mathcal{B}(x_f, r_f - \rho_k)$ is a fixed ball, while instead x'^*_N depends on the specific choice of $z_k^{(1)}, \ldots, z_k^{(N)} \in \mathcal{B}_{\rho_k}$; this means that there are actually as many convex hulls as choices of $z_k^{(1)}, \ldots, z_k^{(N)}$. Moreover, $\mathrm{co}[\mathcal{B}(x_f, r_f - \rho_k) \cup x'^*_N]$ is feasible for problem HP_N with constraints translated by $z_k^{(1)}, \ldots, z_k^{(N)}$ since $\mathcal{B}(x_f, r_f - \rho_k)$ and x'^*_N are. Construct then the closed ball $\mathcal{B}(x'_k, \rho_k) \subseteq \mathrm{co}[\mathcal{B}(x_f, r_f - \rho_k) \cup x'^*_N]$ with radius ρ_k, whose center x'_k is as close as possible to x'^*_N and lies on the line segment connecting x_f with x'^*_N (this ball exists since $\rho_1 < \frac{r_f}{2}$). Note that x'_k depends on $z_k^{(1)}, \ldots, z_k^{(N)}$, too.

Since x'_k is in the feasibility domain of HP_N with constraints translated by $z_k^{(1)}, \ldots, z_k^{(N)}$ at a distance at least ρ_k from where these translated constraints are violated, x'_k is also in the feasibility domain of P_N.

What is different from the previous convex hull construction is that we cannot here easily conclude that $x'_k \to x'^*_N$ as $k \to \infty$ since x'^*_N is not a fixed point (it depends on $z_k^{(1)}, \ldots, z_k^{(N)} \in \mathcal{B}_{\rho_k}$, a ball that changes with k). We can still, however, secure a result that goes along a similar line, namely, that

$$x'_k = \alpha_k x_f + (1-\alpha_k) x'^*_N, \qquad \text{where } \alpha_k = \frac{\rho_k}{r_f - \rho_k} \longrightarrow 0 \text{ as } k \to \infty, \tag{20}$$

as it results from Figure 9 by a simple proportion argument.[6] Reorganizing terms in this equation, we obtain $x'^*_N - x^*_N = -\frac{\alpha_k}{1-\alpha_k}(x_f - x^*_N) + \frac{1}{1-\alpha_k}(x'_k - x^*_N)$, from which

$$\|x'^*_N - x^*_N\| \leq \frac{\alpha_k}{1-\alpha_k} \|x_f - x^*_N\| + \frac{1}{1-\alpha_k} \|x'_k - x^*_N\| .$$

We are now ready to reformulate goal (18) in a different way.

Note that the norm in (18) is the same as the left-hand side of the latter equation. On the right-hand side, $\|x_f - x^*_N\|$ is a fixed quantity multiplied by scalar $\frac{\alpha_k}{1-\alpha_k}$ which goes to zero. So, this first term vanishes. In the second term, scalar $\frac{1}{1-\alpha_k} \to 1$, and hence (18) is equivalent to

$$\sup_{z_k^{(1)}, \ldots, z_k^{(N)} \in \mathcal{B}_{\rho_k}} \|x'_k - x^*_N\| \longrightarrow 0 \quad \text{as } k \to \infty. \tag{21}$$

The goal of establishing (18) is finally achieved by proving (21) by contradiction.

Suppose that (21) is false; then, for a given $\mu > 0$, we can choose translations $\bar{z}_k^{(1)}, \ldots, \bar{z}_k^{(N)} \in \mathcal{B}_{\rho_k}$, $k = 1, 2, \ldots$, such that

$$\left\| x'_k \left(\bar{z}_k^{(1)}, \ldots, \bar{z}_k^{(N)} \right) - x^*_N \right\| > \mu \quad \forall k,$$

where we have here preferred to explicitly indicate dependence of x'_k on $\bar{z}_k^{(1)}, \ldots, \bar{z}_k^{(N)}$.

Note that $x'_k(\bar{z}_k^{(1)}, \ldots, \bar{z}_k^{(N)})$ is asymptotically superoptimal for problem P_N:

$$\begin{aligned}
&\limsup_{k\to\infty} c^T x'_k \left(\bar{z}_k^{(1)}, \ldots, \bar{z}_k^{(N)} \right) \\
&\leq [\text{using (20) and since } \alpha_k \to 0] \\
&\leq \limsup_{k\to\infty} \sup_{z_k^{(1)}, \ldots, z_k^{(N)}} c^T x'^*_N \\
&\leq [\text{using (19)}] \\
&\leq c^T x^*_N. \qquad (22)
\end{aligned}$$

The line segment connecting $x'_k(\bar{z}_k^{(1)}, \ldots, \bar{z}_k^{(N)})$ with x^*_N intersects the surface of the ball with center x^*_N and radius μ in a point that we name x'^S_k. x'^S_k is still feasible for P_N being a convex combination of x^*_N and $x'_k(\bar{z}_k^{(1)}, \ldots, \bar{z}_k^{(N)})$, both feasible points for P_N. In addition, since $x'_k(\bar{z}_k^{(1)}, \ldots, \bar{z}_k^{(N)})$ is asymptotically superoptimal for P_N (see (22)) and x^*_N is the solution of P_N, x'^S_k is asymptotically superoptimal for P_N, too, i.e., $\limsup_{k\to\infty} c^T x'^S_k \leq c^T x^*_N$. Finally, since x'^S_k belongs to a compact, it admits a convergent subsequence to, say, x'^S_∞, a point which is still feasible for P_N due to

[6] Note that (20) does not imply that $x'_k \to x'^*_N$ since x'^*_N could in principle escape to infinity.

the fact that the feasibility domain of P_N is closed. x'^S_∞ would thus be feasible and superoptimal for P_N, so contradicting the uniqueness of the solution of P_N.

This concludes Step 2.

Step 3 (*drawing the conclusions*). The theorem statement that $\mathbb{P}^N\{V(x^*_N) > \epsilon\} \leq \sum_{i=0}^{d-1} \binom{N}{i}\epsilon^i(1-\epsilon)^{N-i}$ is established in this Step 3 along the following lines: by the convergence result (18) in Step 2, a bad multiextraction $(\bar\delta^{(1)}, \ldots, \bar\delta^{(N)})$ (i.e., one such that $V(x^*_N) > \epsilon$) is shown to generate bad heated multiextractions $((\bar\delta^{(1)}, z_k^{(1)}), \ldots, (\bar\delta^{(N)}, z_k^{(N)}))$ for k large enough; we thus have that the probability of bad multiextractions can be bounded by the probability of bad heated multiextractions; by then using the bound for the probability of bad heated multiextractions derived in Step 1, the thesis follows.

Fix a *bad* multiextraction $(\bar\delta^{(1)}, \ldots, \bar\delta^{(N)}) \in \Delta^N$, and consider x^*_N, the solution of the optimization problem P_N with constraints $\bar\delta^{(1)}, \ldots, \bar\delta^{(N)}$. For an additional constraint $\delta \in \Delta$ to be violated by x^*_N, x^*_N must belong to the complement of $\mathcal{X}_\delta$, i.e., $\mathcal{X}^c_\delta$. Since $\mathcal{X}^c_\delta$ is open, we then have the fact that there exists a small enough ball centered in x^*_N fully contained in $\mathcal{X}^c_\delta$. Thus,

$$\{\delta \in \Delta :\ x^*_N \notin \mathcal{X}_\delta\} = \bigcup_{n=1,2,\ldots} \{\delta \in \Delta :\ \mathcal{B}(x^*_N, 1/n) \subseteq \mathcal{X}^c_\delta\}, \tag{23}$$

and

$$\begin{aligned}
\epsilon &< [\text{since } \left(\bar\delta^{(1)}, \ldots, \bar\delta^{(N)}\right) \text{ is bad}] \\
&< V(x^*_N) \\
&= \mathbb{P}\{\delta \in \Delta :\ x^*_N \notin \mathcal{X}_\delta\} \\
&= [\text{using (23)}] \\
&= \mathbb{P}\Big\{ \bigcup_{n=1,2,\ldots} \{\delta \in \Delta :\ \mathcal{B}(x^*_N, 1/n) \subseteq \mathcal{X}^c_\delta\}\Big\} \\
&= \lim_{n\to\infty} \mathbb{P}\{\delta \in \Delta :\ \mathcal{B}(x^*_N, 1/n) \subseteq \mathcal{X}^c_\delta\},
\end{aligned}$$

from which there exists a $\bar n$ such that

$$\mathbb{P}\{\delta \in \Delta :\ \mathcal{B}(x^*_N, 1/\bar n) \subseteq \mathcal{X}^c_\delta\} > \epsilon. \tag{24}$$

Let us now heat the constraints $\bar\delta^{(1)}, \ldots, \bar\delta^{(N)}$ up by translation parameters $z_k^{(1)}, \ldots, z_k^{(N)} \in \mathcal{B}_{\rho_k}$ and ask the following question: is it true that the heated multiextraction $((\bar\delta^{(1)}, z_k^{(1)}), \ldots, (\bar\delta^{(N)}, z_k^{(N)}))$ is bad for HUP with heating parameter ρ_k? It turns out that the answer is positive for k large enough, a fact that is proven next.

Recall that x'^*_N is the solution with constraints $(\bar\delta^{(1)}, z_k^{(1)}), \ldots, (\bar\delta^{(N)}, z_k^{(N)})$, and define $d_k := \sup_{z_k^{(1)}, \ldots, z_k^{(N)} \in \mathcal{B}_{\rho_k}} \|x'^*_N - x^*_N\|$ which, by (18), goes to 0 as $k \to \infty$. Pick a $\bar k$ such that

$$d_k + \rho_k < 1/\bar n \quad \forall k \geq \bar k.$$

All heated solutions x'^*_N are apart from x^*_N by at most d_k, and all heated constraints $(\delta, z) \in \Delta \times \mathcal{B}_{\rho_k}$ are apart from the corresponding unheated constraint δ by at most ρ_k. Thus, if $k \geq \bar k$, all heated versions of a constraint δ in the set $\{\delta \in \Delta :\ \mathcal{B}(x^*_N, 1/\bar n) \subseteq \mathcal{X}^c_\delta\}$ on the left-hand side of (24) are violated by x'^*_N. That is,

$$\{\delta \in \Delta :\ \mathcal{B}(x^*_N, 1/\bar n) \subseteq \mathcal{X}^c_\delta\} \times \mathcal{B}_{\rho_k} \subseteq \{(\delta, z) \in \Delta \times \mathcal{B}_{\rho_k} :\ x'^*_N \notin [\mathcal{X}_\delta + z]\} \quad \forall k \geq \bar k. \tag{25}$$

Then, for any $z_k^{(1)}, \ldots, z_k^{(N)} \in \mathcal{B}_{\rho_k}$ and for any $k \geq \bar{k}$, we have

$$\begin{aligned} V'(x_N'^*) &= \mathbb{P}'\{(\delta, z) \in \Delta \times \mathcal{B}_{\rho_k} : \ x_N'^* \notin [\mathcal{X}_\delta + z]\} \\ &\geq [\text{using (25)}] \\ &\geq \mathbb{P}'\Big\{\{\delta \in \Delta : \ \mathcal{B}(x_N^*, 1/\bar{n}) \subseteq \mathcal{X}_\delta^c\} \times \mathcal{B}_{\rho_k}\Big\} \\ &= [\text{recalling that } \mathbb{P}' = \mathbb{P} \times \mathbb{U}] \\ &= \mathbb{P}\{\delta \in \Delta : \ \mathcal{B}(x_N^*, 1/\bar{n}) \subseteq \mathcal{X}_\delta^c\} \cdot \mathbb{U}\{\mathcal{B}_{\rho_k}\} \\ &> [\text{since } \mathbb{U}\{\mathcal{B}_{\rho_k}\} = 1 \text{ and using (24)}] \\ &> \epsilon, \end{aligned}$$

i.e., $((\bar{\delta}^{(1)}, z_k^{(1)}), \ldots, (\bar{\delta}^{(N)}, z_k^{(N)}))$ is bad for HUP with heating parameter ρ_k for any $z_k^{(1)}, \ldots, z_k^{(N)} \in \mathcal{B}_{\rho_k}$ when $k \geq \bar{k}$. In turn, this entails that

$$\int_{\mathcal{B}_{\rho_k}^N} \mathbb{I}_{\{V'(x_N'^*) > \epsilon\}} \frac{\mathrm{d}z^N}{\mathrm{Vol}(\mathcal{B}_{\rho_k}^N)} = 1 \quad \forall k \geq \bar{k}. \tag{26}$$

Finally,

$$\begin{aligned} &\sum_{i=0}^{d-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i} \\ &\quad \geq \quad [\text{using (17)}] \\ &\quad \geq \quad (\mathbb{P}')^N\{V'(x_N'^*) > \epsilon\} \\ &\quad = \quad \int_{\Delta^N} \left[\int_{\mathcal{B}_{\rho_k}^N} \mathbb{I}_{\{V'(x_N'^*) > \epsilon\}} \frac{\mathrm{d}z^N}{\mathrm{Vol}(\mathcal{B}_{\rho_k}^N)} \right] \mathbb{P}^N(\mathrm{d}\delta^N) \\ &\quad \geq \quad \int_{\{V(x_N^*) > \epsilon\}} \left[\int_{\mathcal{B}_{\rho_k}^N} \mathbb{I}_{\{V'(x_N'^*) > \epsilon\}} \frac{\mathrm{d}z^N}{\mathrm{Vol}(\mathcal{B}_{\rho_k}^N)} \right] \mathbb{P}^N(\mathrm{d}\delta^N) \\ &\overrightarrow{k \to \infty} [\text{recalling (26) and by the dominated convergence theorem [26]}] \\ &\overrightarrow{k \to \infty} \int_{\{V(x_N^*) > \epsilon\}} \mathbb{P}^N(\mathrm{d}\delta^N) \\ &\quad = \quad \mathbb{P}^N\{V(x_N^*) > \epsilon\}. \end{aligned}$$

This concludes the proof. □

Acknowledgments. The authors would like to thank Marco Dalai for a careful reading of an earlier version of this paper and an anonymous referee for many suggestions that helped improve the paper.

REFERENCES

[1] A. Ben-Tal and A. Nemirovski, *Robust truss topology design via semidefinite programming*, SIAM J. Optim., 7 (1997), pp. 991–1016.

[2] A. Ben-Tal and A. Nemirovski, *Robust convex optimization*, Math. Oper. Res., 23 (1998), pp. 769–805.

[3] A. Ben-Tal and A. Nemirovski, *Robust solutions of uncertain linear programs*, Oper. Res. Lett., 25 (1999), pp. 1–13.

[4] A. Ben-Tal, A. Nemirovski, and C. Roos, *Robust solutions of uncertain quadratic and conic-quadratic problems*, SIAM J. Optim., 13 (2002), pp. 535–560.

[5] G. Calafiore and M. C. Campi, *Uncertain convex programs: Randomized solutions and confidence levels*, Math. Program., 102 (2005), pp. 25–46.
[6] G. Calafiore and M. C. Campi, *The scenario approach to robust control design*, IEEE Trans. Automat. Control, 51 (2006), pp. 742–753.
[7] G. Calafiore and F. Dabbene, *A probabilistic framework for problems with real structured uncertainty in systems and control*, Automatica, 38 (2002), pp. 1265–1276.
[8] G. Calafiore, F. Dabbene, and R. Tempo, *Randomized algorithms for probabilistic robustness with real and complex structured uncertainty*, IEEE Trans. Automat. Control, 45 (2000), pp. 2218–2235.
[9] G. Calafiore and B. Polyak, *Stochastic algorithms for exact and approximate feasibility of robust LMIs*, IEEE Trans. Automat. Control, 46 (2001), pp. 1755–1759.
[10] A. Charnes and W. W. Cooper, *Chance constrained programming*, Management Sci., 6 (1959), pp. 73–79.
[11] D. P. de Farias and B. Van Roy, *On constraint sampling in the linear programming approach to approximate dynamic programming*, Math. Oper. Res., 29 (2004), pp. 462–478.
[12] D. Dentcheva, *Optimization models with probabilistic constraints*, in Probabilistic and Randomized Methods for Design under Uncertainty, G. Calafiore and F. Dabbene, eds., Springer, London, 2005, pp. 47–96.
[13] E. Erdogan and G. Iyengar, *Ambiguous chance constrained problems and robust optimization*, Math. Program. Ser. B, 107 (2006), pp. 37–61.
[14] L. El Ghaoui and G. Calafiore, *Robust filtering for discrete-time systems with bounded noise and parametric uncertainty*, IEEE Trans. Automat. Control, 46 (2001), pp. 1084–1089.
[15] L. El Ghaoui and H. Lebret, *Robust solutions to least-squares problems with uncertain data*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 1035–1064.
[16] L. El Ghaoui, F. Oustry, and H. Lebret, *Robust solutions to uncertain semidefinite programs*, SIAM J. Optim., 9 (1998), pp. 33–52.
[17] L. El Ghaoui and S. I. Niculescu, *Robust decision problems in engineering: A linear matrix inequality approach*, in Advances in Linear Matrix Inequality Methods in Control, L. El Ghaoui and S. I. Niculescu, eds., SIAM, Philadelphia, 1999.
[18] V. L. Levin, *Application of E. Helly's theorem to convex programming, problems of best approximation and related questions*, Sb. Math., 8 (1969), pp. 235–247.
[19] R. Lucchetti, *Convexity and Well-posed Problems*, CMS Books Math., Springer, New York, 2006.
[20] A. Nemirovski and A. Shapiro, *Scenario approximations of chance constraints*, in Probabilistic and Randomized Methods for Design under Uncertainty, G. Calafiore and F. Dabbene, eds., Springer, London, 2005, pp. 3–48.
[21] A. Prékopa, *Stochastic Programming*, Kluwer, Boston, 1995.
[22] A. Prékopa, *Probabilistic programming*, in Stochastic Programming, A. Ruszczyǹski and A. Shapiro, eds., Handbooks Oper. Res. Managment Sci. 10, Elsevier, London, UK, 2003, pp. 267–352.
[23] W. Rudin, *Real and Complex Analysis*, 3rd ed., McGraw–Hill, New York, 1987.
[24] A. Ruszczyǹski and A. Shapiro, *Stochasting programming models*, in Stochastic Programming, A. Ruszczyǹski and A. Shapiro, eds., Handbooks Oper. Res. Managment Sci. 10, Elsevier, London, UK, 2003, pp. 1–64.
[25] A. Ruszczyǹski and A. Shapiro, *Lectures on Stochastic Programming*, preprint, 2008.
[26] A. N. Shiryaev, *Probability*, Springer, New York, 1996.
[27] R. L. Smith, *Efficients Monte-Carlo procedures for generating points uniformly distributed over bounded regions*, Oper. Res., 32 (1984), pp. 1296–1308.

SIAM J. OPTIM.
Vol. 19, No. 3, pp. 1231–1249

GLOBAL CONVERGENCE OF FILTER METHODS FOR NONLINEAR PROGRAMMING*

ADEMIR A. RIBEIRO†, ELIZABETH W. KARAS†, AND CLÓVIS C. GONZAGA‡

Abstract. We present a general filter algorithm that allows a great deal of freedom in the step computation. Each iteration of the algorithm consists basically in computing a point which is not forbidden by the filter, from the current point. We prove its global convergence, assuming that the step must be efficient, in the sense that, near a feasible nonstationary point, the reduction of the objective function is "large." We show that this condition is reasonable, by presenting two classical ways of performing the step which satisfy it. In the first one, the step is obtained by the inexact restoration method of Martínez and Pilotta. In the second, the step is computed by sequential quadratic programming.

Key words. filter methods, nonlinear programming, global convergence

AMS subject classifications. 49M37, 65K05, 90C30

DOI. 10.1137/060672285

1. Introduction.

We shall study the nonlinear programming problem

$$(P) \qquad \begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_{\mathcal{E}}(x) = 0 \\ & f_{\mathcal{I}}(x) \leq 0, \end{array}$$

where the index sets $\mathcal{E}$ and $\mathcal{I}$ refer to the equality and inequality constraints, respectively. Let the cardinality of $\mathcal{E} \cup \mathcal{I}$ be m, and assume that the functions $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 0, 1, \ldots, m$, are continuously differentiable.

A nonlinear programming algorithm must deal with two different (and possibly conflicting) criteria, related respectively to optimality and to feasibility. Optimality is measured by the objective function f_0; feasibility is typically measured by penalization of constraint violation, for instance, by the function $h : \mathbb{R}^n \to \mathbb{R}_+$, given by

$$h(x) = \left\| f^+(x) \right\|, \tag{1.1}$$

where $\| \cdot \|$ is an arbitrary norm and $f^+ : \mathbb{R}^n \to \mathbb{R}^m$ is defined by

$$f_i^+(x) = \begin{cases} f_i(x) & \text{if } \ i \in \mathcal{E}, \\ \max\{0, f_i(x)\} & \text{if } \ i \in \mathcal{I}. \end{cases}$$

Both criteria must be optimized and the algorithm should follow a certain balance between them at every step of the iterative process. Several algorithms for nonlinear programming have been designed in which a merit function is a tool to guarantee global convergence [3, 11, 12, 20, 26].

As an alternative to merit function, Fletcher and Leyffer [7] introduced the so-called *filter* to globalize sequential quadratic programming type methods. Filter methods are based on the concept of dominance, borrowed from multicriteria optimization.

*Received by the editors October 13, 2006; accepted for publication (in revised form) June 11, 2008; published electronically November 19, 2008.
http://www.siam.org/journals/siopt/19-3/67228.html

†Department of Mathematics, Federal University of Paraná, Cx. Postal 19081, 81531-980, Curitiba, PR, Brazil (ademir.ribeiro@ufpr.br, ewkaras@ufpr.br).

‡Department of Mathematics, Federal University of Santa Catarina, Cx. Postal 5210, 88040-970, Florianópolis, SC, Brazil (clovis@mtm.ufsc.br). This author is supported by CNPq.

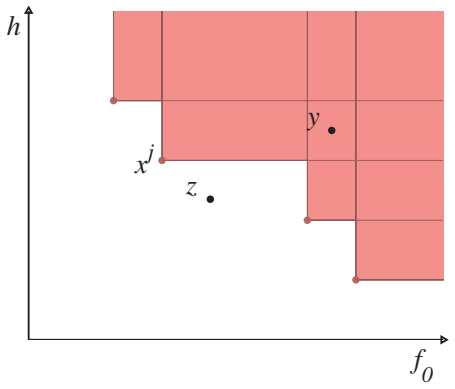


FIG. 1.1. *A filter with four pairs.*

A filter algorithm defines a *forbidden region*, by memorizing pairs $(f_0(x^j), h(x^j))$, chosen conveniently from former iterations and then avoids points dominated by these by the modified Pareto domination rule:

$$y \text{ is dominated by } x \text{ if, and only if, } f_0(y) \geq f_0(x) \text{ and } h(y) \geq h(x).$$

Figure 1.1 shows a filter with four pairs, where we have simplified the notation by using x to represent the pair $(f_0(x), h(x))$. The point y is in the forbidden region and z is not.

The filter methods were also applied to sequential linear programming (SLP). The works of Chin and Fletcher [5] and Fletcher, Leyffer, and Toint [8] present global convergence proofs of the method.

For SQP-filter methods, global convergence has been proved by Fletcher, Leyffer, and Toint [9], assuming that the quadratic subproblems are solved globally. Without this requirement, that is, allowing approximate solutions of the subproblems, Fletcher et al. [6] have also proved convergence to first-order critical points. Their approach uses a composite-step SQP method similar in spirit to the ones pioneered by Byrd [2] and Omojokun [22]. Another SQP-filter algorithm, using line search, was proposed by Wächter and Biegler [28], where global convergence was obtained.

In the context of interior points, Ulbrich, Ulbrich, and Vicente [24] have proposed a globally convergent primal-dual interior-point filter method. However, the filter entries have components that take into account the centrality and complementarity measures arising from interior-point techniques.

The filter was also studied by Gonzaga, Karas, and Vanti [13], in an algorithm that resembles the inexact restoration method of Martínez and Pilotta [19, 20]. By suitable rules for building the filter they prove stationarity of all qualified accumulation points.

The good performance of these methods [7, 29] has motivated their use in other problems, like nonlinear systems of equations [14, 16, 17], unconstrained optimization [15], and nonsmooth convex constrained optimization [18]. This last work, by Karas et al., combines the ideas of the proximal bundle methods [23] with the filter strategy. We can find a survey of filter methods in [10].

Although we know that filter methods may suffer from the Maratos effect, we shall not discuss local convergence issues in this work. Some strategies can be found in [4, 25, 27] to ensure a fast rate of convergence.

In this paper, we propose a general filter algorithm that does not depend on the particular method used for the step computation. The only requirement is that the points generated must be acceptable for the filter and that near a feasible nonstationary point, the reduction of the objective function be large. This efficiency condition, stated below as Hypothesis H3, is the main tool of the global convergence analysis. It is a weaker version of the one introduced by Gonzaga, Karas, and Vanti [13] in their inexact restoration filter method. Under this hypothesis, we prove that every sequence generated by the algorithm has at least one stationary accumulation point. Furthermore, we show how to compute the step in order to fulfill this hypothesis. One way to do this is by inexact restoration, for which H3 was proven in [13]. Another way for computing the step is by a sequential quadratic programming algorithm. We prove in this work that this approach also satisfies the efficiency condition H3.

The paper is organized as follows. Our general filter algorithm and its convergence analysis are described in section 2. In section 3 we present the SQP method for computing the step and prove that Hypothesis H3 is satisfied.

2. The algorithm. In this section we present a general filter algorithm that allows a great deal of freedom in the step computation. Afterwards we state an assumption on the performance of the step, and prove that any sequence generated by the algorithm has a stationary accumulation point. In the next section we show that this condition is reasonable, by presenting a classical way of performing the step, satisfying this condition.

The algorithm constructs a sequence of *filter* sets $F_0 \subset F_1 \subset \cdots \subset F_k$, composed of pairs $(\tilde{f}_0^j, \tilde{h}^j) \in \mathbb{R}^2$. We also mention in the algorithm the sets $\mathcal{F}_k \subset \mathbb{R}^n$, which are formally defined in each step for clarity, but are never actually constructed.

ALGORITHM 2.1. *General filter algorithm model*

Given: $x^0 \in \mathbb{R}^n$, $F_0 = \emptyset$, $\mathcal{F}_0 = \emptyset$, $\alpha \in (0,1)$.

$k = 0$

REPEAT

$(\tilde{f}_0, \tilde{h}) = (f_0(x^k) - \alpha h(x^k),\ (1-\alpha)h(x^k))$.

Set $\bar{F}_k = F_k \bigcup \{(\tilde{f}_0, \tilde{h})\}$ and define

$\bar{\mathcal{F}}_k = \mathcal{F}_k \bigcup \{x \in \mathbb{R}^n \mid f_0(x) \geq \tilde{f}_0, h(x) \geq \tilde{h}\}$.

Step:

IF x^k is stationary, stop with success

ELSE, compute $x^{k+1} \notin \bar{\mathcal{F}}_k$.

Filter update:

IF $f_0(x^{k+1}) < f_0(x^k)$,

$F_{k+1} = F_k, \quad \mathcal{F}_{k+1} = \mathcal{F}_k$ (f_0-iteration: the new entry is discarded)

ELSE,

$F_{k+1} = \bar{F}_k, \quad \mathcal{F}_{k+1} = \bar{\mathcal{F}}_k$ (h-iteration: the new entry becomes permanent)

$k = k + 1$.

At the beginning of each iteration, the pair $(\tilde{f}_0, \tilde{h})$ is temporarily introduced in the filter. After the complete iteration, this entry will become permanent in the filter only if the iteration *does not* produce a decrease in f_0.

Note that the forbidden region was slightly modified by subtracting the expression $\alpha h(x^k)$ from both filter pair components. This prevents the acceptance of trial pairs (f_0, h) arbitrarily close to old iterates $(f_0(x^j), h(x^j))$. Figure 2.1 illustrates the effect of this modification which adds a small margin around the border of region already

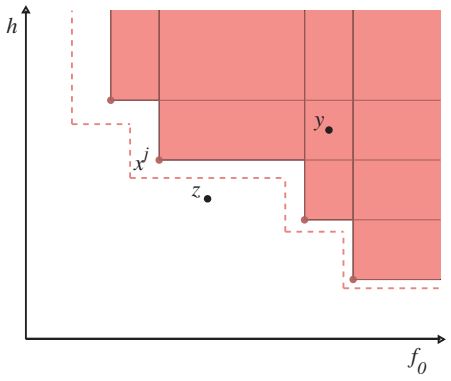


FIG. 2.1. *The filter margin.*

defined by the Pareto domination rule. Now, all points above the dashed line are forbidden.

In order to obtain our global convergence result, we shall assume that the algorithm generates an infinite sequence $(x^k)_{k\in\mathbb{N}}$ in $\mathbb{R}^n$ and the following hypotheses are satisfied.

H1. *All the functions $f_i(\cdot)$, $i = 0, 1, \ldots, m$, are twice continuously differentiable.*

H2. *The sequence $(x^k)_{k\in\mathbb{N}}$ remains in a convex compact domain $X \subset \mathbb{R}^n$.*

H 3. *Given a feasible nonstationary point $\bar{x} \in X$, there exist $M > 0$ and a neighborhood V of $\bar{x}$ such that for any iterate $x^k \in V$,*

$$f_0\left(x^k\right) - f_0\left(x^{k+1}\right) \geq M v_k,$$

where $v_k = \min\{1, \min\{\tilde{h}^j \mid (\tilde{f}_0^j, \tilde{h}^j) \in F_k\}\}$.

The first ones are standard hypotheses, while Hypothesis H3 requires some discussion. With this hypothesis we are assuming that the step must be efficient, in the sense that, near a feasible nonstationary point, the reduction of the objective function is "large." This assumption is a weaker version of the decrease condition introduced by Gonzaga, Karas, and Vanti [13] who used

$$H_k = \min\left\{1, \min\left\{\tilde{h}^j \mid \left(\tilde{f}_0^j, \tilde{h}^j\right) \in F_k, \tilde{f}_0^j \leq f_0\left(x^k\right)\right\}\right\} \tag{2.1}$$

instead of v_k. Figure 2.2 shows the variables v_k and H_k.

We start our analysis with some relations which follow directly from the hypotheses and the construction of the algorithm. In the next section we shall state methods which satisfy Hypothesis H3.

LEMMA 2.2. *Given $k \in \mathbb{N}$, the following statements hold:*

(i) $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ *(Inclusion property).*

(ii) $x^{k+p} \notin \mathcal{F}_{k+1}$ *for all* $p \geq 1$.

(iii) *At least one of the two conditions below occurs:*

1. $f_0(x^{k+1}) < f_0(x^k) - \alpha h(x^k)$,
2. $h(x^{k+1}) < (1-\alpha)h(x^k)$.

(iv) $\tilde{h} > 0$ *for all* $\left(\tilde{f}_0, \tilde{h}\right) \in F_k$.

Proof. The first statement follows directly from the filter update criterion. By (i) and the definition of $\bar{\mathcal{F}}$, we have $\mathcal{F}_{k+1} \subset \mathcal{F}_{k+p} \subset \bar{\mathcal{F}}_{k+p-1}$. Then the second statement follows from $x^{k+p} \notin \bar{\mathcal{F}}_{k+p-1}$. The third one is clear, since $x^{k+1} \notin \bar{\mathcal{F}}_k$. Finally, note

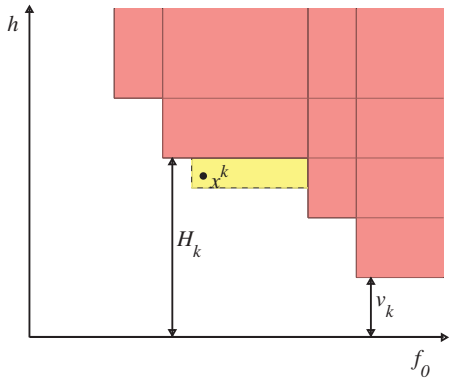


FIG. 2.2. *The filter height v_k and the filter slack H_k.*

that if $h(x^k) = 0$, using (iii), we obtain

$$f_0\left(x^{k+1}\right) < f_0\left(x^k\right) - \alpha h\left(x^k\right) = f_0\left(x^k\right),$$

that is, the iteration k is an f_0-iteration. Thus the pair $(\tilde{f}_0, \tilde{h})$ can be added to the filter only if $h(x^k) > 0$, or equivalently if $\tilde{h} > 0$. This completes the proof. ∎

For the purpose of our analysis, we shall consider

$$\mathcal{K}_a = \left\{k \in \mathbb{N} \mid \left(f_0\left(x^k\right) - \alpha h\left(x^k\right),\ (1-\alpha)h\left(x^k\right)\right) \text{ is added to the filter}\right\}, \tag{2.2}$$

the set of indices of h-iterations. First, we analyze what happens when this set is infinite.

LEMMA 2.3. *If the set $\mathcal{K}_a$ is infinite, then*

$$h\left(x^k\right) \overset{\mathcal{K}_a}{\to} 0.$$

Proof. Given k, we denote $\left(f_0(x^k), h(x^k)\right)$ by $\left(f_0^k, h^k\right)$. Assume by contradiction that, for some $\delta > 0$, the set

$$\mathcal{K} = \left\{k \in \mathcal{K}_a \mid h\left(x^k\right) \geq \delta\right\}$$

is infinite. The continuity of (f_0, h), implied by H1, and the compactness assumption H2 ensure that there exists a convergent subsequence $(f_0^k, h^k)_{k\in\mathcal{K}_1}$, $\mathcal{K}_1 \subset \mathcal{K}$. Therefore, since $\alpha \in (0,1)$, we can take indices $j, k \in \mathcal{K}_1$, $j < k$ such that

$$\left\|\left(f_0^k, h^k\right) - \left(f_0^j, h^j\right)\right\| < \alpha\delta \leq \alpha h(x^j).$$

This means that $x^k \in \bar{\mathcal{F}}_j = \mathcal{F}_{j+1}$, contradicting Lemma 2.2(ii) and completing the proof. ∎

We now prove that the objective function decreases along the iterations, whenever the iterates stay near a nonstationary point.

LEMMA 2.4. *Let $\bar{x} \in X$ be a nonstationary point. Then there exist $\bar{k} \in \mathbb{N}$ and a neighborhood V of $\bar{x}$ such that whenever $k > \bar{k}$ and $x^k \in V$, the iteration k is an f_0-iteration, that is, $k \notin \mathcal{K}_a$.*

Proof. If $\bar{x}$ is a feasible point, then by Hypothesis H3 there exist $M > 0$ and a neighborhood V of $\bar{x}$ such that for all $x^k \in V$,

$$f_0\left(x^k\right) - f_0\left(x^{k+1}\right) \geq M v_k.$$

Using Lemma 2.2(iv), we conclude that $v_k > 0$, consequently $f_0(x^{k+1}) < f_0(x^k)$ and k is an f_0-iteration.

Now, assume that $\bar{x}$ is infeasible and suppose by contradiction that there exists an infinite set $\mathcal{K} \subset \mathcal{K}_a$ such that $x^k \overset{\mathcal{K}}{\to} \bar{x}$. Since h is continuous, we have $h(x^k) \overset{\mathcal{K}}{\to} h(\bar{x})$. On the other hand, Lemma 2.3 ensures that $h(x^k) \overset{\mathcal{K}}{\to} 0$. Thus $h(\bar{x}) = 0$, contradicting that $\bar{x}$ is infeasible and completing the proof. ∎

Our global convergence result is presented in the next theorem.

THEOREM 2.5. *The sequence $(x^k)_{k\in\mathbb{N}}$ has a stationary accumulation point.*

Proof. Let $\mathcal{K}_a$ be the set defined in (2.2). If $\mathcal{K}_a$ is infinite, then by H2 there exist $\mathcal{K}_1 \subset \mathcal{K}_a$ and $\bar{x} \in X$ such that $x^k \overset{\mathcal{K}_1}{\to} \bar{x}$. From Lemma 2.4, $\bar{x}$ must be stationary.

On the other hand, if $\mathcal{K}_a$ is finite, there exists $k_0 \in \mathbb{N}$ such that every iteration $k \geq k_0$ is an f_0-iteration. Thus $(f_0(x^k))_{k\geq k_0}$ is decreasing and by H1 and H2,

$$f_0\left(x^k\right) - f_0\left(x^{k+1}\right) \to 0. \tag{2.3}$$

Moreover, by construction, $F_k = F_{k_0}$ for all $k \geq k_0$. Therefore, the sequence $(v_k)_{k\in\mathbb{N}}$, defined in Hypothesis H3, satisfies

$$v_k = v_{k_0} > 0 \tag{2.4}$$

for all $k \geq k_0$. If the set

$$\mathcal{K}_2 = \left\{k \in \mathbb{N} \mid \alpha h\left(x^k\right) < f_0\left(x^k\right) - f_0\left(x^{k+1}\right)\right\}$$

is infinite, using (2.3), we conclude that $h(x^k) \overset{\mathcal{K}_2}{\to} 0$. Otherwise, Lemma 2.2(iii) ensures that there exists $k_1 \in \mathbb{N}$ such that $h(x^{k+1}) < (1-\alpha)h(x^k)$ for all $k \geq k_1$, which in turn implies that $h(x^k) \to 0$. Anyway, $(x^k)_{k\in\mathbb{N}}$ has a feasible accumulation point $\bar{x}$. Now we prove that this point is stationary. Let $\mathcal{K}$ be a set of indices such that $x^k \overset{\mathcal{K}}{\to} \bar{x}$ and assume by contradiction that $\bar{x}$ is nonstationary. By Hypothesis H3, there exist $k_2 \in \mathbb{N}$ and $M > 0$ such that

$$f_0\left(x^k\right) - f_0\left(x^{k+1}\right) \geq M v_k$$

for all $k \in \mathcal{K}$, $k \geq k_2$. This together with (2.4) contradicts (2.3), completing the proof. ∎

As we have seen above, the hypothesis H3 is crucial for the convergence analysis. It is a very strong assumption and we must show that there exist methods satisfying this condition. One of them is the inexact restoration method of Martínez and Pilotta [20]. Gonzaga, Karas, and Vanti [13] have proved in their inexact restoration filter method a condition that implies our hypothesis.

We now discuss another way of performing the step, satisfying H3. It uses sequential quadratic programming and decomposes the step into its normal and tangential components.

3. Sequential quadratic programming. In this section we present an SQP method based on that proposed by Fletcher et al. [6], which computes the overall step in two phases. First, a feasibility phase aims at reducing the infeasibility measure h, satisfying a linear approximation of the constraints. Then an optimality phase computes a trial point reducing a quadratic model of the objective function in the linearization of the feasible set. We prove that this approach satisfies Hypothesis H3.

The step computation. Given the current iterate x^k and a trust-region radius $\Delta > 0$, we compute the step by solving the quadratic subproblem

$$(QP_k) \qquad \begin{array}{ll} \text{minimize} & m_k\left(x^k + d\right) \\ \text{subject to} & x^k + d \in \mathcal{L}\left(x^k\right) \\ & \|d\| \leq \Delta, \end{array}$$

where

$$m_k\left(x^k + d\right) = f_0\left(x^k\right) + \nabla f_0\left(x^k\right)^T d + \frac{1}{2} d^T B_k d, \tag{3.1}$$

with B_k symmetric, and

$$\mathcal{L}\left(x^k\right) = \left\{x^k + d \in \mathbb{R}^n \mid f_{\mathcal{E}}\left(x^k\right) + A_{\mathcal{E}}\left(x^k\right) d = 0,\ f_{\mathcal{I}}\left(x^k\right) + A_{\mathcal{I}}\left(x^k\right) d \leq 0\right\}. \tag{3.2}$$

The matrix B_k may be chosen as an approximation of the Hessian of some Lagrangian function or any other symmetric matrix, provided that the sequence (B_k) remains uniformly bounded. See the hypothesis H6 below.

The solution of (QP_k) yields a trial point $x^k + d_\Delta$ that will be evaluated by the filter. To be accepted as the new iterate, this point must not be forbidden.

In fact, we will see the step d_Δ as the sum of two components, a feasibility step n^k and a tangential (optimality) step t_Δ. We now discuss each one of these steps.

Feasibility step and compatibility of (QP_k). The feasibility step n^k must satisfy the constraints of (QP_k) and has the purpose of reducing the infeasibility measure h. This can be done, for example, by

$$n^k = P_{\mathcal{L}(x^k)}\left(x^k\right) - x^k,$$

where $P_{\mathcal{L}(x)}(\cdot)$ is the projection onto the set $\mathcal{L}(x)$. However, we do not use this particular choice, but we shall assume a certain efficiency in this phase, given by the following hypothesis.

H 4. *There exist constants $\delta_h > 0$ and $c_n > 0$ such that for all $k \geq 0$ with $h(x^k) \leq \delta_h$, a step n^k can be computed, satisfying*

$$\left\|n^k\right\| \leq c_n h\left(x^k\right).$$

This assumption means that the feasibility step must be reasonably scaled with respect to the constraints. In particular, $n^k = 0$ whenever x^k is feasible. This hypothesis is discussed by Martínez [19], who presents a feasibility algorithm which satisfies it under reasonable conditions, like some regularity of the constraints and the absence of a stationary point $\bar{x}$ for h, with $h(\bar{x}) \neq 0$.

The step n^k is only useful if it is not too close to the trust-region boundary because, otherwise, the tangential step is unlikely to produce a sufficient decrease in the model m_k. We say that the subproblem (QP_k) is *compatible* when $\mathcal{L}(x^k) \neq \emptyset$ and

$$\left\|n^k\right\| \leq \xi\Delta, \tag{3.3}$$

where $\xi \in (0, 1)$ is a constant.

In our analysis, we shall consider

$$z^k = x^k + n^k, \tag{3.4}$$

the point obtained in the feasibility phase. Note that, from (3.1) and (3.4), we have

$$m_k\left(z^k\right) = m_k\left(x^k + n^k\right) = f_0\left(x^k\right) + \nabla f_0\left(x^k\right)^T n^k + \frac{1}{2} {n^k}^T B_k n^k. \tag{3.5}$$

Tangential step. If the subproblem (QP_k) is compatible, we anticipate a satisfactory decrease in the model when performing a tangential step t_Δ, approximate solution of the quadratic problem

$$(TP_k) \qquad \begin{array}{ll} \text{minimize} & \left(\nabla f_0\left(x^k\right) + B_k n^k\right)^T t + \frac{1}{2} t^T B_k t \\ \text{subject to} & A_{\mathcal{E}}\left(x^k\right) t = 0 \\ & f_{\mathcal{I}}\left(x^k\right) + A_{\mathcal{I}}\left(x^k\right)\left(n^k + t\right) \leq 0 \\ & \left\|n^k + t\right\| \leq \Delta. \end{array}$$

This problem is equivalent to (QP_k) with $d = n^k + t$.

Given the current iterate x^k and a trust-region radius $\Delta > 0$, if (QP_k) is compatible, the trial point is

$$x^k + d_\Delta = z^k + t_\Delta,$$

where $z^k = x^k + n^k$ is the point which comes from the feasibility phase and t_Δ is the tangential step.

Restoration procedure. If the subproblem (QP_k) is not compatible, the algorithm calls a *restoration procedure*, whose aim is to obtain a point $x^{k+1} \notin \bar{\mathcal{F}}_k$ with $h(x^{k+1}) < h(x^k)$, where the function h is the infeasibility measure defined by (1.1). This can be done by taking steps of some algorithm for solving the nonsmooth problem

$$\begin{array}{ll} \text{minimize} & h(x) \\ x \in \mathbb{R}^n & \end{array}.$$

We can now summarize the above discussion in the following algorithm for the step computation. After stating the algorithm we shall make some comments about its features.

ALGORITHM 3.1. *Computation of* $x^{k+1} \notin \bar{\mathcal{F}}_k$

Data: $x^k \in \mathbb{R}^n$, the current filter $\bar{\mathcal{F}}_k$, $0 < \Delta_{\min} < \Delta_{\max}$, $\Delta \in [\Delta_{\min}, \Delta_{\max}]$ and $c_p, \xi, \eta, \gamma \in (0,1)$.

IF $\mathcal{L}(x^k) = \emptyset$,
  use the restoration procedure to obtain $x^{k+1} \notin \bar{\mathcal{F}}_k$.
ELSE
  compute a feasibility step n^k such that $x^k + n^k \in \mathcal{L}(x^k)$
  REPEAT (while the point x^{k+1} is not obtained)
    IF $\left\|n^k\right\| > \xi\Delta$,
      use the restoration procedure to obtain $x^{k+1} \notin \bar{\mathcal{F}}_k$.
      determine B_{k+1} symmetric
    ELSE
      compute the tangential step t_Δ as above and define $d_\Delta = n^k + t_\Delta$
      set $ared = f_0(x^k) - f_0(x^k + d_\Delta)$ and $pred = m_k(x^k) - m_k(x^k + d_\Delta)$

IF $\{x^k + d_\Delta \in \bar{\mathcal{F}}_k\}$ OR $\{pred \geq c_p(h(x^k))^2$ AND $ared < \eta\, pred\}$
  $\Delta = \gamma\Delta$
ELSE
  $x^{k+1} = x^k + d_\Delta$
  determine B_{k+1} symmetric
$\Delta_k = \Delta$

Algorithm 3.1 was inspired in the SQP-filter algorithm proposed by Fletcher et al. [6]. However, there exist some differences between them, which we now point out. The first one is that here the step computation is made separately from the main filter algorithm, presented in section 2. This simplifies the study of the step properties and leaves the convergence analysis of the main algorithm in a clean framework. Another difference is in the trust-region radius. Algorithm 3.1 starts with a radius $\Delta \in [\Delta_{\min}, \Delta_{\max}]$, where $\Delta_{\min}, \Delta_{\max} > 0$ are constants. This procedure is not used in [6], making the convergence proofs involved. To overcome some difficulties they impose a condition like

$$\left\|n^k\right\| \leq c\Delta^{1+\mu},$$

where $c > 0$ and $\mu \in (0,1)$, to accept the normal step and to proceed with the tangential step. In our algorithm, this condition is replaced by (3.3); that is,

$$\left\|n^k\right\| \leq \xi\Delta,$$

where $\xi \in (0,1)$ is a constant. This requirement is usual in the composite-step approaches that we are considering.

We mention that the choice of a minimum radius $\Delta_{\min}$ may cause practical disadvantages, like the rejection of many trial points before the progress of the algorithm. On the other hand, it simplifies the analysis and enhances the chance of taking a pure Newton step.

Remarks. At iteration k, we denote by d_Δ the trial step obtained with the trust-region radius $\Delta \geq \Delta_k$. The point x^{k+1} can be computed in two different ways: by means of a restoration procedure or by $x^{k+1} = x^k + d_{\Delta_k}$. We also have two possibilities for rejecting the trial step d_Δ:

$$x^k + d_\Delta \in \bar{\mathcal{F}}_k \tag{3.6}$$

or

$$pred \geq c_p \left(h\left(x^k\right)\right)^2 \quad \text{and} \quad ared < \eta\, pred. \tag{3.7}$$

In both cases the trust-region radius is reduced and a new step is computed. Thus, in order to accept the step d_Δ, it is not enough to pass the filter criterion. It also must ensure a sufficient decrease in the objective function whenever the predicted reduction is more significant than the constraint violation. In particular, if all iterates are feasible, the first inequality in (3.7) will be always true, because $n^k = 0$ in this case. Furthermore, if $ared \geq \eta\, pred$, then $x^k + d_\Delta \notin \bar{\mathcal{F}}_k$. So, the step acceptance criterion reduces to $ared \geq \eta\, pred$, and the algorithm may be viewed as a classical unconstrained trust-region method.

We now prove that Hypothesis H3 is satisfied if Algorithm 2.1 is applied to problem (P) and the step is obtained by Algorithm 3.1. For that, we shall introduce a

function used as a stationarity measure. Given $x, z \in X$ and the set $\mathcal{L}(x)$ defined in (3.2), we denote

$$d^c(x,z) = P_{\mathcal{L}(x)}\left(z - \nabla f_0(x)\right) - z \tag{3.8}$$

the *projected gradient direction* and the function $\varphi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, given by

$$\varphi(x,z) = \begin{cases} -\nabla f_0(x)^T \dfrac{d^c(x,z)}{\|d^c(x,z)\|} & \text{if } d^c(x,z) \neq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{3.9}$$

the stationarity measure. According to [13] we have, at a feasible point $\bar{x}$, that the KKT conditions are equivalent to $d^c(\bar{x},\bar{x}) = 0$. Furthermore, if $\bar{x}$ is nonstationary, then $\varphi(\bar{x},\bar{x}) > 0$.

The projected gradient direction given above is based on a direction introduced by Martínez and Svaiter [21] to define a new optimality condition, called *AGP property* (Approximate Gradient Projection), which implies, and is strictly stronger than, the Fritz–John optimality conditions. Unlike the KKT conditions, it is satisfied by local minimizers of nonlinear programming problems, independently of constraint qualifications.

Note. Let us give an interpretation for the direction $d^c(x,z)$ when $z \in \mathcal{L}(x)$ (which is the case in the algorithm). It is an approximation to

$$d_B(z) = P_{\mathcal{L}(x)}\left(z - \nabla f_0(z)\right) - z.$$

This is the projected Cauchy direction defined by Bertsekas [1] for the minimization of $f_0(\cdot)$ in $\mathcal{L}(x)$, and $d_B(z) = 0$ implies that z is stationary for this problem if $z \in \mathcal{L}(x)$. If, in addition, z is feasible for (P) it is also stationary for (P). The direction $d^c(x,z)$ may be a good descent direction for (P) if

$$\frac{\nabla f_0(x)}{\|\nabla f_0(x)\|} \approx \frac{\nabla f_0(z)}{\|\nabla f_0(z)\|},$$

but otherwise it may be meaningless (possibly null). If $d^c(x^k,z^k) \neq 0$, we consider $d_1^c = \frac{d^c(x^k,z^k)}{\|d^c(x^k,z^k)\|}$. To continue our analysis we define the *generalized Cauchy step* given by

$$t^c = \begin{cases} \underset{\lambda \geq 0}{\operatorname{argmin}} \left\{ m_k\left(z^k + \lambda d_1^c\right) \mid \left\| z^k + \lambda d_1^c - x^k \right\| \leq \Delta \right\} & \text{if } d^c\left(x^k, z^k\right) \neq 0, \\ 0 & \text{otherwise,} \end{cases}$$

and we assume the following hypotheses related to Algorithm 3.1.

H5. *If the subproblem (QP_k) is compatible, then the model decrease at the tangential step t_Δ satisfies*

$$m_k\left(z^k\right) - m_k\left(z^k + t_\Delta\right) \geq m_k\left(z^k\right) - m_k\left(z^k + t^c\right).$$

H 6. *The matrices B_k are uniformly bounded, that is, there exists a constant $\beta > 0$ such that $\|B_k\| \leq \beta$ for all $k \geq 0$.*

The assumption H5 says that the tangential step must be at least as good as the generalized Cauchy step t^c. We also consider a very standard condition on the Hessians B_k, stated in Hypothesis H6.

We start our task by evaluating the infeasibility measure before and after the trial step.

LEMMA 3.2. *Suppose that Hypotheses* H1 *and* H2 *hold. There exists a constant* $c_h > 0$ *such that for any* $x^k \in X$ *and* $\Delta > 0$ *so that a trial step* d_Δ *is obtained by Algorithm* 3.1*,*

$$h\left(x^k\right) \leq c_h \Delta \qquad \textit{and} \qquad h\left(x^k + d_\Delta\right) \leq c_h \Delta^2.$$

Proof. It follows from Hypotheses H1 and H2 that there exists a constant $c_h > 0$ such that

$$\|\nabla f_i(x)\| \leq c_h \qquad \text{and} \qquad \left\|\nabla^2 f_i(x)\right\| \leq c_h \tag{3.10}$$

for all $x \in X$ and $i = 1, \ldots, m$. Consider $x^k \in X$ and $\Delta > 0$ so that a trial step d_Δ is obtained by Algorithm 3.1. Thus the feasibility step n^k also was computed by the algorithm. Taking, without loss of generality, the norm $\|\cdot\|_\infty$ in (1.1) and using the fact that $x^k + n^k \in \mathcal{L}(x^k)$, we conclude that

$$h\left(x^k\right) = \left|f_i\left(x^k\right)\right| = \left|-\nabla f_i\left(x^k\right)^T n^k\right|$$

for some $i \in \mathcal{E}$, or

$$h\left(x^k\right) = f_i\left(x^k\right) \leq -\nabla f_i\left(x^k\right)^T n^k$$

for some $i \in \mathcal{I}$. Hence, from the Cauchy–Schwarz inequality, (3.10), and the trust-region boundedness of n^k, we obtain

$$h\left(x^k\right) \leq \left\|\nabla f_i\left(x^k\right)\right\| \left\|n^k\right\| \leq c_h \Delta,$$

proving the first claim in the lemma.

To prove the other inequality, note that by Taylor's theorem and the fact that $x^k + d_\Delta \in \mathcal{L}(x^k)$,

$$f_i\left(x^k + d_\Delta\right) = \frac{1}{2} d_\Delta^T \nabla^2 f_i\left(x^k + \theta_i d_\Delta\right) d_\Delta,$$

for $i \in \mathcal{E}$, and

$$f_i\left(x^k + d_\Delta\right) \leq \frac{1}{2} d_\Delta^T \nabla^2 f_i\left(x^k + \theta_i d_\Delta\right) d_\Delta,$$

for $i \in \mathcal{I}$, where $\theta_i \in (0, 1)$. Because the trust-region radius is bounded, we may assume without loss of generality that the trial points also remain in the compact set X. Thus, from (3.10), the Cauchy–Schwarz inequality and since $\|d_\Delta\| \leq \Delta$,

$$h\left(x^k + d_\Delta\right) \leq c_h \Delta^2,$$

completing the proof. ∎

We next assess the model and the objective function growth in the feasibility step computed by Algorithm 3.1.

LEMMA 3.3. *Suppose that Hypotheses* H1, H2, H4, *and* H6 *hold. Let* δ_h *be the constant given by* H4*. Given a feasible point* $\bar{x} \in X$*, there exist* $N > 0$ *and a neighborhood* V_1 *of* $\bar{x}$ *such that if* $x^k \in V_1$ *and* $z^k = x^k + n^k$*, then*

(i) $h(x^k) \leq \delta_h$.
(ii) $\left|m_k(x^k) - m_k(z^k)\right| \leq N h(x^k)$.
(iii) $\left|f_0(x^k) - f_0(z^k)\right| \leq N h(x^k)$.

In particular, by (i)*, the hypothesis* H4 *is applicable; that is, the step* n^k *can be computed and satisfies* $\|n^k\| \leq c_n h(x^k)$.

Proof. Let c_n and β be the constants given by H4 and H6, respectively. Consider the constant $c_g = \max\{\|\nabla f_0(x)\| \mid x \in X\}$, whose existence is ensured by H1 and H2. Since $h(\bar{x}) = 0$ and h is continuous, there exists a neighborhood V_1 of $\bar{x}$ such that if $x^k \in V_1$, then

$$h\left(x^k\right) \leq \delta_h \qquad \text{and} \qquad \frac{1}{2}\beta c_n^2 \left(h\left(x^k\right)\right)^2 \leq c_g c_n h\left(x^k\right). \tag{3.11}$$

The first inequality in (3.11) proves (i). By (3.5), we have

$$m_k\left(z^k\right) = m_k\left(x^k\right) + \nabla f_0\left(x^k\right)^T n^k + \frac{1}{2} n^{kT} B_k n^k.$$

Using the Cauchy–Schwarz inequality, H4 and H6, we obtain

$$\begin{aligned} |m_k\left(x^k\right) - m_k\left(z^k\right)| &\leq c_g \left\|n^k\right\| + \tfrac{1}{2}\beta \left\|n^k\right\|^2 \\ &\leq c_g c_n h\left(x^k\right) + \tfrac{1}{2}\beta c_n^2 \left(h\left(x^k\right)\right)^2. \end{aligned}$$

From the second inequality in (3.11), it follows that

$$|m_k\left(x^k\right) - m_k\left(z^k\right)| \leq 2 c_g c_n h\left(x^k\right).$$

On the other hand, by Hypotheses H1 and H2, there exists a constant $L > 0$ such that

$$\left|f_0\left(x^k\right) - f_0\left(z^k\right)\right| \leq L \left\|x^k - z^k\right\|.$$

This together with H4 yields

$$\left|f_0\left(x^k\right) - f_0\left(z^k\right)\right| \leq L c_n h\left(x^k\right).$$

Taking $N = \max\{2c_g c_n, L c_n\}$, we complete the proof. ∎

Now we prove that the model and the objective function reductions are large near a feasible nonstationary point, if we ignore the filter. The first lemma looks only at the tangential step, while the second lemma considers the whole step (feasibility and tangential).

LEMMA 3.4. *Suppose that Hypotheses* H1, H2, and H4–H6 *hold. Let* $\bar{x} \in X$ *be a feasible nonstationary point and* $\bar{\eta} \in (0,1)$*. Consider the neighborhood* V_1 *and the constant* $\Delta_{\min}$ *given by Lemma* 3.3 *and Algorithm* 3.1*, respectively. Then there exist constants* $\Delta_\rho \in (0, \Delta_{\min}]$*,* $\tilde{c} > 0$*, and a neighborhood* $V_2 \subset V_1$ *of* $\bar{x}$ *such that whenever* $x^k \in V_2$*,* $z^k = x^k + n^k$*, and a tangential trial step* t_Δ *is obtained by the algorithm, we have*

(i) $m_k(z^k) - m_k(z^k + t_\Delta) \geq \tilde{c}\Delta'$ *for all* Δ, Δ' *such that* $\Delta' \leq \min\{\Delta, \Delta_\rho\}$.

(ii) $f_0(z^k) - f_0(z^k + t_\Delta) \geq \bar{\eta}\left(m_k(z^k) - m_k(z^k + t_\Delta)\right)$ *for all* $\Delta \in (0, \Delta_\rho]$.

Proof. Let $\Delta > 0$ and $\lambda_{\Delta'} = (1 - \xi)\Delta'$, where ξ is given by (3.3) and $\Delta' \leq \Delta$. First, note that the vector d_1^c, defined before H5, satisfies $\|d_1^c\| = 1$. Consequently,

$$\left\|z^k + \lambda_{\Delta'} d_1^c - x^k\right\| = \left\|n^k + \lambda_{\Delta'} d_1^c\right\| \leq \left\|n^k\right\| + \lambda_{\Delta'} \leq \xi\Delta + (1 - \xi)\Delta' \leq \Delta.$$

Using the assumption on the Cauchy point H5, we obtain

$$m_k\left(z^k\right) - m_k\left(z^k + t_\Delta\right) \geq m_k\left(z^k\right) - m_k\left(z^k + t^c\right) \geq m_k\left(z^k\right) - m_k\left(z^k + \lambda_{\Delta'} d_1^c\right).$$

Developing the quadratic model (3.1) in the right-hand side, we conclude that

$$m_k\left(z^k\right) - m_k\left(z^k + t_\Delta\right) \geq \lambda_{\Delta'}\left(-\nabla f_0\left(x^k\right)^T d_1^c - {n^k}^T B_k d_1^c - \frac{1}{2}\lambda_{\Delta'} {d_1^c}^T B_k d_1^c\right).$$

By (3.9), $\varphi(x^k, z^k) = -\nabla f_0(x^k)^T d_1^c$ and by H6, $\|B_k\| \leq \beta$. Hence

$$m_k\left(z^k\right) - m_k\left(z^k + t_\Delta\right) \geq \lambda_{\Delta'}\left(\varphi\left(x^k, z^k\right) - \left\|n^k\right\|\beta - \frac{1}{2}\lambda_{\Delta'}\beta\right). \tag{3.12}$$

Since $\bar{x}$ is feasible nonstationary, the continuous function φ satisfies $\varphi(\bar{x}, \bar{x}) > 0$. Using the fact that $\|n^k\| \leq c_n h(x^k)$ by H4, we conclude that there exist a neighborhood V_2 of $\bar{x}$ and $\Delta_0 \in (0, \Delta_{\min}]$ such that for any $x^k \in V_2$ and $\Delta' \in (0, \Delta_0]$,

$$\varphi\left(x^k, z^k\right) \geq \frac{1}{2}\varphi\left(\bar{x}, \bar{x}\right) \qquad \text{and} \qquad \left\|n^k\right\|\beta + \frac{1}{2}\lambda_{\Delta'}\beta \leq \frac{1}{4}\varphi\left(\bar{x}, \bar{x}\right).$$

Thus, by (3.12), we obtain for $\Delta' \leq \min\{\Delta, \Delta_0\}$,

$$m_k\left(z^k\right) - m_k\left(z^k + t_\Delta\right) \geq \frac{1}{4}\lambda_{\Delta'}\varphi(\bar{x}, \bar{x}) = \frac{1}{4}\left(1 - \xi\right)\varphi\left(\bar{x}, \bar{x}\right)\Delta'.$$

This proves (i) for any $\Delta_\rho \leq \Delta_0$ and $\tilde{c} = \frac{1}{4}(1-\xi)\varphi(\bar{x}, \bar{x})$.

To prove (ii), note that by the mean value theorem,

$$ared_{z^k} \stackrel{\text{def}}{=} f_0\left(z^k\right) - f_0\left(z^k + t_\Delta\right) = -\nabla f_0\left(z^k + \theta t_\Delta\right)^T t_\Delta$$

for some $\theta \in (0, 1)$. On the other hand,

$$pred_{z^k} \stackrel{\text{def}}{=} m_k\left(z^k\right) - m_k\left(z^k + t_\Delta\right) = -\nabla f_0\left(x^k\right)^T t_\Delta - t_\Delta^T B_k n^k - \frac{1}{2} t_\Delta^T B_k t_\Delta.$$

By H1 and H2, we can apply the mean value inequality to ∇f_0 to conclude that there exists a constant $L > 0$ such that

$$\left\|\nabla f_0\left(x^k\right) - \nabla f_0\left(z^k + \theta t_\Delta\right)\right\| \leq L\left\|z^k - x^k + \theta t_\Delta\right\|,$$

so, using the facts that $\|B_k\| \leq \beta$ and $\|t_\Delta\| \leq \Delta$, we obtain

$$\begin{aligned} |ared_{z^k} - pred_{z^k}| &\leq L\left\|z^k - x^k + \theta t_\Delta\right\|\|t_\Delta\| + \beta\left\|n^k\right\|\|t_\Delta\| + \tfrac{1}{2}\beta\|t_\Delta\|^2 \\ &\leq L\left\|n^k\right\|\Delta + L\Delta^2 + \beta\left\|n^k\right\|\Delta + \tfrac{1}{2}\beta\Delta^2 \\ &= (L+\beta)\left\|n^k\right\|\Delta + \left(L + \tfrac{1}{2}\beta\right)\Delta^2. \end{aligned}$$

We can restrict the neighborhood V_2, if necessary, and take $\Delta_\rho \leq \Delta_0$ such that for any $x^k \in V_2$ and $\Delta \in (0, \Delta_\rho]$,

$$\frac{(L+\beta)\left\|n^k\right\|}{\tilde{c}} \leq \frac{1-\bar{\eta}}{2} \qquad \text{and} \qquad \frac{\left(L + \frac{1}{2}\beta\right)\Delta}{\tilde{c}} \leq \frac{1-\bar{\eta}}{2}.$$

Consequently, using (i) with $\Delta' = \Delta$,

$$\left|\frac{ared_{z^k}}{pred_{z^k}} - 1\right| = \left|\frac{ared_{z^k} - pred_{z^k}}{pred_{z^k}}\right| \leq \frac{(L+\beta)\left\|n^k\right\|\Delta + \left(L + \frac{1}{2}\beta\right)\Delta^2}{\tilde{c}\Delta} \leq 1 - \bar{\eta},$$

completing the proof. □

In the next lemma we extend for the whole step the properties of the tangential step near a feasible nonstationary point.

LEMMA 3.5. *Suppose that Hypotheses* H1, H2, and H4–H6 *hold. Let* $\bar{x} \in X$ *be a feasible nonstationary point and* $0 < \eta < 1$. *Consider the constant* γ *given in Algorithm* 3.1, *the neighborhood* V_2 *and the constant* Δ_ρ *given in Lemma* 3.4. *Then there exists a neighborhood* $V_3 \subset V_2$ *of* $\bar{x}$ *such that whenever* $x^k \in V_3$, $z^k = x^k + n^k$, *and a tangential trial step* t_Δ *is obtained by the algorithm, we have for all* $\Delta \in [\gamma^2\Delta_\rho, \Delta_\rho]$,

(i) $m_k(x^k) - m_k(z^k + t_\Delta) \geq \frac{1}{2}\tilde{c}\Delta$,

(ii) $f_0(x^k) - f_0(z^k + t_\Delta) \geq \eta(m_k(x^k) - m_k(z^k + t_\Delta))$.

Proof. Let $\bar{\eta} \in (\eta, 1)$ and $\tau = \frac{\bar{\eta}-\eta}{\bar{\eta}+\eta}$. Consider the constants N and $\tilde{c}$ given by Lemmas 3.3 and 3.4, respectively, and $V_3 \subset V_2$ a neighborhood of $\bar{x}$ such that for all $x \in V_3$,

$$Nh(x) \leq \min\left\{\frac{1}{2}\tilde{c}\gamma^2\Delta_\rho, \tau\bar{\eta}\tilde{c}\gamma^2\Delta_\rho\right\}. \tag{3.13}$$

Hence, if $x^k \in V_3$ and $\Delta \in [\gamma^2\Delta_\rho, \Delta_\rho]$, we can apply Lemma 3.3 to conclude that

$$\left|m_k\left(x^k\right) - m_k\left(z^k\right)\right| \leq Nh\left(x^k\right) \leq \frac{1}{2}\tilde{c}\gamma^2\Delta_\rho \leq \frac{1}{2}\tilde{c}\Delta.$$

It follows from this and Lemma 3.4(i) with $\Delta' = \Delta$ that

$$m_k\left(x^k\right) - m_k\left(z^k + t_\Delta\right) = m_k\left(x^k\right) - m_k\left(z^k\right) + m_k\left(z^k\right) - m_k\left(z^k + t_\Delta\right) \geq \frac{1}{2}\tilde{c}\Delta,$$

proving (i).

(ii) Applying again Lemmas 3.3 and 3.4 together with (3.13), we obtain

$$\left|f_0\left(x^k\right) - f_0\left(z^k\right)\right| \leq Nh\left(x^k\right) \leq \tau\bar{\eta}\tilde{c}\gamma^2\Delta_\rho \leq \tau\bar{\eta}\tilde{c}\Delta \leq \tau\left(f_0\left(z^k\right) - f_0\left(z^k + t_\Delta\right)\right)$$

and

$$m_k\left(x^k\right) - m_k\left(z^k\right) \leq Nh\left(x^k\right) \leq \tau\tilde{c}\gamma^2\Delta_\rho \leq \tau\tilde{c}\Delta \leq \tau\left(m_k\left(z^k\right) - m_k\left(z^k + t_\Delta\right)\right).$$

Consequently,

$$\begin{aligned} f_0\left(x^k\right) - f_0\left(z^k + t_\Delta\right) &= f_0\left(x^k\right) - f_0\left(z^k\right) + f_0\left(z^k\right) - f_0\left(z^k + t_\Delta\right) \\ &\geq (1-\tau)\left(f_0\left(z^k\right) - f_0\left(z^k + t_\Delta\right)\right) \end{aligned} \tag{3.14}$$

and

$$\begin{aligned} m_k\left(x^k\right) - m_k\left(z^k + t_\Delta\right) &= m_k\left(x^k\right) - m_k\left(z^k\right) + m_k\left(z^k\right) - m_k\left(z^k + t_\Delta\right) \\ &\leq (1+\tau)\left(m_k\left(z^k\right) - m_k\left(z^k + t_\Delta\right)\right). \end{aligned} \tag{3.15}$$

Therefore, if $x^k \in V_3$ and $\Delta \in [\gamma^2\Delta_\rho, \Delta_\rho]$, using (3.14), (3.15), and Lemma 3.4(ii), we obtain

$$\begin{aligned} f_0\left(x^k\right) - f_0\left(z^k + t_\Delta\right) &\geq (1-\tau)\bar{\eta}\left(m_k\left(z^k\right) - m_k\left(z^k + t_\Delta\right)\right) \\ &\geq \frac{(1-\tau)\bar{\eta}}{(1+\tau)}\left(m_k\left(x^k\right) - m_k\left(z^k + t_\Delta\right)\right) \\ &= \eta\left(m_k\left(x^k\right) - m_k\left(z^k + t_\Delta\right)\right), \end{aligned}$$

completing the proof. ∎

In the next two results we shall use the *filter slack* H_k, defined in (2.1). First we show that, near a feasible nonstationary point, the rejection of a step is due to a large increase of the infeasibility.

LEMMA 3.6. *Suppose that Hypotheses* H1, H2, and H4–H6 *hold. Let* $\bar{x} \in X$ *be a feasible nonstationary point and consider the constants* γ *and* Δ_ρ *given by Algorithm* 3.1 *and Lemma* 3.4, *respectively, and the neighborhood* V_3 *given by Lemma* 3.5. *Then there exists a neighborhood* $V \subset V_3$ *of* $\bar{x}$ *such that whenever* $x^k \in V$, $z^k = x^k + n^k$, *and a tangential trial step* t_Δ *is obtained by the algorithm, we have*

$$h\left(z^k + t_\Delta\right) \geq H_k$$

for any $\Delta \in [\gamma^2 \Delta_\rho, \Delta_\rho]$ *that was rejected by Algorithm* 3.1.

Proof. Let α, η, N, and $\tilde{c}$ be the constants given by Algorithms 2.1, 3.1 and Lemmas 3.3, 3.4, respectively. Consider $V \subset V_3$ a neighborhood of $\bar{x}$ such that for all $x \in V$,

$$Nh(x) \leq \frac{1}{2}\tilde{c}\gamma^2\Delta_\rho \qquad \text{and} \qquad \alpha h(x) \leq \frac{1}{2}\eta\tilde{c}\gamma^2\Delta_\rho. \tag{3.16}$$

Hence, if $x^k \in V$ and $\Delta \in [\gamma^2\Delta_\rho, \Delta_\rho]$, we can apply Lemma 3.3 to obtain

$$\left|m_k\left(x^k\right) - m_k\left(z^k\right)\right| \leq Nh\left(x^k\right) \leq \frac{1}{2}\tilde{c}\gamma^2\Delta_\rho \leq \frac{1}{2}\tilde{c}\Delta,$$

which together with Lemma 3.4 yields

$$m_k\left(x^k\right) - m_k\left(z^k + t_\Delta\right) \geq \frac{1}{2}\tilde{c}\Delta \geq \frac{1}{2}\tilde{c}\gamma^2\Delta_\rho. \tag{3.17}$$

Using Lemma 3.5, (3.16), and (3.17), we obtain

$$\begin{aligned} f_0\left(x^k\right) - f_0\left(z^k + t_\Delta\right) &\geq \eta\left(m_k\left(x^k\right) - m_k\left(z^k + t_\Delta\right)\right) \\ &\geq \frac{1}{2}\eta\tilde{c}\gamma^2\Delta_\rho \\ &\geq \alpha h\left(x^k\right). \end{aligned}$$

Since $z^k + t_\Delta = x^k + d_\Delta$, it follows that

$$f_0\left(x^k\right) - f_0\left(x^k + d_\Delta\right) \geq \eta\left(m_k\left(x^k\right) - m_k\left(x^k + d_\Delta\right)\right) \tag{3.18}$$

and

$$f_0\left(x^k + d_\Delta\right) \leq f_0\left(x^k\right) - \alpha h\left(x^k\right). \tag{3.19}$$

Therefore, if the trial step d_Δ was rejected by Algorithm 3.1, then $x^k + d_\Delta \in \bar{\mathcal{F}}_k$ because of (3.18). We thus conclude from (3.19) that

$$h\left(z^k + t_\Delta\right) \geq H_k,$$

completing the proof. ∎

We now prove the main result of this section: *Hypothesis* H3 *is satisfied.* Indeed, we give a sufficient condition to ensure H3. As we saw in Theorem 2.5, this hypothesis was crucial in the convergence analysis of section 2.

For the purpose of our analysis, we shall consider the set of restoration iterations

$$\mathcal{K}_r = \left\{k \in \mathbb{N} \mid \mathcal{L}\left(x^k\right) = \emptyset \text{ or } \left\|n^k\right\| > \xi\Delta_k\right\}, \tag{3.20}$$

where $\mathcal{L}(x^k)$ is defined by (3.2). We also assume the following hypothesis.

H7. *Every feasible accumulation point* $\bar{x} \in X$ *of* $(x^k)_{k\in\mathbb{N}}$ *satisfies the Mangasarian–Fromovitz constraint qualification; namely, the gradients* $\nabla f_i(\bar{x})$ *for* $i \in \mathcal{E}$ *are linearly independent, and there exists a direction* $d \in \mathbb{R}^n$ *such that* $A_{\mathcal{E}}(\bar{x})d = 0$ *and* $A_{\bar{\mathcal{I}}}(\bar{x})d < 0$, *where* $\bar{\mathcal{I}} = \{i \in \mathcal{I} \mid f_i(\bar{x}) = 0\}$.

THEOREM 3.7. *Suppose that Algorithm* 2.1 *is applied to problem* (P)*, with the step computed by Algorithm* 3.1*, and that Hypotheses* H1, H2, and H4–H7 *hold. Given a feasible nonstationary point* $\bar{x} \in X$*, there exist* $M > 0$ *and a neighborhood* V *of* $\bar{x}$ *such that if* $x^k \in V$*, then*

$$f_0\left(x^k\right) - f_0\left(x^{k+1}\right) \geq M\sqrt{H_k}.$$

In particular, since $\sqrt{H_k} \geq v_k$*, the hypothesis* H3 *is satisfied.*

Proof. Let $\bar{x}$ be a feasible nonstationary point. Consider the neighborhood V given by Lemma 3.6 and the constant Δ_ρ given by Lemma 3.4. Without loss of generality, we can assume that

$$\Delta_\rho \leq \frac{\gamma^2}{c_h} \min\left\{\frac{\xi}{c_n}, \frac{\tilde{c}}{2N}, \frac{\tilde{c}}{2c_p}, \frac{\eta\tilde{c}}{2\alpha}\right\}, \tag{3.21}$$

where α is the constant given in Algorithm 2.1, ξ, γ, c_p, and η are given in Algorithm 3.1, c_n is given in Hypothesis H4, and c_h, N, and $\tilde{c}$ are given by Lemmas 3.2, 3.3, and 3.4, respectively. By the constraint qualification hypothesis H7, we can assume that if $x^k \in V$, then $\mathcal{L}(x^k) \neq \emptyset$. Thus, Algorithm 3.1 starts with the radius $\Delta \geq \Delta_{\min}$ and ends with $\Delta_k = \gamma^r\Delta$, where r is the number of times that the radius was reduced in the algorithm. We shall consider two cases, respectively, with $\Delta_k \geq \gamma^2\Delta_\rho$ and $\Delta_k < \gamma^2\Delta_\rho$.

First case: $\Delta_k \geq \gamma^2\Delta_\rho$. In this case, using the hypothesis H4 and restricting the neighborhood V, if necessary, we have

$$\left\|n^k\right\| \leq c_n h\left(x^k\right) \leq \xi\gamma^2\Delta_\rho \leq \xi\Delta_k.$$

So, Algorithm 3.1 does not enter the restoration phase during the iteration k, that is, $k \notin \mathcal{K}_r$. Therefore, applying Lemma 3.4(i) with $\Delta' = \gamma^2\Delta_\rho$, we obtain

$$m_k\left(z^k\right) - m_k\left(x^{k+1}\right) = m_k\left(z^k\right) - m_k\left(z^k + t_{\Delta_k}\right) \geq \tilde{c}\gamma^2\Delta_\rho. \tag{3.22}$$

On the other hand, by Lemma 3.3,

$$\left|m_k\left(x^k\right) - m_k\left(z^k\right)\right| \leq Nh\left(x^k\right). \tag{3.23}$$

We can restrict again the neighborhood V, if necessary, so that

$$Nh\left(x^k\right) \leq \frac{1}{2}\tilde{c}\gamma^2\Delta_\rho, \quad c_p\left(h\left(x^k\right)\right)^2 \leq \frac{1}{2}\tilde{c}\gamma^2\Delta_\rho \quad \text{and} \quad h\left(x^k\right) \leq 1. \tag{3.24}$$

By (3.22)–(3.24), we have

$$pred_k \stackrel{\text{def}}{=} m_k\left(x^k\right) - m_k\left(x^{k+1}\right) \geq \frac{1}{2}\tilde{c}\gamma^2\Delta_\rho \geq c_p\left(h\left(x^k\right)\right)^2.$$

Then the mechanism of Algorithm 3.1 and the fact that $H_k \leq 1$ imply that

$$f_0\left(x^k\right) - f_0\left(x^{k+1}\right) \stackrel{\text{def}}{=} ared_k \geq \eta\, pred_k \geq \frac{1}{2}\eta\tilde{c}\gamma^2\Delta_\rho \geq \frac{1}{2}\eta\tilde{c}\gamma^2\Delta_\rho\sqrt{H_k}. \tag{3.25}$$

Second case: now, assume that $\Delta_k < \gamma^2\Delta_\rho$. In this case we shall analyze two possibilities. In the first one, we suppose that $h(x^k + d_\Delta) \geq H_k$ for all $\Delta \leq \gamma\Delta_\rho$ such that the trial step d_Δ has been computed. Let $\tilde{\Delta} = \frac{\Delta_k}{\gamma}$. Since $\Delta_k < \Delta_{\min}$, the trial step $\tilde{d} = d_{\tilde{\Delta}}$ was computed. Furthermore, $h(x^k + \tilde{d}) \geq H_k$ because $\tilde{\Delta} < \gamma\Delta_\rho$. So, using Lemma 3.2 and the definition of H_k, it follows that

$$c_h\Delta_k^2 = c_h\gamma^2\tilde{\Delta}^2 \geq \gamma^2 h\big(x^k + \tilde{d}\big) \geq \gamma^2 H_k \geq \gamma^2 h\left(x^k\right). \tag{3.26}$$

From Hypothesis H4, (3.21), and (3.26), we obtain

$$\left\|n^k\right\| \leq c_n h\left(x^k\right) \leq \frac{c_n c_h}{\gamma^2}\Delta_k^2 \leq \xi\Delta_k,$$

meaning that Algorithm 3.1 does not enter the restoration phase during the iteration k, that is, $k \notin \mathcal{K}_r$. Therefore, by Lemma 3.4(i) with $\Delta' = \Delta_k$, we have

$$m_k\left(z^k\right) - m_k\left(x^{k+1}\right) = m_k\left(z^k\right) - m_k\left(z^k + t_{\Delta_k}\right) \geq \tilde{c}\Delta_k. \tag{3.27}$$

Moreover, (3.23) remains true in this case and together with (3.21) and (3.26) yields

$$\left|m_k\left(x^k\right) - m_k\left(z^k\right)\right| \leq Nh\left(x^k\right) \leq \frac{Nc_h}{\gamma^2}\Delta_k^2 \leq \frac{1}{2}\tilde{c}\Delta_k. \tag{3.28}$$

Combining (3.27) and (3.28), we obtain

$$pred_k = m_k\left(x^k\right) - m_k\left(x^{k+1}\right) \geq \frac{1}{2}\tilde{c}\Delta_k. \tag{3.29}$$

By (3.21), (3.24), and (3.26),

$$pred_k \geq \frac{c_p c_h}{\gamma^2}\Delta_k^2 \geq c_p h\left(x^k\right) \geq c_p\left(h\left(x^k\right)\right)^2.$$

Thus, the mechanism of Algorithm 3.1, (3.26), and (3.29) imply that

$$f_0\left(x^k\right) - f_0\left(x^{k+1}\right) = ared_k \geq \eta\, pred_k \geq \frac{1}{2}\eta\tilde{c}\Delta_k \geq \frac{\eta\tilde{c}\gamma}{2\sqrt{c_h}}\sqrt{H_k}. \tag{3.30}$$

Let us see now the second possibility; that is, there exists $\Delta \leq \gamma\Delta_\rho$ such that $h(x^k + d_\Delta) < H_k$. Let $\hat{\Delta}$ be the first Δ satisfying such a condition. We shall show that $\hat{\Delta} = \Delta_k$. Let $\bar{d} = d_{\bar{\Delta}}$ be the trial step obtained with $\bar{\Delta} = \frac{\hat{\Delta}}{\gamma}$. We claim that

$$h\left(x^k + \bar{d}\right) \geq H_k. \tag{3.31}$$

Indeed, if $\bar{\Delta} \leq \gamma\Delta_\rho$, the definition of $\hat{\Delta}$ ensures the claim. On the other hand, if $\bar{\Delta} > \gamma\Delta_\rho$, then $\bar{\Delta} \in [\gamma^2\Delta_\rho, \Delta_\rho]$ and, applying Lemma 3.6, we have

$$h\left(x^k + \bar{d}\right) = h\left(z^k + t_{\bar{\Delta}}\right) \geq H_k.$$

So, the inequality (3.31) holds. As above, we can therefore prove that

$$c_h\hat{\Delta}^2 \geq \gamma^2 H_k \geq \gamma^2 h\left(x^k\right) \tag{3.32}$$

and

$$pred_{\hat{\Delta}} \stackrel{\text{def}}{=} m_k\left(x^k\right) - m_k\left(z^k + t_{\hat{\Delta}}\right) \geq \frac{1}{2}\tilde{c}\hat{\Delta}. \tag{3.33}$$

Now, by the same reasoning as in the proof of Lemma 3.5(ii), using (3.32) and (3.33), we obtain

$$ared_{\hat{\Delta}} \stackrel{\text{def}}{=} f_0\left(x^k\right) - f_0\left(z^k + t_{\hat{\Delta}}\right) \geq \eta\, pred_{\hat{\Delta}} \geq \frac{1}{2}\eta\tilde{c}\hat{\Delta}, \tag{3.34}$$

which together with (3.21) and (3.32) yields

$$ared_{\hat{\Delta}} > \frac{\alpha c_h}{\gamma^2}\hat{\Delta}^2 \geq \alpha h\left(x^k\right). \tag{3.35}$$

The definition of $\hat{\Delta}$ and (3.35) ensure that $z^k + t_{\hat{\Delta}}$ is accepted by the filter. Therefore, using (3.34), we conclude that $z^k + t_{\hat{\Delta}} = x^{k+1}$. Moreover, (3.32) and (3.34) imply that

$$f_0\left(x^k\right) - f_0\left(z^k + t_{\hat{\Delta}}\right) \geq \frac{1}{2}\eta\tilde{c}\hat{\Delta} \geq \frac{\eta\tilde{c}\gamma}{2\sqrt{c_h}}\sqrt{H_k},$$

that is,

$$f_0\left(x^k\right) - f_0\left(x^{k+1}\right) \geq \frac{\eta\tilde{c}\gamma}{2\sqrt{c_h}}\sqrt{H_k}. \tag{3.36}$$

Since (3.25), (3.30), and (3.36) run out all possibilities, by defining

$$M = \min\left\{\frac{1}{2}\eta\tilde{c}\gamma^2\Delta_\rho, \frac{\eta\tilde{c}\gamma}{2\sqrt{c_h}}\right\},$$

we complete the proof. □

4. Conclusions. In this work we have studied filter methods for nonlinear programming. These methods seem to be a successful strategy for globalizing algorithms without the use of merit functions. Since its appearance in 1997, the filter technique has been applied to many problems, including sequential linear programming (SLP), sequential quadratic programming (SQP), inexact restoration, interior-point methods, nonlinear systems of equations, unconstrained optimization, and nonsmooth convex constrained optimization.

Our purpose here was to present a general globally convergent filter algorithm that leaves the step computation separate from the main algorithm. This technique cleans the convergence analysis and accepts any method for computing the step, as long as this internal algorithm is efficient in the sense that the hypothesis H3 is satisfied. For completeness, we have shown that there are methods which satisfy the referred hypothesis.

Acknowledgment. We thank the referees for their valuable comments and suggestions which very much improved this paper.

REFERENCES

[1] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1995.
[2] R. H. Byrd, *Robust Trust Region Methods for Constrained Optimization*, Third SIAM Conference on Optimization, 1987.
[3] R. H. Byrd, J. C. Gilbert, and J. Nocedal, *A trust region method based on interior point techniques for nonlinear programming*, Math. Program., 89 (2000), pp. 149–185.
[4] C. M. Chin, *A local convergence theory of a filter line search method for nonlinear programming*, Technical report, Numerical Optimization Report, Department of Statistics, University of Oxford, England, January 2003.
[5] C. M. Chin and R. Fletcher, *On the global convergence of an SLP-filter algorithm that takes EQP steps*, Math. Program., 96 (2003), pp. 161–177.
[6] R. Fletcher, N. Gould, S. Leyffer, P. L. Toint, and A. Wächter, *Global convergence of a trust-region SQP-filter algorithm for general nonlinear programming*, SIAM J. Optim., 13 (2002), pp. 635–659.

[7] R. Fletcher and S. Leyffer, *Nonlinear programming without a penalty function*, Math. Program. Ser. A, 91 (2002), pp. 239–269.
[8] R. Fletcher, S. Leyffer, and P. L. Toint, *On the global convergence of an SLP-filter algorithm*, Technical Report NA/183, Dept. of Mathematics, Dundee University, 1998.
[9] R. Fletcher, S. Leyffer, and P. L. Toint, *On the global convergence of a filter-SQP algorithm*, SIAM J. Optim., 13 (2002), pp. 44–59.
[10] R. Fletcher, S. Leyffer, and P. L. Toint, *A brief history of filter method*, SIAG/Optimization Views-and-News, 18 (2007), pp. 2–12.
[11] D. M. Gay, M. L. Overton, and M. H. Wright, *A primal-dual interior method for nonconvex nonlinear programming*, in Advances in Nonlinear Programming, Y. Y, ed., Kluwer Academic Publishers, Dordretch, 1998, pp. 31–56.
[12] F. A. M. Gomes, M. C. Maciel, and J. M. Martínez, *Nonlinear programming algorithms using trust regions and augmented Lagrangians with nonmonotone penalty parameters*, Math. Program., 84 (1999), pp. 161–200.
[13] C. C. Gonzaga, E. W. Karas, and M. Vanti, *A globally convergent filter method for nonlinear programming*, SIAM J. Optim., 14 (2003), pp. 646–669.
[14] N. I. M. Gould, S. Leyffer, and P. L. Toint, *A multidimensional filter algorithm for nonlinear equations and nonlinear least-squares*, SIAM J. Optim., 15 (2004), pp. 17–38.
[15] N. I. M. Gould, C. Sainvitu, and P. L. Toint, *A filter-trust-region method for unconstrained optimization*, SIAM J. Optim., 16 (2006), pp. 341–357.
[16] N. I. M. Gould and P. L. Toint, *The filter idea and its application to the nonlinear feasibility problem*, in Proceedings of the 20th Biennal Conference on Numerical Analysis, D. F. Griffiths and G. A. Watson, eds., Scotland, 2003, pp. 73–79.
[17] N. I. M. Gould and P. L. Toint, *FILTRANE, a Fortran* 95 *filter-trust-region package for solving nonlinear feasibility problems*, Trans. ACM Math. Software, 33 (2007), pp. 3–25.
[18] E. W. Karas, A. A. Ribeiro, C. Sagastizábal, and M. Solodov, *A bundle-filter method for nonsmooth convex constrained optimization*, Math. Program. Ser. B, 116 (2009), pp. 297–320.
[19] J. M. Martínez, *Inexact-restoration method with Lagrangian tangent decrease and a new merit function for nonlinear programming*, J. Optim. Theory Appl., 111 (2001), pp. 39–58.
[20] J. M. Martínez and E. A. Pilotta, *Inexact restoration algorithm for constrained optimization*, J. Optim. Theory Appl., 104 (2000), pp. 135–163.
[21] J. M. Martínez and B. F. Svaiter, *A practical optimality condition without constraint qualifications for nonlinear programming*, J. Optim. Theory Appl., 118 (2003), pp. 117–133.
[22] E. Omojokun, *Trust Region Algorithms for Optimization with Nonlinear Equality and Inequality Constraints*, Ph.D. thesis, Dept. of Computer Science, University of Colorado, 1991.
[23] C. Sagastizábal and M. Solodov, *An infeasible bundle method for nonsmooth convex constrained optimization without a penalty function or a filter*, SIAM J. Optim., 16 (2005), pp. 146–169.
[24] M. Ulbrich, S. Ulbrich, and L. N. Vicente, *A globally convergent primal-dual interior-point filter method for nonlinear programming*, Math. Program., Ser. A, 100 (2004), pp. 379–410.
[25] S. Ulbrich, *On the superlinear local convergence of a filter-SQP method*, Math. Program., Ser. B, 100 (2004), pp. 217–245.
[26] R. J. Vanderbei and D. F. Shanno, *An interior-point algorithm for nonconvex nonlinear programming*, Comput. Optim. Appl., 13 (1999), pp. 231–252.
[27] A. Wächter and L. T. Biegler, *Line search filter methods for nonlinear programming: Local convergence*, SIAM J. Optim., 16 (2005), pp. 32–48.
[28] A. Wächter and L. T. Biegler, *Line search filter methods for nonlinear programming: Motivation and global convergence*, SIAM J. Optim., 16 (2005), pp. 1–31.
[29] A. Wächter and L. T. Biegler, *On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming*, Math. Program., 106 (2006), pp. 25–57.

SIAM J. OPTIM.
Vol. 19, No. 3, pp. 1250–1269

PARAMETERIZED MINIMAX PROBLEM: ON LIPSCHITZ-LIKE DEPENDENCE OF THE SOLUTION WITH RESPECT TO THE PARAMETER*

MARC QUINCAMPOIX† AND NADIA ZLATEVA‡

Abstract. We study Lipschitz continuity with respect to the parameter of the set of solutions of a parameterized minimax problem on a product Banach space. We present a sufficient condition, ensuring that the map which to any value of the parameter assigns the set of solutions of the problem (possibly multi-valued, and unbounded) possesses Aubin property.

Key words. parameterized minimax problem, saddle point, set-valued map, Aubin property, pseudo-Lipschitz continuity

AMS subject classifications. 90C31, 90C47, 46N10

DOI. 10.1137/060653718

1. Introduction.

Consider the parameterized minimax problem

$$M(\lambda) \qquad \inf_{x\in K} \sup_{y\in L} f(x,y,\lambda),$$

where $\lambda \in \Lambda$ is a parameter.

Here K and L are nonempty closed subsets of the Banach spaces X and Y, respectively; $\{f(\cdot,\cdot,\lambda) : X \times Y \to \mathbb{R}, \lambda \in \Lambda\}$ is a family of real-valued functions parameterized by $\lambda \in \Lambda$, where Λ is a subset of the Banach space Z.

Saddle point of $f(\cdot,\cdot,\lambda)$ on $K \times L$ is any point $(\overline{x},\overline{y}) \in K \times L$ that satisfies

$$f(\overline{x},y,\lambda) \le f(\overline{x},\overline{y},\lambda) \le f(x,\overline{y},\lambda) \quad \forall x \in K, \quad \forall y \in L.$$

A saddle point $(\overline{x},\overline{y})$ of $f(\cdot,\cdot,\lambda)$ on $K \times L$ can be considered as a solution of the minimax problem $M(\lambda)$ by reason of $(\overline{x},\overline{y}) \in K \times L$ and $f(\overline{x},\overline{y}) = \inf_{x\in K} \sup_{y\in L} f(x,y,\lambda)$. Let us denote the (possibly empty) set of all saddle points of the function $f(\cdot,\cdot,\lambda)$ on $K \times L$ by

$$\mathcal{S}(\lambda) := \{(\overline{x},\overline{y}) \in K\times L : f(\overline{x},y,\lambda) \le f(\overline{x},\overline{y},\lambda) \le f(x,\overline{y},\lambda),\ \forall x \in K,\ \forall y \in L\}. \tag{1.1}$$

That $\mathcal{S}(\lambda)$ is nonempty can be ensured in several cases. For example, if K and L are convex sets, $f(x,y,\lambda)$ is convex and lower semicontinuous in x, concave and upper semicontinuous in y, and there are $x_0 \in K$ and $y_0 \in L$ such that $f(\cdot,y_0,\lambda)$ is inf-compact and $f(x_0,\cdot,\lambda)$ is sup-compact, then $\mathcal{S}(\lambda) \neq \emptyset$ by a minimax result due to Hartung [17], Theorem 1 (see also [3], Theorem 6.2.8). When, moreover, $f(x,y,\lambda)$ is strictly convex in x and strictly concave in y, then $\mathcal{S}(\lambda)$ is a singleton.

In the present work we presume the existence of saddle points for $M(\lambda)$ and focus our attention on studying Lipschitz-like dependence of the solution set $\mathcal{S}(\lambda)$ on the

*Received by the editors March 6, 2006; accepted for publication (in revised form) June 24, 2008; published electronically November 19, 2008. This research has been supported by European's Community Human Potential Programme HPRN-CT-2002-00281 (Evolution Equations).
http://www.siam.org/journals/siopt/19-3/65371.html

†Laboratoire de Mathématiques, UMR CNRS 6205, Université de Bretagne Occidentale, 6, avenue Victor Le Gorgeu, 29200 Brest, France (Marc.Quincampoix@univ-brest.fr).

‡Faculty of Mathematics and Informatics, St. Kliment Ohridski University of Sofia, 5, James Bourchier Blvd, 1164 Sofia, Bulgaria (zlateva@fmi.uni-sofia.bg).

parameter λ. That is, we find sufficient conditions for Lipschitz-like continuity of the set-valued map

$$\mathcal{S} : \lambda \rightrightarrows \mathcal{S}(\lambda)$$

from Λ to nonempty subsets of $K \times L$.

Of course, when the map $\mathcal{S}$ is single-valued, the Lipschitz continuity is understood in the classical sense. However, the map $\mathcal{S}$ could be multivalued. Moreover, its values $\mathcal{S}(\lambda)$ could be unbounded sets. A notion of Lipschitz-like continuity very appropriate for such a case is due to Aubin [1, 2]:

The multivalued map $S : \Lambda \rightrightarrows X$ has *Aubin property*, or it is *Aubin continuous, near* $(\overline{\lambda}, \overline{x}) \in \operatorname{gph} S$, if there are positive constant κ and neighborhoods U of $\overline{x}$, and V of $\overline{\lambda}$, such that

$$e(S(\lambda) \cap U, S(\mu)) \leq \kappa \|\lambda - \mu\|, \quad \forall \lambda, \mu \in \Lambda \cap V, \tag{1.2}$$

where $e(A, B) := \sup_{x \in A} d(x, B)$ is the *excess* from set A to set B with $e(\emptyset, B) = +\infty$. S is said to be *Aubin continuous* if S is Aubin continuous near any point $(\overline{\lambda}, \overline{x}) \in \operatorname{gph} S$.

For various applications of Aubin continuity in the field of nonlinear analysis and optimization the reader is referred, e.g., to [1, 2, 4, 23]. The Aubin property of a map S near $(\overline{\lambda}, \overline{x})$ is known to be equivalent to the metric regularity of S^{-1} near $(\overline{x}, \overline{\lambda})$ and was originally introduced in [2] under the name of *pseudo-Lipschitz continuity*. For bibliographical details see [23].

Whenever S is locally bounded, Aubin continuity coincides with the classical notion for Lipschitz continuity of set-valued maps [4, 23]

$$e(S(\lambda), S(\mu)) \leq \kappa \|\lambda - \mu\|, \quad \forall \lambda, \mu,$$

but Aubin property works without any boundedness imposed on the values of S. Aubin property is in fact Lipschitzean property localized in the range space, as well as in the domain space.

In the present paper we establish quite general sufficient conditions for Aubin continuity of the saddle point map $\mathcal{S} : \lambda \rightrightarrows \mathcal{S}(\lambda)$ arising from the parameterized minimax problem $M(\lambda)$. Examples illustrating this condition are presented. Several corollaries related to the case of convex-concave smooth data are also sketched.

The paper is organized as follows. In section 2, after a short subsection devoted to preliminaries, we formulate and prove a sufficient condition for Aubin continuity of the solution map $S : \Lambda \rightrightarrows X$ of a parameterized minimization problem

$$P(\lambda) \qquad \inf_{x \in K} f(x, \lambda).$$

Many authors study Lipschitz-like dependence on λ of the solutions of the associated generalized Euler equation

$$0 \in \nabla_x f(x, \lambda) + N_K(x);$$

see [7, 15, 25] and the references therein for recent developments. Here we do not follow that approach, because the map $St : \lambda \rightrightarrows St(\lambda)$, which to any λ assigns the set $St(\lambda)$ of solutions of the generalized Euler equation, does not inherit Aubin continuity property from S (see Example 2.6 for a parameterized problem such that the corresponding S is Aubin continuous while St is not).

In section 3 we present our main result (Theorem 3.2), which is a sufficient condition for Aubin continuity of the saddle point map $\mathcal{S} : \Lambda \rightrightarrows X \times Y$ of a parameterized minimax problem $M(\lambda)$.

It is clear that the results of section 2 are contained in the more general framework of section 3. Nevertheless, we think that presenting the proof of the former simple case will help the understanding of the more technical proof of the latter general case.

Section 4 relates the obtained results to some questions in the field of two-player zero sum differential games.

2. Parameterized minimization problem.

2.1. Preliminaries. As already said, X stands for a Banach space. We denote its norm by $\|\cdot\|$, and its open unit ball by B°. The dual space is denoted by X^*, while for the duality brackets notation $\langle\cdot,\cdot\rangle$ is used.

For $C \subset X$ the *distance function* to C is $d(x,C) := \inf_{c\in C} \|x-c\|$ if $C \neq \emptyset$, and $d(x,C) := +\infty$ if $C = \emptyset$.

Function $f : X \to \mathbb{R}$ is *Gâteaux differentiable* at $\overline{x} \in X$ if there exists $\nabla f(\overline{x}) \in X^*$, called the *Gâteaux derivative* of f at $\overline{x}$, such that for any $h \in X$,

$$\lim_{t\to 0} \frac{f(\overline{x}+th) - f(\overline{x})}{t} = \langle \nabla f(\overline{x}), h\rangle.$$

Also, f is said to be *strictly differentiable* at $\overline{x}$ whenever

$$\lim_{\substack{x\to\overline{x}\\ t\to 0}} \frac{f(x+th) - f(x)}{t} = \langle \nabla f(\overline{x}), h\rangle.$$

Given an open set $U \subset X$ we denote by $C^{1,\alpha}(U)$ the class of all Gâteaux differentiable functions $f : U \to \mathbb{R}$ such that $\nabla f : U \to X^*$ is α-*Hölder* on U, that is, for some constant $L > 0$,

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x-y\|^\alpha, \quad \forall x, y \in U.$$

Let Z be a Banach space, whose norm is also denoted by $\|\cdot\|$. Let S be a *map* from $\Lambda \subset Z$ to X. If not stated otherwise, *map* means *set-valued map*. In order to outline the multivaluedness we write $S : Z \rightrightarrows X$. The *inverse* $S^{-1} : X \rightrightarrows Z$ of S is defined by $\lambda \in S^{-1}(x) \iff x \in S(\lambda)$. The *graph*, *domain*, and *range* sets of S are given by

$$\operatorname{gph} S := \{(\lambda, x) \mid x \in S(\lambda)\}, \quad \operatorname{dom} S := \{\lambda \mid S(\lambda) \neq \emptyset\}, \quad \operatorname{rge} S := \operatorname{dom} S^{-1},$$

respectively.

Any product space $X \times Z$ of Banach spaces X and Z is considered with the supremum norm $\|(x,z)\| := \max\{\|x\|, \|z\|\}$.

2.2. Assumptions. Let $\{f(\cdot,\lambda) : X \to \mathbb{R}, \lambda \in \Lambda\}$ be a family of functions parameterized by $\lambda \in \Lambda \subset Z$. We look for sufficient conditions to ensure Aubin continuity of the solutions of the parameterized family of constrained minimization problems:

$$P(\lambda) \qquad \inf_{x\in K} f(x,\lambda),$$

where K is a given nonempty closed set in X.

For $\lambda \in \Lambda$, the (possibly empty) set of solutions of the minimization problem $P(\lambda)$ is denoted by

$$S(\lambda) := \left\{ \overline{x} \in K : f(\overline{x}, \lambda) = \inf_{x \in K} f(x, \lambda) \right\},$$

and its optimal value by

$$m(\lambda) := \inf_{x \in K} f(x, \lambda).$$

It is well known that even for smooth parameterized problem $P(\lambda)$ the solution $S : \Lambda \rightrightarrows X$ may fail Lipschitz continuity. For example, for $f(x, \lambda) = \frac{1}{4}x^4 - \lambda x$, where $x, \lambda \in \mathbb{R}$, and $K = [-1, 1]$, we see that for $\lambda \in (-1, 1)$ the solution is $S(\lambda) = \{\sqrt[3]{\lambda}\}$, and it is not Lipschitz continuous at $\lambda = 0$ ([7], Example 4.31).

Hence, to establish Lipschitz behavior of S one needs something more than the standard requirements. We now turn to relevant analysis of $P(\lambda)$.

DEFINITION 2.1. *Let X and Z be Banach spaces. Let $U \subset X$, $V \subset Z$ be nonempty. We denote by $\mathfrak{L}^{\alpha,\beta}(U; V)$, $\alpha, \beta \in [0, 1]$, the class of all functions $g : U \times U \times V \to \mathbb{R}$ such that there exists a constant $k_g > 0$ such that for all $x, x' \in U$ and all $\lambda, \lambda' \in V$,*

$$|g(x, x', \lambda) - g(x, x', \lambda')| \leq k_g \|x - x'\|^\alpha \|\lambda - \lambda'\|^\beta.$$

For example, $g \in \mathfrak{L}^{1,1}(U; V)$ means that $g(x, x', \cdot)$ is Lipschitz on V and its best Lipschitz constant $L(x, x')$ satisfies $L(x, x') \leq k\|x - x'\|$ for some positive constant k and all $x, x' \in U$.

With the parameterized family of functions $\{f(\cdot, \lambda), \lambda \in \Lambda\}$ one may associate two difference functions: the function $f_1 : X \times X \times \Lambda \to \mathbb{R}$ defined by

$$f_1(x, x', \lambda) := f(x, \lambda) - f(x', \lambda),$$

and the function $f_2 : \Lambda \times \Lambda \times X \to \mathbb{R}$ defined by

$$f_2(\lambda, \lambda', x) := f(x, \lambda) - f(x, \lambda').$$

The above notions are linked through the following:

PROPOSITION 2.2. *For any $U \subset X$, $V \subset Z$ the function $f_1 \in \mathfrak{L}^{\alpha,\beta}(U; V)$ if and only if $f_2 \in \mathfrak{L}^{\beta,\alpha}(V; U)$.*

Proof. Let $f_1 \in \mathfrak{L}^{\alpha,\beta}(U; V)$. Take any $x, x' \in U$, and any $\lambda, \lambda' \in V$. Since

$$\begin{aligned} f_2(\lambda, \lambda', x) - f_2(\lambda, \lambda', x') &= [f(x, \lambda) - f(x, \lambda')] - [f(x', \lambda) - f(x', \lambda')] \\ &= [f(x, \lambda) - f(x', \lambda)] - [f(x, \lambda') - f(x', \lambda')] \\ &= f_1(x, x', \lambda) - f_1(x, x', \lambda') \leq k_{f_1} \|x - x'\|^\alpha \|\lambda - \lambda'\|^\beta, \end{aligned}$$

one can take $k_{f_2} := k_{f_1}$ to conclude that $f_2 \in \mathfrak{L}^{\beta,\alpha}(V; U)$. The proof of the other direction is similar. ∎

We are ready to present the sufficient condition for Aubin continuity of the solution map.

Given a $(\overline{\lambda}, \overline{x}) \in \operatorname{gph} S$, consider the following local assumption A at $(\overline{\lambda}, \overline{x})$:

$$(A) \quad \begin{cases} \text{there exist neighborhoods } U \text{ of } \overline{x} \text{ and } V \text{ of } \overline{\lambda}, \text{ such that} \\ 1. \quad S(\lambda) \cap U \neq \emptyset \text{ for all } \lambda \in \Lambda \cap V; \\ \text{and there exist constants } c > 0 \text{ and } \alpha \in [0, 1] \text{ such that} \\ 2. \quad f(x, \lambda') \geq m(\lambda') + c d^{1+\alpha}(x, S(\lambda')), \ \forall \lambda, \lambda' \in \Lambda \cap V, \ \forall x \in S(\lambda) \cap U; \\ 3. \quad f_1 \in \mathfrak{L}^{\alpha,1}(K; \Lambda \cap V). \end{cases}$$

By Proposition 2.2 it is clear that assumption $A3$ could be replaced with $f_2 \in \mathfrak{L}^{1,\alpha}(\Lambda \cap V; K)$.

It is clear that $A1$ implies that $m(\lambda)$ is finite for all $\lambda \in \Lambda \cap V$.

In the case $\alpha = 1$, assumption $A2$ can be considered as a relaxed (in x) uniform (in λ') version of the so-called second-order growth condition. One says that the *second-order growth condition* holds for the problem

$$\inf_{x \in K} f(x)$$

in a neighborhood N of the solution set S_0, if there exists a constant $c > 0$ such that

$$f(x) \geq \inf_K f + cd^2(x, S_0), \quad \forall x \in K \cap N. \tag{2.1}$$

This condition is involved in a number of works (see [5, 6, 7, 19, 24]) in order to ensure Lipschitz stability of the solution map S of the constrained minimization problem. Let us recall that S is said to be *Lipschitz stable* or, equivalently, *upper Lipschitz* at a point $\overline{\lambda} \in \Lambda$, if there exist a constant $\kappa > 0$ and a neighborhood V of $\overline{\lambda}$ such that it holds that

$$e(S(\lambda), S(\overline{\lambda})) \leq \kappa \left\|\lambda - \overline{\lambda}\right\|, \quad \forall \lambda \in \Lambda \cap V.$$

Let us note that Lipschitz stability is a point property: it holds for S at a fixed point $\overline{\lambda}$, while Aubin continuity we wish to obtain, is a local property, and it holds uniformly at all points μ in some neighborhood V of the referenced point $\overline{\lambda}$. Obviously, Aubin continuity of S near $(\overline{\lambda}, \overline{x})$ implies Lipschitz stability of $S \cap U$ at $\overline{\lambda}$ while the opposite implication is not always true.

A stronger version of uniform second-order growth condition than $A2$ with $\alpha = 1$ is given in [7], Definition 5.16. It implies single-valuedness and local Lipschitz continuity of S (cf. [7], Theorem 5.17 and Remark 5.19). In contrast, assumption $A2$ does not imply neither single-valuedness nor local boundedness of the solution map S (see Example 2.6).

We would now give a few examples of parameterized families of functions $\{f(\cdot, \lambda), \lambda \in \Lambda\}$ for which A holds, in this way showing the consistency of our main assumption.

Obviously, the compactness of K and lower semicontinuity of $f(\cdot, \lambda)$ on K are sufficient to ensure $A1$ (note that weak compactness and weak lower semicontinuity would do just as well).

$A2$ with $\alpha = 1$ is satisfied at any $(\overline{\lambda}, \overline{x}) \in \operatorname{gph} S$ provided that, for example, $U = X$, $V = Z$, and $K \subset X$ is a nonempty closed convex set, the functions $f(\cdot, \lambda)$ are lower semicontinuous, and uniformly on $\lambda \in \Lambda$ strongly convex on K, that is, for some constant $c > 0$ the inequality

$$f(tx' + (1-t)x'', \lambda) \leq tf(x', \lambda) + (1-t)f(x'', \lambda) - ct(1-t)\|x' - x''\|^2$$

holds for every $t \in [0, 1]$, every $x', x'' \in K$ and every $\lambda \in \Lambda$.

Lemma 2.3 below provides examples of parameterized families of functions satisfying $A3$. However, we need a few more definitions before stating this lemma.

Recall that the *Clarke generalized derivative* of Lipschitz function $f : X \to \mathbb{R}$ at $\overline{x} \in X$ in direction $h \in X$ is

$$f^\circ(\overline{x}; h) := \limsup_{\substack{x \to \overline{x} \\ t \downarrow 0}} \frac{f(x + th) - f(x)}{t},$$

and the *Clarke subdifferential* at $\overline{x}$ is the nonempty w^* compact set

$$\partial f(\overline{x}) := \{x^* \in X^* : \langle x^*, h\rangle \le f^\circ(\overline{x}; h),\ \forall h \in X\};$$

see [12]. It is well known that for any $h \in X$ there exists some $x^* \in \partial f(\overline{x})$ such that

$$\langle x^*, h\rangle = f^\circ(\overline{x}; h).$$

Lipschitz function $f : U \to \mathbb{R}$ is said to be *regular* on an open set $U \subset X$ if for any $h \in X$ and any $\overline{x} \in U$ its *directional derivative*

$$f'(\overline{x}; h) := \lim_{t \downarrow 0} \frac{f(\overline{x} + th) - f(\overline{x})}{t}$$

exists and is equal to $f^\circ(\overline{x}; h)$. Convex continuous functions and strictly differentiable functions are examples of regular functions.

Let $f(x, \lambda)$ be Lipschitz on each variable bivariate function. Denote by $f_x^\circ(\overline{x}, \overline{\lambda}; h)$ and by $f_x'(\overline{x}, \overline{\lambda}; h)$ the generalized derivative and the directional derivative of $f(\cdot, \overline{\lambda})$ at $\overline{x}$ in direction h, respectively. Also, denote by $\partial_x f(\overline{x}, \overline{\lambda})$ the partial Clarke subdifferential of $f(\cdot, \overline{\lambda})$ at $\overline{x}$, and by $\partial_\lambda f(\overline{x}, \overline{\lambda})$ the partial Clarke subdifferential of $f(\overline{x}, \cdot)$ at $\overline{\lambda}$.

LEMMA 2.3. *Let $(\overline{\lambda}, \overline{x}) \in \operatorname{gph} S$ and let $U \subset X$ and $V \subset Z$ be convex neighborhoods of K and $\overline{\lambda}$, respectively. Consider the conditions:*

$$(F1) \qquad \begin{cases} \text{for } \lambda \in \Lambda \cap V,\ f(\cdot, \lambda) \text{ is Lipschitz and regular on } U \text{ and} \\ \partial_x f(x, \cdot) : \Lambda \cap V \to X^* \text{ is a } k\text{-Lipschitz map on } \Lambda \cap V \\ \text{with } k \text{ that does not depend on } x \in U, \end{cases}$$

$$(F2) \qquad \begin{cases} \text{for } x \in K,\ f(x, \cdot) \text{ is Lipschitz and regular on } V \text{ and} \\ \partial_\lambda f(\cdot, \lambda) : K \to Z^* \text{ is a } k\text{-Lipschitz map on } K \\ \text{with } k \text{ that does not depend on } \lambda \in V. \end{cases}$$

If f satisfies $F1$ or $F2$, then $A3$ holds with $\alpha = 1$.

Proof. Let f satisfy $F1$. Fix $x, y \in K$ and $\lambda, \mu \in \Lambda \cap V$. Consider the function $r(t) := f(y + t(x - y), \lambda)$ which is well-defined on an open interval I containing $[0, 1]$. Since the function $f(\cdot, \lambda)$ is assumed to be Lipschitz on U, we have that r is Lipschitz on I. By Rademacher's theorem, for almost all $t \in [0, 1]$ there exists

$$\begin{aligned} r'(t) &= \lim_{s \to 0} \frac{r(t+s) - r(t)}{s} = \lim_{s \downarrow 0} \frac{f(y + t(x-y) + s(x-y), \lambda) - f(y + t(x-y), \lambda)}{s} \\ &= f_x'(y + t(x-y), \lambda; x - y) = f_x^\circ(y + t(x-y), \lambda; x - y). \end{aligned}$$

The last equality holds because $f(\cdot, \lambda)$ is regular on U.

Hence,

$$f(x, \lambda) - f(y, \lambda) = r(1) - r(0) = \int_0^1 r'(t)\, dt = \int_0^1 f_x^\circ(y + t(x-y), \lambda; x - y)\, dt. \tag{2.2}$$

Similarly,

$$f(x, \mu) - f(y, \mu) = \int_0^1 f_x^\circ(y + t(x-y), \mu; x - y)\, dt. \tag{2.3}$$

There exists $x^*_\lambda(t) \in \partial_x f(y + t(x-y), \lambda)$ such that $f^\circ_x(y + t(x-y), \lambda; x-y) = \langle x^*_\lambda(t), x-y\rangle$, so (2.2) becomes

$$f(x,\lambda) - f(y,\lambda) = \int_0^1 \langle x^*_\lambda(t), x-y\rangle \, dt. \tag{2.4}$$

Since $x_\lambda(t) \in \partial_x f(y + t(x-y), \lambda)$ and the multivalued map $\partial_x f(x,\cdot) : \Lambda \cap V \to X^*$ is k–Lipschitz continuous with w^* compact images, there is $x^*_\mu(t) \in \partial_x f(y+t(x-y), \mu)$ such that $\|x^*_\lambda(t) - x^*_\mu(t)\| \le k\|\lambda - \mu\|$. Note that k does not depend on either $t \in [0,1]$ or $x, y \in U$.

Let us use these for estimating $f_1(x,y,\lambda) - f_1(x,y,\mu)$.

From (2.4) we get

$$\begin{aligned} f(x,\lambda) - f(y,\lambda) &= \int_0^1 \langle x^*_\lambda(t) - x^*_\mu(t), x-y\rangle \, dt + \int_0^1 \langle x^*_\mu(t), x-y\rangle \, dt \\ &\le \int_0^1 \|x^*_\lambda(t) - x^*_\mu(t)\| \, \|x-y\| \, dt + \int_0^1 \langle x^*_\mu(t), x-y\rangle \, dt \\ &\le k\|\lambda-\mu\| \, \|x-y\| + \int_0^1 \langle x^*_\mu(t), x-y\rangle \, dt. \end{aligned}$$

Since $x^*_\mu(t) \in \partial_x f(y+t(x-y), \mu)$, it holds that $\langle x^*_\mu(t), x-y\rangle \le f^\circ_x(y+t(x-y), \mu; x-y)$, and by (2.3) we have

$$\int_0^1 \langle x^*_\mu(t), x-y\rangle \, dt \le \int_0^1 f^\circ_x(y + t(x-y), \mu; x-y) \, dt = f(x,\mu) - f(y,\mu).$$

Hence,

$$f(x,\lambda) - f(y,\lambda) \le f(x,\mu) - f(y,\mu) + k\|\lambda-\mu\|\|x-y\|;$$

that is, $f_1(x,y,\lambda) \le f_1(x,y,\mu) + k\|\lambda-\mu\|\|x-y\|$, or

$$f_1(x,y,\lambda) - f_1(x,y,\mu) \le k\|\lambda-\mu\|\|x-y\|,$$

which means that $f_1 \in \mathfrak{L}^{1,1}(K; \Lambda \cap V)$.

If f satisfies $F2$, then by the same reasoning one obtains that $f_2 \in \mathfrak{L}^{1,1}(\Lambda \cap V; K)$ and by Proposition 2.2, $A3$ holds. □

It is interesting to note here that the regularity (in particular, the differentiability) can be asked for the argument x as in $F1$, or for the parameter λ as in $F2$.

It is clear that both $F1$ and $F2$ hold whenever $f \in C^{1,1}(U \times V)$.

2.3. Lipschitz-like continuity of the solution map. Here we prove that given $(\overline{\lambda}, \overline{x}) \in \operatorname{gph} S$, assumption A is sufficient to ensure Aubin continuity of the solution map S near $(\overline{\lambda}, \overline{x})$.

PROPOSITION 2.4. *Assume that X and Z are Banach spaces and consider a family of constraint minimization problems $P(\lambda)$ parameterized by $\lambda \in \Lambda$, a nonempty subset of Z.*

If for some $(\overline{\lambda}, \overline{x}) \in \operatorname{gph} S$ assumption A holds, then

$$e(S(\lambda) \cap U, S(\mu)) \le \frac{k_{f_1}}{c}\|\lambda - \mu\|, \quad \forall \lambda, \mu \in \Lambda \cap V, \tag{2.5}$$

and the solution map S is Aubin continuous near $(\overline{\lambda}, \overline{x}) \in \operatorname{gph} S$.

Proof. Take any $\lambda \in \Lambda \cap V$ and any $x_\lambda \in S(\lambda) \cap U$ (which is a nonempty set thanks to $A1$). By $A2$, for arbitrary $\mu \in \Lambda \cap V$

$$f(x_\lambda, \mu) \geq m(\mu) + cd^{1+\alpha}(x_\lambda, S(\mu)). \tag{2.6}$$

Since by $A1$ the set $S(\mu) \cap U$ is nonempty, for any $\varepsilon > 0$ there exists some $x_\mu^\varepsilon \in S(\mu)$ such that

$$\|x_\lambda - x_\mu^\varepsilon\| \leq d(x_\lambda, S(\mu)) + \varepsilon. \tag{2.7}$$

As $x_\mu^\varepsilon \in S(\mu)$ we have $m(\mu) = f(x_\mu^\varepsilon, \mu)$ and inequality (2.6) reads

$$f(x_\lambda, \mu) \geq f(x_\mu^\varepsilon, \mu) + cd^{1+\alpha}(x_\lambda, S(\mu)). \tag{2.8}$$

Since $x_\lambda \in S(\lambda)$, we have that

$$f(x_\mu^\varepsilon, \lambda) \geq f(x_\lambda, \lambda). \tag{2.9}$$

By adding (2.8) and (2.9) and rearranging, we obtain

$$[f(x_\lambda, \mu) - f(x_\mu^\varepsilon, \mu)] - [f(x_\lambda, \lambda) - f(x_\mu^\varepsilon, \lambda)] \geq cd^{1+\alpha}(x_\lambda, S(\mu)).$$

That is,

$$f_1\left(x_\lambda, x_\mu^\varepsilon, \mu\right) - f_1\left(x_\lambda, x_\mu^\varepsilon, \lambda\right) \geq cd^{1+\alpha}(x_\lambda, S(\mu)). \tag{2.10}$$

Using $A3$, that is, $f_1 \in \mathfrak{L}^{\alpha,1}(K; \Lambda \cap V)$, we estimate the left-hand side of (2.10):

$$f_1\left(x_\lambda, x_\mu^\varepsilon, \mu\right) - f_1\left(x_\lambda, x_\mu^\varepsilon, \lambda\right) \leq k_{f_1}\|x_\lambda - x_\mu^\varepsilon\|^\alpha \|\lambda - \mu\|.$$

Hence, we have that $k_{f_1}\|\lambda - \mu\| \, \|x_\lambda - x_\mu^\varepsilon\|^\alpha \geq cd^{1+\alpha}(x_\lambda, S(\mu))$. From this and (2.7) it follows that

$$k_{f_1}\|\lambda - \mu\|[d(x_\lambda, S(\mu)) + \varepsilon]^\alpha \geq cd^{1+\alpha}(x_\lambda, S(\mu)).$$

Letting $\varepsilon \downarrow 0$ and then dividing by $d^\alpha(x_\lambda, S(\mu)) > 0$ (if $= 0$ the inequality below is trivial), we obtain $k_{f_1}\|\lambda - \mu\| \geq cd(x_\lambda, S(\mu))$, or

$$d(x_\lambda, S(\mu)) \leq \frac{k_{f_1}}{c}\|\lambda - \mu\|.$$

As x_λ was an arbitrary point in $S(\lambda) \cap U$, the latter yields

$$e(S(\lambda) \cap U, S(\mu)) \leq \frac{k_{f_1}}{c}\|\lambda - \mu\|,$$

completing the proof. □

2.4. Examples and corollaries. The following is a basic example of nonsmooth parameterized minimization problem with Lipschitz continuous solution map with unbounded values. We show that it is within the scope of Proposition 2.4.

Example 2.5. Let $K = \mathbb{R}^2$ and

$$f(x_1, x_2, \lambda) := |x_1 - x_2 - \lambda|,$$

$x_1, x_2, \lambda \in \mathbb{R}$. Consider the parameterized family of unconstrained minimization problems over the plane

$$P(\lambda) \qquad \inf_{x_1,x_2} f(x_1, x_2, \lambda).$$

Then the solution map $S : \lambda \rightrightarrows S(\lambda)$ is Lipschitz continuous.

Proof. Obviously, for any $\lambda \in \mathbb{R}$ the solution set consists of a single line, i.e., $S(\lambda) = \{(x_1, x_2) : x_1 - x_2 = \lambda\}$. Moreover, for λ and μ the solution sets $S(\lambda)$ and $S(\mu)$ are parallel lines. The distance between $S(\lambda)$ and $S(\mu)$ is the distance from any point $(\overline{x}_1, \overline{x}_2) \in S(\lambda)$ to the line $x_1 - x_2 = \mu$ which is equal to $\frac{|\overline{x}_1 - \overline{x}_2 - \mu|}{\sqrt{2}} = \frac{|\lambda - \mu|}{\sqrt{2}}$, so the map S is Lipschitz continuous with Lipschitz constant $\frac{1}{\sqrt{2}}$.

Note that the sufficient condition A holds. Indeed

$A1$ holds with $U \equiv \mathbb{R}^2$;

$A2$ holds with $\alpha = 0$, $c = \sqrt{2}$, and $U = \mathbb{R}^2$, $V = \mathbb{R}$;

$A3$ holds because $f_1 \in \mathfrak{L}^{0,1}(\mathbb{R}^2, \mathbb{R})$ with $k_{f_1} = 2$.

The Lipschitz constant provided by Proposition 2.4 is $\frac{k_{f_1}}{c} = \sqrt{2}$. □

The next example shows that studying the generalized Euler equation may sometimes be inadequate for obtaining Aubin continuity of the solution map. This is because the set of the stationary points may be larger than the set of minima.

Example 2.6. Let $K = \mathbb{R}^2$ and

$$f(x_1, x_2, \lambda) := (x_1 + \lambda x_2 - 1)^2 (x_2 + \lambda x_1 + 1)^2,$$

$x_1, x_2, \lambda \in \mathbb{R}$. Consider the parameterized family of unconstrained minimization problems over the plane

$$P(\lambda) \qquad \inf_{x_1,x_2} f(x_1, x_2, \lambda).$$

Then at the point $\overline{\lambda} = 1$ the set of solutions $S(\lambda)$ is smaller than the set of stationary points $St(\lambda) := \{x \in \mathbb{R}^2 : 0 \in \nabla_x f(x, \lambda)\}$. Moreover, the map S is Aubin continuous near any point in his graph while St is not Aubin continuous near the point $(\overline{\lambda}, \overline{x}) \in \operatorname{gph} St$ where $\overline{\lambda} = 1$ and $\overline{x} = (0, 0)$.

Proof. Straightforward computations show that for any $\lambda \in \mathbb{R}$ the solution set

$$S(\lambda) = \{(x_1, x_2) : x_2 + \lambda x_1 = -1, \text{ or } x_1 + \lambda x_2 = 1\}$$

is the union of two lines—the line $p_1(\lambda)$ with equation $x_2 + \lambda x_1 = -1$ and the line $p_2(\lambda)$ with equation $x_1 + \lambda x_2 = 1$. Because of

$$\begin{aligned} \nabla_x f(x, \lambda) = [&2(x_1 + \lambda x_2 - 1)(x_2 + \lambda x_1 + 1)(2\lambda x_1 + (1 + \lambda^2)x_2 + 1 - \lambda), \\ &2(x_1 + \lambda x_2 - 1)(x_2 + \lambda x_1 + 1)((1 + \lambda)^2 x_1 + 2\lambda x_2 + \lambda - 1)], \end{aligned}$$

the set of the stationary points at $\overline{\lambda} = 1$ consists of three parallel lines

$$St(1) = \{(x_1, x_2) : x_1 + x_2 = 1, \text{ or } x_1 + x_2 = -1, \text{ or } x_1 + x_2 = 0\},$$

while for $\lambda \neq 1$, $St(\lambda) \equiv S(\lambda)$.

It is not difficult to see that S is Aubin continuous near an arbitrary point $(\lambda, x) \in \operatorname{gph} S$ (we note, by the way, that S is not Lipschitz continuous). Indeed, fix $\tilde{\lambda} \in \mathbb{R}$ and take $\tilde{x} = (\tilde{x}_1, \tilde{x}_2) \in S(\tilde{\lambda}) = p_1(\tilde{\lambda}) \cup p_2(\tilde{\lambda})$. Obviously, $\tilde{x} \neq 0$.

Take λ such that $|\lambda-\tilde{\lambda}|<1/2$. If $\tilde{x}\in p_1(\tilde{\lambda})$, then

$$d(\tilde{x},S(\lambda))\le d(\tilde{x},p_1(\lambda))=\frac{\left|\left(\lambda-\tilde{\lambda}\right)\tilde{x}_1\right|}{\sqrt{1+\lambda^2}}\le\left|\lambda-\tilde{\lambda}\right|\,|\tilde{x}_1|\,,$$

and if $\tilde{x}\in p_2(\tilde{\lambda})$, then

$$d(\tilde{x},S(\lambda))\le d(\tilde{x},p_2(\lambda))=\frac{\left|\left(\lambda-\overline{\lambda}\right)\tilde{x}_2\right|}{\sqrt{1+\lambda^2}}\le\left|\lambda-\tilde{\lambda}\right|\,|\tilde{x}_2|\,,$$

which yields

$$d(\tilde{x},S(\lambda))\le\left|\lambda-\tilde{\lambda}\right|\max\{|\tilde{x}_1|,|\tilde{x}_2|\}\le\left|\lambda-\tilde{\lambda}\right|\,\|\tilde{x}\|<\|\tilde{x}\|/2.$$

This implies that for all λ such that $|\lambda-\tilde{\lambda}|<1/2$ the intersection of $S(\lambda)$ with the neighborhood $U:=\tilde{x}+\|\tilde{x}\|B^\circ$ is nonempty.

Take $x=(x_1,x_2)\in S(\lambda)\cap U$ and μ such that $|\mu-\tilde{\lambda}|<1/2$. Similarly we get

$$d(x,S(\mu))\le|\lambda-\mu|\,\|x\|\le|\lambda-\mu|\,[\|x-\tilde{x}\|+\|\tilde{x}\|]\le2\|\tilde{x}\|\,|\lambda-\mu|.$$

Hence,

$$e(S(\lambda)\cap U,S(\mu))\le2\|\tilde{x}\|\,|\lambda-\mu|,\quad\forall\lambda,\mu\in\tilde{\lambda}+\frac{1}{2}B^\circ,$$

which means that S is Aubin continuous near $(\tilde{\lambda},\tilde{x})\in\operatorname{gph}S$.

In contrast, St is not Aubin continuous near the point, $(\overline{\lambda},\overline{x})\in\operatorname{gph}St$ where $\overline{\lambda}=1$ and $\overline{x}=(0,0)$. Indeed, if St is Aubin continuous near that point, then $d(\overline{x},St(\lambda))$ tends to zero as λ tends to 1. But the distance

$$d(\overline{x},St(\lambda))=\min\{d(\overline{x},p_1(\lambda)),d(\overline{x},p_2(\lambda))\}=\frac{1}{\sqrt{1+\lambda^2}}$$

tends to $\frac{1}{\sqrt{2}}$ as λ tends to 1, which means that St is not Aubin continuous near $(\overline{\lambda},\overline{x})\in\operatorname{gph}St$. □

As an immediate consequence of Proposition 2.4 we get the following.

COROLLARY 2.7. *Let for the parameterized family of minimization problems $P(\lambda)$ the following assumption hold*

$$(A')\qquad\begin{cases}\text{for all }\lambda\in\Lambda,\ \text{all }x\in K,\ \text{and some }c>0\\ 1.\ S(\lambda)\neq\emptyset;\\ 2.\ f(x,\lambda)\ge m(\lambda)+cd^2(x,S(\lambda));\\ 3.\ f\in C^{1,1}(X\times Z).\end{cases}$$

Then the solution map $S:\Lambda\rightrightarrows X$ is Lipschitz continuous on Λ.

In a Banach space X with separable dual X^* the notion of a second-order subdifferential for a function $f\in C^{1,1}(X)$ is introduced in [16] (see also the previous work [18] for the finite dimensional case). For any $x\in X$ the second-order subdifferential $\partial^2f(x)$ of f at x is a nonempty, convex, and w^* compact set in $\mathcal{L}(X\times X)$ (the Banach space of all bilinear continuous functionals $M:X\times X\to\mathbb{R}$ with the norm $\|M\|:=\sup_{\|h_1\|=\|h_2\|=1}|M[h_1,h_2]|$), which is singleton exactly when f is twice strictly Gâteaux differentiable at x.

Setting a simple condition on the second subdifferential is sufficient to get a family of functions satisfying assumption A' in the above corollary.

Indeed, let X be a Banach space with separable dual. Let in the parameterized family of minimization problems $P(\lambda)$, $f \in C^{1,1}(X \times Z)$, and let the constraint set K be closed and convex. If there exist $c > 0$ with

$$\langle M(y-x), y-x\rangle \geq c\|y-x\|^2 \quad \text{for all } \lambda \in \Lambda,\ x, y \in K,\ M \in \partial^2 f(\cdot,\lambda)(x), \tag{2.11}$$

then the solution map $S : \Lambda \to X$ will be single-valued and Lipschitz continuous on Λ.

It is easily seen that (2.11) implies uniform on $\lambda \in \Lambda$ strong convexity of $f(\cdot,\lambda)$ on K. By this and continuity of $f(\cdot,\lambda)$, for every λ the infimum of $f(\cdot,\lambda)$ is attained at unique $x_\lambda \in K$ and $A'1$ holds.

For any $x \in K$ and $\lambda \in \Lambda$ there exists some $z_\lambda \in K$ and $M_{z_\lambda} \in \partial^2 f(\cdot,\lambda)(z_\lambda)$ with

$$f(x,\lambda) = f(x_\lambda,\lambda) + \langle \nabla_x f(x_\lambda,\lambda), x - x_\lambda\rangle + \frac{1}{2}\langle M_{z_\lambda}(x - x_\lambda), x - x_\lambda\rangle$$

(see [16]). Since x_λ is a minimum point for $f(\cdot,\lambda)$ on K and K is convex, then for all $x \in K$, $\langle \nabla_x f(x_\lambda,\lambda), x - x_\lambda\rangle \geq 0$ and from above equality and (2.11)

$$f(x,\lambda) \geq m(\lambda) + \frac{1}{2}c\|x - x_\lambda\|^2,$$

so $A'2$ holds with $\alpha = 1$.

We will use Corollary 2.7 to obtain existence and Lipschitz continuity of the optimal solution for a linearly perturbed optimization problem, assuming a slightly weaker version (see (2.12) below) of the uniform second-order growth condition (Definition 5.19 in [7]), and $C^{1,1}$ data. In this way we extend [7], Theorem 5.17 (see also [7], Remark 5.19), where C^2 data are assumed.

Recall that the Banach space X has *Radon–Nikodym property* (RNP) if for every bounded set C and every $\varepsilon > 0$, there exists an $x \in C$ that does not belong to the closed convex hull of $C \setminus \{x + \varepsilon B^\circ\}$. All Banach spaces which have separable dual and all reflexive Banach spaces have RNP. In [13], p. 157 there is a long list of equivalent definitions of RNP. A good introductory survey on RNP is [14].

An efficient tool in dealing with minimization problems on Banach space X with RNP is *Stegall's variational principle* [26] (see also [21], Theorem 5.15): Let $C \subset X$ be a non-empty closed and bounded convex set and let $f : C \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous function, bounded below on C, then for every $\varepsilon > 0$, there exists $x^* \in X^*$ with $\|x^*\| \leq \varepsilon$ such that $f + x^*$ attains its strong minimum on C. Let us remind that $x_0 \in C$ is said to be a *strong minimum* for function $g : C \to \mathbb{R} \cup \{+\infty\}$ on the set C if $g(x_0) = \inf_C g$ and $\|x_n - x_0\| \to 0$ whenever $g(x_n) \to g(x_0)$.

Corollary 2.8. *Let the Banach space X have Radon–Nikodym property. Consider a parameterized family of minimization problems $P(\lambda)$, where the parameter space is X^* and $f : X \times X^* \to \mathbb{R}$ is defined by $f(x,\lambda) := f(x) + \langle \lambda, x\rangle$.*

Assume that the constraint set K is closed and convex, $f \in C^{1,1}(X)$, and $S(0)$ is nonempty.

Suppose that there exist neighborhood V of the origin 0 of X^* and a constant $c > 0$ such that for all $\lambda \in V$ and all $x_\lambda \in S(\lambda)$ it holds that

$$f(x,\lambda) \geq f(x_\lambda,\lambda) + c\|x - x_\lambda\|^2, \quad \forall x \in K. \tag{2.12}$$

Then there exists a neighborhood W of the origin 0 of X^* such that $S(\lambda)$ is single-valued and Lipschitz continuous on W.

Proof. From (2.12) it is clear that $S(\lambda)$ contains at most one point for $\lambda \in V$.

We will show that $S(\lambda)$ is nonempty for λ belonging to some neighborhood of 0. Fix $\gamma > 0$ such that $W := 2\gamma B^\circ \subset V$.

Given $\lambda \in \gamma B^\circ$, let ε_k be a sequence of positive numbers less than γ, tending to zero.

Thanks to (2.12) with $\lambda = 0$, $f(\cdot, \lambda)$ is bounded below on K.

If K is bounded we could apply directly Stegall's variational principle for the function $f(\cdot, \lambda) : K \to \mathbb{R}$ and ε_k to find $x_k^* \in X^*$ with $\|x_k^*\| \le \varepsilon_k$ and a strong minimum x_k of $f(\cdot, \lambda) + x_k^*$ on K.

If K is not bounded, a variant of Stegall's variational principle still holds thanks to (2.12). Indeed, (2.12) for $\lambda = 0$ reads

$$f(x) \ge f(x_0) + c\|x - x_0\|^2, \quad \forall x \in K,$$

which yields that for all $x \in K$,

$$\begin{aligned} f(x, \lambda) &\ge f(x_0) + \langle \lambda, x\rangle + c\|x - x_0\|^2 = f(x_0, \lambda) + \langle \lambda, x - x_0\rangle + c\|x - x_0\|^2 \\ &\ge f(x_0, \lambda) + \|x - x_0\|[c\|x - x_0\| - \|\lambda\|] \\ &\ge f(x_0, \lambda) + \|x - x_0\|[c\|x - x_0\| - \gamma]. \end{aligned} \tag{2.13}$$

Set $r := \frac{3\gamma}{c}$. Now, we apply Stegall's variational principle for the function $f(\cdot, \lambda)$ on the closed bounded set $K \cap \{x_0 + rB\}$ and ε_k. Thus, there exists $x_k^* \in X^*$, $\|x_k^*\| < \varepsilon_k$, and a point $x_k \in K \cap \{x_0 + rB\}$ such that $f(\cdot, \lambda) + x_k^*$ attains a strong minimum on $K \cap \{x_0 + rB\}$ at x_k. Moreover, x_k is a strong minimum of $f(\cdot, \lambda) + x_k^*$ on K. Indeed, if we assume that $x \in K$ is such that

$$f(x, \lambda) + \langle x_k^*, x\rangle \le f(x_k, \lambda) + \langle x_k^*, x_k\rangle = \inf_{K \cap \{x_0 + rB\}} f(\cdot, \lambda) + x_k^* \le f(x_0, \lambda) + \langle x_k^*, x_0\rangle,$$

then by (2.13) we will have

$$\|x - x_0\|[c\|x - x_0\| - \gamma] \le \|x_k^*\|\|x - x_0\| \le \varepsilon_k\|x - x_0\| < \gamma\|x - x_0\|,$$

or

$$\|x - x_0\| \le \frac{2\gamma}{c} < r,$$

which means that $x \in x_0 + rB$ and clearly entails $x = x_k$.

However, in both cases for any k we found $x_k^* \in X^*$ with $\|x_k^*\| \le \varepsilon_k$ and unique $x_k \in K$ satisfying

$$f(x) + \langle \lambda + x_k^*, x\rangle \ge f(x_k) + \langle \lambda + x_k^*, x_k\rangle, \quad \forall x \in K.$$

This means that $S(\lambda + x_k^*) = \{x_k\}$ and since $\lambda + x_k^* \in V$, (2.12) reads

$$f(x) + \langle \lambda + x_k^*, x\rangle \ge f(x_k) + \langle \lambda + x_k^*, x_k\rangle + c\|x - x_k\|^2, \quad \forall x \in K. \tag{2.14}$$

Substitute $x = x_n$ and rearrange to obtain

$$f(x_n) - f(x_k) \ge \langle \lambda + x_k^*, x_k - x_n\rangle + c\|x_n - x_k\|^2.$$

Also, swapping k and n we get

$$f(x_k) - f(x_n) \ge \langle \lambda + x_n^*, x_n - x_k\rangle + c\|x_n - x_k\|^2.$$

Adding the above two, we obtain $2c\|x_n - x_k\|^2 \le \langle x_k^* - x_n^*, x_n - x_k\rangle \le (\|x_k^*\| + \|x_n^*\|)\|x_n - x_k\|$. That is, $2c\|x_n - x_k\| \le \varepsilon_k + \varepsilon_n$, which means that x_k is a Cauchy sequence. Let $x_\lambda \in K$ be its limit. Passing to limit in (2.14) we see that $x_\lambda \in S(\lambda)$.

A straightforward application of Corollary 2.7 completes the proof. ∎

3. Parameterized minimax problem. In this section we study the behavior of the saddle points set of a parameterized family of minimax problems.

3.1. Preliminaries and statement of the problem. Let X and Y be Banach spaces, and let $\{f(\cdot,\cdot,\lambda) : X \times Y \to \mathbb{R}, \lambda \in \Lambda\}$ be a family of functions defined on the product space $X \times Y$, parameterized by $\lambda \in \Lambda \subset Z$.

Let us consider the parameterized family of minimax problems

$$M(\lambda) \qquad \inf_{x\in K} \sup_{y\in L} f(x,y,\lambda),$$

where the constraints are nonempty closed sets $K \subset X$ and $L \subset Y$. Denote the optimal value of $M(\lambda)$ by $\mathfrak{m}(\lambda)$ and recall that the (possibly empty) set of saddle points of $f(\cdot,\cdot,\lambda)$ on $K \times L$ is given by (1.1).

For a set $C \subset X \times Y$ we denote by $\pi_X C$ and $\pi_Y C$ the *projections* of C on the spaces X and Y, respectively. More precisely, $x \in \pi_X C$ whenever there exists some $y \in Y$ with $(x,y) \in C$ and $y \in \pi_Y C$ whenever there exists some $x \in X$ with $(x,y) \in C$.

It is well known that the saddle point set is a product set; that is,

$$\mathcal{S}(\lambda) = \pi_X \mathcal{S}(\lambda) \times \pi_Y \mathcal{S}(\lambda). \tag{3.1}$$

To the parameterized family of functions $\{f(\cdot,\cdot,\lambda), \lambda \in \Lambda\}$ one naturally associates three difference functions:

$$\begin{aligned}
\mathfrak{f}_1(x,x',\lambda,y) &:= f(x,y,\lambda) - f(x',y,\lambda),\\
\mathfrak{f}_2(y,y',\lambda,x) &:= f(x,y,\lambda) - f(x,y',\lambda),\\
\mathfrak{f}_3(\lambda,\lambda',x,y) &:= f(x,y,\lambda) - f(x,y,\lambda').
\end{aligned}$$

By analogy with Definition 2.1 we write $\mathfrak{f}_1 \in \mathfrak{L}_W^{\alpha,\beta}(U;V)$ whenever the functions $f_1^y(x,x',\lambda) := \mathfrak{f}_1(x,x',\lambda,y)$ are such that for all $y \in W$, $f_1^y \in \mathfrak{L}^{\alpha,\beta}(U;V)$, and $\sup_{y\in W} k_{f_1^y}$ is finite. We set $k_{\mathfrak{f}_1} := \sup_{y\in W} k_{f_1^y}$.

Easy computations as those done in Proposition 2.2 show that $\mathfrak{f}_1 \in \mathfrak{L}_W^{\alpha,\beta}(U;V)$ exactly when $\mathfrak{f}_3 \in \mathfrak{L}_W^{\beta,\alpha}(V;U)$ and that $\mathfrak{f}_2 \in \mathfrak{L}_U^{\alpha,\beta}(W;V)$ exactly when $\mathfrak{f}_3 \in \mathfrak{L}_U^{\beta,\alpha}(V;W)$.

Now we are ready to state the sufficient condition for Aubin continuity of the saddle points map $\mathcal{S} : \Lambda \rightrightarrows X \times Y$.

Let $(\overline{\lambda},\overline{x},\overline{y}) \in \operatorname{gph}\mathcal{S}$. We set the following local assumption $\mathcal{A}$ at $(\overline{\lambda},\overline{x},\overline{y})$:

$$(\mathcal{A}) \quad \begin{cases}
\text{there exist neighborhoods } U \text{ of } \overline{x},\ W \text{ of } \overline{y}, \text{ and } V \text{ of } \overline{\lambda}, \text{such that}\\
1.\quad \mathcal{S}(\lambda) \cap [U \times W] \neq \emptyset \text{ for all } \lambda \in \Lambda \cap V;\\
\text{and there exist constants } c > 0 \text{ and } \alpha \in [0,1] \text{ such that}\\
2.\quad f(x,y',\lambda') \geq \mathfrak{m}(\lambda') + c d^{1+\alpha}(x, \pi_X \mathcal{S}(\lambda')),\\
\qquad f(x',y,\lambda') \leq \mathfrak{m}(\lambda') - c d^{1+\alpha}(y, \pi_Y \mathcal{S}(\lambda')),\\
\qquad \forall \lambda,\lambda' \in \Lambda \cap V, \forall (x,y) \in \mathcal{S}(\lambda) \cap [U \times W], \forall (x',y') \in \mathcal{S}(\lambda');\\
3.\ \mathfrak{f}_1 \in \mathfrak{L}_{L\cap W}^{\alpha,1}(K;\Lambda\cap V) \text{ and } \mathfrak{f}_2 \in \mathfrak{L}_{K\cap U}^{\alpha,1}(L;\Lambda\cap V).
\end{cases}$$

Clearly, condition $\mathcal{A}3$ could be replaced by

$$\mathfrak{f}_3 \in \mathfrak{L}_{L\cap W}^{1,\alpha}(\Lambda\cap V;K) \cap \mathfrak{L}_{K\cap U}^{1,\alpha}(\Lambda\cap V;L).$$

$\mathcal{A}1$ implies that $\mathfrak{m}(\lambda)$ is finite for $\lambda \in \Lambda \cap V$.

We would show the consistency of our main hypothesis by giving some examples of parameterized families of functions $\{f(\cdot,\cdot,\lambda), \lambda \in \Lambda\}$ for which $\mathcal{A}$ is satisfied.

One gets a parameterized family of functions $\{f(\cdot,\cdot,\lambda), \lambda\in\Lambda\}$ satisfying $\mathcal{A}2$, for example, by assuming that $K \subset X$ and $L \subset Y$ are nonempty closed convex sets; the function $f(\cdot, y, \lambda)$ is lower semicontinuous and uniformly on $(y,\lambda) \in L \times \Lambda$ strongly convex on K, i.e., such that for some constant $c > 0$ the inequality

$$f(tx + (1-t)x', y, \lambda) \le tf(x, y, \lambda) + (1-t)f(x', y, \lambda) - ct(1-t)\|x - x'\|^2$$

holds for every $t \in [0,1]$, every $x, x' \in K$, and every $(y,\lambda) \in L \times \Lambda$; the function $f(x,\cdot,\lambda)$ is upper semicontinuous and uniformly on $(x,\lambda) \in K \times \Lambda$ strongly concave on L, i.e., such that the inequality

$$f(x, ty + (1-t)y', \lambda) \ge tf(x, y, \lambda) + (1-t)f(x, y', \lambda) + ct(1-t)\|y - y'\|^2$$

holds for every $t \in [0,1]$, every $y, y' \in L$, and every $(x,\lambda) \in K \times \Lambda$.

Then it is routine to see that $\mathcal{A}2$ holds at any $(\overline{\lambda}, \overline{x}, \overline{y}) \in \operatorname{gph}\mathcal{S}$ with $\alpha = 1$, $V = Z$, $U = X$, and $W = Y$.

Examples of parameterized families of functions satisfying $\mathcal{A}3$ are given by the following.

LEMMA 3.1. *Let $(\overline{\lambda}, \overline{x}, \overline{y}) \in \operatorname{gph}\mathcal{S}$ and let $U \subset X$, $W \subset Y$ and $V \subset Z$ be convex neighborhoods of K, L, and $\overline{\lambda}$, respectively. Consider the conditions:*

$(\mathcal{F}1)$ $\begin{cases} \text{for any } (y,\lambda) \in L \times [\Lambda \cap V],\ f(\cdot, y, \lambda) \text{ is Lipschitz and regular} \\ \text{function on } U \text{ and } \partial_x f(x, y, \cdot) : \Lambda \cap V \to X^* \text{ is a } k\text{–Lipschitz map on} \\ \Lambda \cap V \text{ with } k \text{ that does not depend on } (x,y) \in K \times L, \end{cases}$

$(\mathcal{F}2)$ $\begin{cases} \text{for any } (x,\lambda) \in K \times [\Lambda \cap V],\ f(x, \cdot, \lambda) \text{ is Lipschitz and regular} \\ \text{function on } W \text{ and } \partial_y f(x, y, \cdot) : \Lambda \cap V \to Y^* \text{ is a } k\text{–Lipschitz map on} \\ \Lambda \cap V \text{ with } k \text{ that does not depend on } (x,y) \in K \times L, \end{cases}$

$(\mathcal{F}3)$ $\begin{cases} \text{for any } (x,y) \in K \times L,\ f(x, y, \cdot) \text{ is Lipschitz and regular} \\ \text{function on } V \text{ and } \partial_\lambda f(\cdot, y, \lambda) : K \to \Lambda^* \text{ is a } k\text{–Lipschitz map on} \\ K \text{ with } k \text{ that does not depend on } (y,\lambda) \in L \times [\Lambda \cap V], \end{cases}$

$(\mathcal{F}4)$ $\begin{cases} \text{for any } (x,y) \in K \times L,\ f(x, y, \cdot) \text{ is Lipschitz and regular} \\ \text{function on } V \text{ and } \partial_\lambda f(x, \cdot, \lambda) : L \to \Lambda^* \text{ is a } k\text{–Lipschitz map on} \\ L \text{ with } k \text{ that does not depend on } (x,\lambda) \in K \times [\Lambda \cap V]. \end{cases}$

If f satisfies $\mathcal{F}1 - \mathcal{F}2$ or $\mathcal{F}3 - \mathcal{F}4$, then $\mathcal{A}3$ holds with $\alpha = 1$.

Proof. We follow the same steps as in the proof of Lemma 2.3.

If f satisfies $\mathcal{F}1$, then $\mathfrak{f}_1 \in \mathfrak{L}^{1,1}_L(K; \Lambda \cap V)$.

If f satisfies $\mathcal{F}2$, then $\mathfrak{f}_2 \in \mathfrak{L}^{1,1}_K(L; \Lambda \cap V)$.

If f satisfies $\mathcal{F}3$, then $\mathfrak{f}_3 \in \mathfrak{L}^{1,1}_L(\Lambda \cap V; K)$.

If f satisfies $\mathcal{F}4$, then $\mathfrak{f}_3 \in \mathfrak{L}^{1,1}_K(\Lambda \cap V; L)$. ∎

Obviously, if $f \in C^{1,1}(U \times W \times V)$, then $\mathcal{F}1$ to $\mathcal{F}4$ hold.

3.2. Lipschitz-like continuity of the saddle point map. Here we will prove that assumption $\mathcal{A}$ is sufficient for Aubin continuity of the saddle point map $\mathcal{S}$. Let us note that the result cannot be derived (or at least not in an obvious manner) from the case of minimization only. Indeed, if $f(x, y, \lambda)$ satisfies assumption $\mathcal{A}$, then the function $f(x,\lambda) := \sup_{y\in L} f(x, y, \lambda)$ satisfies assumption $A2$ but $A3$ for this $f(x,\lambda)$ cannot be derived from $\mathcal{A}3$ since the differences of suprema involved do not yield themselves to rearrangement.

THEOREM 3.2. *Assume that for the parameterized family of minimax problems $M(\lambda)$ the assumption $\mathcal{A}$ holds at some $(\overline{\lambda},\overline{x},\overline{y}) \in \operatorname{gph}\mathcal{S}$. Then for all $\lambda,\mu \in \Lambda \cap V$*

$$e(\mathcal{S}(\lambda)\cap[U\times W],\mathcal{S}(\mu)) \le \frac{2k}{c}\|\lambda-\mu\|, \tag{3.2}$$

where $k := \max\{k_{\mathfrak{f}_1}, k_{\mathfrak{f}_2}\}$, hence the saddle point map $\mathcal{S} : \Lambda \rightrightarrows X \times Y$ is Aubin continuous near $(\overline{\lambda},\overline{x},\overline{y}) \in \operatorname{gph}\mathcal{S}$.

Proof. By $\mathcal{A}1$ for all $\lambda \in \Lambda \cap V$ the set $\mathcal{S}(\lambda)\cap[U\times W]$ is nonempty.

Fix $\lambda \in \Lambda \cap V$ and take some $(x_\lambda, y_\lambda) \in \mathcal{S}(\lambda)\cap[U\times W]$.

Pick any other $\mu \in \Lambda \cap V$.

Since $\mathcal{S}(\mu)$ is a nonempty set we find some $x^\varepsilon_\mu \in \pi_X\mathcal{S}(\mu)$ such that

$$\|x_\lambda - x^\varepsilon_\mu\| \le d(x_\lambda, \pi_X\mathcal{S}(\mu)) + \varepsilon.$$

Similarly, there is $y^\varepsilon_\mu \in \pi_Y\mathcal{S}(\mu)$ such that

$$\|y_\lambda - y^\varepsilon_\mu\| \le d(y_\lambda, \pi_Y\mathcal{S}(\mu)) + \varepsilon.$$

By the product form of the saddle point set, $(x^\varepsilon_\mu, y^\varepsilon_\mu) \in \mathcal{S}(\mu)$. The first inequality of $\mathcal{A}2$ for $(x_\lambda, y_\lambda) \in \mathcal{S}(\lambda)\cap[U\times W]$ and $(x^\varepsilon_\mu, y^\varepsilon_\mu) \in \mathcal{S}(\mu)$ reads

$$f\left(x_\lambda, y^\varepsilon_\mu, \mu\right) \ge \mathfrak{m}(\mu) + cd^{1+\alpha}(x_\lambda, \pi_X\mathcal{S}(\mu)), \tag{3.3}$$

in particular,

$$f\left(x_\lambda, y^\varepsilon_\mu, \mu\right) \ge \mathfrak{m}(\mu), \tag{3.4}$$

while the second inequality of $\mathcal{A}2$ states

$$\mathfrak{m}(\mu) \ge f\left(x^\varepsilon_\mu, y_\lambda, \mu\right) + cd^{1+\alpha}(y_\lambda, \pi_Y\mathcal{S}(\mu)), \tag{3.5}$$

in particular,

$$\mathfrak{m}(\mu) \ge f\left(x^\varepsilon_\mu, y_\lambda, \mu\right). \tag{3.6}$$

Combining (3.3) with (3.6) and (3.4) with (3.5), we get

$$\begin{aligned} f\left(x_\lambda, y^\varepsilon_\mu, \mu\right) &\ge f\left(x^\varepsilon_\mu, y_\lambda, \mu\right) + cd^{1+\alpha}(x_\lambda, \pi_X\mathcal{S}(\mu)),\\ f\left(x_\lambda, y^\varepsilon_\mu, \mu\right) &\ge f\left(x^\varepsilon_\mu, y_\lambda, \mu\right) + cd^{1+\alpha}(y_\lambda, \pi_Y\mathcal{S}(\mu)), \end{aligned}$$

which yields

$$f\left(x_\lambda, y^\varepsilon_\mu, \mu\right) - f\left(x^\varepsilon_\mu, y_\lambda, \mu\right) \ge c[\max\{d(x_\lambda, \pi_X\mathcal{S}(\mu)), d(y_\lambda, \pi_Y\mathcal{S}(\mu))\}]^{1+\alpha}.$$

By the definition of the supremum norm and since $\mathcal{S}(\mu)$ is a product set, it is obvious that

$$\begin{aligned} d((x_\lambda, y_\lambda), \mathcal{S}(\mu)) &= d((x_\lambda, y_\lambda), \pi_X\mathcal{S}(\mu)\times\pi_Y\mathcal{S}(\mu))\\ &= \max\{d(x_\lambda, \pi_X\mathcal{S}(\mu)), d(y_\lambda, \pi_Y\mathcal{S}(\mu))\}, \end{aligned} \tag{3.7}$$

and the above inequality can be rewritten as

$$f\left(x_\lambda, y^\varepsilon_\mu, \mu\right) - f\left(x^\varepsilon_\mu, y_\lambda, \mu\right) \ge cd^{1+\alpha}((x_\lambda, y_\lambda), \mathcal{S}(\mu)).$$

We transform the left-hand side to get

$$f\left(x_\lambda, y_\mu^\varepsilon, \mu\right) - f(x_\lambda, y_\lambda, \mu) + f(x_\lambda, y_\lambda, \mu) - f\left(x_\mu^\varepsilon, y_\lambda, \mu\right) \geq cd^{1+\alpha}((x_\lambda, y_\lambda), \mathcal{S}(\mu)),$$

which is

$$-\mathfrak{f}_2\left(x_\lambda, y_\lambda, y_\mu^\varepsilon, \mu\right) - \mathfrak{f}_1\left(x_\mu^\varepsilon, x_\lambda, y_\lambda, \mu\right) \geq cd^{1+\alpha}((x_\lambda, y_\lambda), \mathcal{S}(\mu)). \tag{3.8}$$

On the other hand, since $(x_\lambda, y_\lambda) \in \mathcal{S}(\lambda)$, the saddle point inequalities give

$$f(x, y_\lambda, \lambda) \geq f(x_\lambda, y_\lambda, \lambda) \geq f(x_\lambda, y, \lambda), \quad \forall x \in K, \forall y \in L.$$

In particular, for $x = x_\mu^\varepsilon \in K$ we have

$$f(x_\mu^\varepsilon, y_\lambda, \lambda) \geq f(x_\lambda, y_\lambda, \lambda),$$

which is

$$\mathfrak{f}_1(x_\mu^\varepsilon, x_\lambda, y_\lambda, \lambda) \geq 0, \tag{3.9}$$

and for $y = y_\mu^\varepsilon \in L$ we get

$$f(x_\lambda, y_\lambda, \lambda) \geq f(x_\lambda, y_\mu^\varepsilon, \lambda),$$

which is

$$\mathfrak{f}_2(x_\lambda, y_\lambda, y_\mu^\varepsilon, \lambda) \geq 0. \tag{3.10}$$

Adding the inequalities (3.8), (3.9), and (3.10) and rearranging we obtain

$$\begin{aligned}\left[\mathfrak{f}_1\left(x_\mu^\varepsilon, x_\lambda, y_\lambda, \lambda\right) - \mathfrak{f}_1\left(x_\mu^\varepsilon, x_\lambda, y_\lambda, \mu\right)\right] + \left[\mathfrak{f}_2\left(x_\lambda, y_\lambda, y_\mu^\varepsilon, \lambda\right) - \mathfrak{f}_2\left(x_\lambda, y_\lambda, y_\mu^\varepsilon, \mu\right)\right] \\ \geq cd^{1+\alpha}((x_\lambda, y_\lambda), \mathcal{S}(\mu)).\end{aligned} \tag{3.11}$$

Since by $\mathcal{A}3$, $\mathfrak{f}_1 \in \mathfrak{L}^{\alpha,1}_{L\cap W}(K; \Lambda \cap V)$, the term in first brackets in (3.11) is estimated by

$$\mathfrak{f}_1\left(x_\mu^\varepsilon, x_\lambda, y_\lambda, \mu\right) - \mathfrak{f}_1\left(x_\mu^\varepsilon, x_\lambda, y_\lambda, \lambda\right) \leq k_{\mathfrak{f}_1}\|x_\mu^\varepsilon - x_\lambda\|^\alpha\|\lambda - \mu\|, \tag{3.12}$$

and since $\mathfrak{f}_2 \in \mathfrak{L}^{\alpha,1}_{K\cap U}(L; \Lambda \cap V)$ the term in second brackets in (3.11) is estimated by

$$\mathfrak{f}_2\left(x_\lambda, y_\lambda, y_\mu^\varepsilon, \lambda\right) - \mathfrak{f}_2\left(x_\lambda, y_\lambda, y_\mu^\varepsilon, \mu\right) \leq k_{\mathfrak{f}_2}\|y_\lambda - y_\mu^\varepsilon\|^\alpha\|\lambda - \mu\|. \tag{3.13}$$

Using (3.13) and (3.12) in (3.11) and setting $k := \max\{k_{\mathfrak{f}_1}, k_{\mathfrak{f}_2}\}$, we get

$$k\|\lambda - \mu\|[\|x_\lambda - x_\mu^\varepsilon\|^\alpha + \|y_\lambda - y_\mu^\varepsilon\|^\alpha] \geq cd^{1+\alpha}((x_\lambda, y_\lambda), \mathcal{S}(\mu)).$$

By the choice of x_μ^ε and y_μ^ε, we have that

$$\begin{aligned}k\|\lambda - \mu\|[(d(x_\lambda, \pi_X\mathcal{S}(\mu)) + \varepsilon)^\alpha + (d(y_\lambda, \pi_Y\mathcal{S}(\mu)) + \varepsilon)^\alpha] \\ \geq cd^{1+\alpha}((x_\lambda, y_\lambda), \mathcal{S}(\mu)).\end{aligned}$$

Passing to limit $\varepsilon \downarrow 0$ we obtain

$$k\|\lambda - \mu\|[d^\alpha(x_\lambda, \pi_X\mathcal{S}(\mu)) + d^\alpha(y_\lambda, \pi_Y\mathcal{S}(\mu))] \geq cd^{1+\alpha}((x_\lambda, y_\lambda), \mathcal{S}(\mu)). \tag{3.14}$$

By (3.7) we get

$$d^\alpha(y_\lambda, \pi_Y\mathcal{S}(\mu)) + d^\alpha(x_\lambda, \pi_X\mathcal{S}(\mu)) \le 2\left[\max\{d(y_\lambda, \pi_Y\mathcal{S}(\mu)), d(x_\lambda, \pi_X\mathcal{S}(\mu))\}\right]^\alpha$$
$$= 2d^\alpha((x_\lambda, y_\lambda), \mathcal{S}(\mu)),$$

and from (3.14) we obtain

$$2k\|\lambda - \mu\| d^\alpha((x_\lambda, y_\lambda), \mathcal{S}(\mu)) \ge c d^{1+\alpha}((x_\lambda, y_\lambda), \mathcal{S}(\mu)).$$

This yields

$$\frac{2k}{c}\|\lambda - \mu\| \ge d((x_\lambda, y_\lambda), \mathcal{S}(\mu)),$$

and since (x_λ, y_λ) was an arbitrary element of $\mathcal{S}(\lambda) \cap [U \times W]$ the latter implies

$$e(\mathcal{S}(\lambda) \cap [U \times W], \mathcal{S}(\mu)) \le \frac{2k}{c}\|\lambda - \mu\|.$$

The proof is completed. □

As an immediate consequence of Theorem 3.2 and Lemma 3.1 one deduces the following.

COROLLARY 3.3. *Let for the parameterized family of minimax problems $M(\lambda)$ the following assumption hold:*

$$(\mathcal{A}') \quad \begin{cases} 1. & \mathcal{S}(\lambda) \ne \emptyset \text{ for any } \lambda \in \Lambda; \\ 2. & \text{for some constant } c > 0 \text{ and all } \lambda \in \Lambda, (x,y) \in \mathcal{S}(\lambda), (x',y') \in K \times L: \\ & f(x', y, \lambda) \ge \mathfrak{m}(\lambda) + c d^2(x', \pi_X\mathcal{S}(\lambda)), \\ & f(x, y', \lambda) \le \mathfrak{m}(\lambda) - c d^2(y', \pi_Y\mathcal{S}(\lambda)); \\ 3. & f \in C^{1,1}(X \times Y \times Z). \end{cases}$$

Then the saddle point map $\mathcal{S} : \Lambda \to X \times Y$ is single-valued and Lipschitz continuous.

As we pointed out after Corollary 2.7 we could deduce the single-valuedness and Lipschitz continuity of the saddle point map $\mathcal{S}$ when X and Y has separable duals, the sets K and L are convex, $f \in C^{1,1}(X \times Y \times Z)$, and there exists a constant $c > 0$ such that for all $\lambda \in \Lambda$, $x, z \in K$, $y, w \in L$,

$$\langle M(z-x), z-x\rangle \ge c\|z-x\|^2, \quad \langle N(w-y), w-y\rangle \le -c\|w-y\|^2$$

for all $M \in \partial^2 f(\cdot, y, \lambda)(x)$ and all $N \in \partial^2 f(x, \cdot, \lambda)(y)$.

4. Lipschitz continuity of the saddle points map in context of two-player zero sum differential games. In this section we briefly consider a differential game for which our result might be of relevance.

In differential games, open-loop strategies are of low interest in many examples. One major reason is that differential games with open-loop strategies do not satisfy, in general, the dynamic programming principle [9, 11, 22]. It is well known now that to solve many problems in differential games (existence of a value, characterization of the game through Hamilton–Jacobi equations), one needs a more general class of strategies which contains the feedback strategies.[1] Such class of strategies is the

[1] It has been shown in [8] that the class of regular feedback is not rich enough to solve differential games at a satisfactory level of generality.

class of nonanticipative strategies introduced by Elliot–Roxin–Varaiya–Kalton (cf. for instance [10]); another possible class of strategies are the positional strategies discussed in [20]. The class of nonanticipative strategies is nice enough to prove the existence of the value, but it is hard to implement for the players. So it is important to know when the nonanticipative strategies giving the value of the game can be reduced to feedback strategies. We will explain in this part how the main result of the paper can lead to a partial answer to this question.

We consider the following differential game with dynamic described by the differential equation:

$$\begin{cases} x'(t) = f(x(t), u(t)), \;\; y'(t) \in g(y(t), v(t)), \\ u(t) \in U, \; v(t) \in V, \end{cases} \tag{4.1}$$

where $f : \mathbb{R}^n \times U \to \mathbb{R}^n$ and $g : \mathbb{R}^n \times V \to \mathbb{R}^n$ are (globally) Lipschitz, $U \subset \mathbb{R}^m$, $V \subset \mathbb{R}^p$ being the control sets of the players. The first player—Ursula, playing with u—wants to minimize a given cost. The goal of the second player—Victor, playing with v—is to maximize the cost

$$J(x_0, y_0, u(\cdot), v(\cdot)) := \int_0^\infty e^{-rt} l(x(t), y(t))\, dt,$$

where $(x(\cdot), y(\cdot))$ is the unique solution starting at $t = 0$ from (x_0, y_0) and $r > 0$ is fixed. Observe that the game is in a separable form, i.e., each player acts in his own dynamics. This is the case, for instance, for pursuit games. Moreover, the integral cost does not depend directly on the control but only on the trajectories.

We work here in the framework of the *nonanticipative strategies* (also called Varaiya–Roxin–Elliot–Kalton strategies). Let

$$\mathcal{U} = L^1([0, +\infty[, U), \;\; \mathcal{V} = L^1([0, +\infty[, V) \tag{4.2}$$

be the sets of time-measurable controls of the first (Ursula) and the second (Victor) player, respectively. We denote $t \mapsto (x(t, x_0, u(t)), y(t, y_0, v(t)))$ the solution of (4.1) starting at $t = 0$ from (x_0, y_0).

DEFINITION 4.1 (nonanticipative strategies). *A map $\alpha : \mathcal{V} \to \mathcal{U}$ is a nonanticipative strategy (for Ursula) if it satisfies the following condition: For any $s \geq 0$, for any $v_1(\cdot)$ and $v_2(\cdot)$ belonging to $\mathcal{V}$ such that $v_1(\cdot)$ and $v_2(\cdot)$ coincide almost everywhere on $[0, s]$, the images $\alpha(v_1(\cdot))$ and $\alpha(v_2(\cdot))$ coincide almost everywhere on $[0, s]$.*

Nonanticipative strategies $\beta : \mathcal{U} \to \mathcal{V}$ (for Victor) are defined in the symmetric way.

Assume now that f and g are continuous and Lipschitz with respect to x and y. Then, we know that the game has a value (cf. [11]), namely,

$$V(x_0, y_0) = \inf_\alpha \sup_{v \in \mathcal{V}} J(x_0, y_0, \alpha(v(\cdot)), v(\cdot)) = \sup_\beta \inf_{u \in \mathcal{U}} J(x_0, y_0, u(\cdot), \beta(u(\cdot))).$$

Let us denote by $R(t)$ the attainable set of the dynamics (4.1) at moment t; i.e.,

$$R(t) = \{(x(t), y(t)) \in \mathbb{R}^n \times \mathbb{R}^n : \exists\, u \in \mathcal{U}, v \in \mathcal{V} \text{ such that}$$
$$(x(\cdot), y(\cdot)) \text{ is the solution of (4.1) starting at } t = 0 \text{ from } (x_0, y_0)\}.$$

Now, suppose that U and V are convex and compact. Saddle point of the function $l(\cdot, \cdot)$ on $R(t)$ will be any point $(\overline{x}, \overline{y}) \in R(t)$ that satisfies

$$l(\overline{x}, y) \leq l(\overline{x}, \overline{y}) \leq l(x, \overline{y}), \quad \forall (x, y) \in R(t),$$

and, because of $e^{-rt} > 0$, the saddle points of $l(\cdot,\cdot)$ on $R(t)$ will be the same as the saddle points of $e^{-rt}l(\cdot,\cdot)$ on $R(t)$.

Let us denote the (possibly empty) set of all saddle points of the function $l(\cdot,\cdot)$ on $R(t)$ by $\mathcal{S}(t) := \{(\overline{x},\overline{y}) \in R(t) : l(\overline{x},y) \le l(\overline{x},\overline{y}) \le l(x,\overline{y}), \quad \forall (x,y) \in R(t)\}$.

Let us suppose that the parameterized by t family of functions $\{e^{-rt}l(\cdot,\cdot), t \in [0,\infty)\}$ satisfies an assumption slightly stronger than assumption $\mathcal{A}$, namely:

$$\begin{cases} 1. \quad \mathcal{S}(t) \neq \emptyset, \ \forall t \ge 0; \\ \text{and there exist constants } k, c > 0 \text{ and } \alpha \in [0,1] \text{ such that} \\ \forall t, t' \ge 0, \forall (x,y) \in \mathcal{S}(t), \forall (x',y') \in \mathcal{S}(t') \text{ it holds :} \\ 2. \quad l(x',y) \ge l(x,y) + ce^{rt}\|x'-x\|^{1+\alpha}, \\ \qquad l(x,y') \le l(x,y) - ce^{rt}\|y'-y\|^{1+\alpha}; \\ 3. \quad |l(x,y) - l(x',y)| \le k\|x-x'\|^{\alpha}, \\ \qquad |l(x,y) - l(x,y')| \le k\|y-y'\|^{\alpha}. \end{cases}$$

This assumption guarantees that for any $t \in [0,\infty)$ the saddle point mapping $\mathcal{S}(t)$ is single-valued and Lipschitz continuous; i.e., for all positive t, the function $e^{-rt}l(\cdot,\cdot)$ has a saddle point $(x(t),y(t))$ on the attainable set $R(t)$ of the dynamics (4.1), which depends in a Lipschitz way on t.

Therefore, if it turns out that so-obtained single valued saddle point mapping is a trajectory $(x(\cdot),y(\cdot))$ of (4.1), then it is an optimal feedback strategy of the game.

For example, under the above assumptions in the case when $m = p = n$ and $f(x,u) = u$, $g(y,v) = v$, the Lipschitz continuity on t of the saddle point map implies that the corresponding controls u and v belong to $\mathcal{U}$ and $\mathcal{V}$, respectively, and, hence, they generate a trajectory of the differential game.

Acknowledgment. The authors express their gratitude to anonymous referees for their valuable comments and suggestions.

REFERENCES

[1] J.-P. Aubin, *Comportement lipschitzien des solutions de problémes de minimisation convexes*, Compt. Rend. Acad. Sci. Paris, Sér. I, 295 (1982), pp. 235–238.

[2] J.-P. Aubin, *Lipschitz behavior of solutions to convex minimization problems*, Math. Oper. Res., 9 (1984), pp. 87–111.

[3] J.-P. Aubin and I. Ekeland, *Applied Nonlinear Analysis*, John Wiley & Sons, New York, 1984.

[4] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser, New York, 1990.

[5] J. Bonnans and A. Ioffe, *Quadratic growth and stability in convex programming problems with multiple solutions*, J. Convex Anal., 2 (1995), pp. 41–57.

[6] J. Bonnans and A. Shapiro, *Optimization problems with perturbations: A guided tour*, SIAM Rev., 40 (1998), pp. 228–264.

[7] J. Bonnans and A. Shapiro, *Perturbation analysis of optimization problems*, Springer Series in Operations Research, Springer, New York, 2000.

[8] P. Cardaliaguet, *A differential game with two players and one target*, SIAM J. Control Optim., 34 (1996), pp. 1441–1460.

[9] P. Cardaliaguet, M. Quincampoix, and P. Saint-Pierre, *Numerical methods for differential games*, in Stochastic and Differential Games: Theory and Numerical Methods, Annals of the International Society of Dynamic Games, M. Bardi, T. E. S. Raghavan, and T. Parthasarathy, eds., Birkhäuser, 1999, pp. 177–247.

[10] P. Cardaliaguet, M. Quincampoix, and P. Saint-Pierre, *Pursuit differential games with state constraints*, SIAM J. Control and Optim., 39 (2001), pp. 1615–1632.

[11] P. Cardaliaguet, M. Quincampoix, and P. Saint-Pierre, *Differential games through viability theory: Old and recent results*, Annals of the International Society of Dynamic Games, Birkhäuser, 9 (2007), pp. 2–23.

[12] F. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley & Sons, New York, 1983.
[13] J. Diestel and J. J. Uhl, *Vector measures*, Mathematical Surveys, No 15, AMS, Providence, RI, 1977.
[14] J. Diestel and J. J. Uhl, *The Radon-Nikodym theorem for Banach space valued measures*, Rocky Mt. J. Math., 6 (1976), pp. 1–46.
[15] A. L. Dontchev and R. T. Rockafellar, *Ample parameterization of variational inclusions*, SIAM J. Optim., 12 (2001), pp. 170–187.
[16] P. G. Georgiev and N. P. Zlateva, *Second-order subdifferentials of $C^{1,1}$ functions and optimality conditions*, Set-Valued Anal., 4 (1996), pp. 101–117.
[17] J. Hartung, *An extension of Sion's minimax theorem with an application to a method for constrained games*, Pac. J. Math., 103 (1982), pp. 401–408.
[18] J.-B. Hiriart-Urruty, J.-J. Strodiot, and V. H. Nguyen, *Generalized Hessian matrix and second-order optimality conditions for problems with $C^{1,1}$ data*, Appl. Math. Optim., 11 (1984), pp. 43–56.
[19] D. Klatte and R. Henrion, *Regularity and stability in nonlinear semi-infinite optimization*, in Semi-infinite Programming, Reemtsen et al., eds., Workshop, Cottbus, Germany, September 1996, Kluwer Academic Publishers, Boston; Nonconvex Optim. Appl., 25 (1998), pp. 69–102.
[20] N. N. Krasovskii and A. I. Subbotin, *Game-Theorical Control Problems*, Springer-Verlag, New York, 1988.
[21] R. R. Phelps, *Convex functions, monotone operators and differentiability*, 2nd ed., Lecture Notes in Mathematics 1364, Springer-Verlag, Berlin, 1993.
[22] S. Plaskacz and M. Quincampoix, *Value-functions for differential games and control systems with discontinuous terminal cost*, SIAM J. Control and Optim., 39 (2001), pp. 1485–1498.
[23] R. T. Rockafellar and R. Wets, *Variational Analysis*, Springer, Berlin, 1998.
[24] A. Shapiro, *On Lipschitzian stability of optimal solutions of parametrized semi-infinite programs*, Math. Oper. Res., 19 (1994), pp. 743–752.
[25] A. Shapiro, *Sensitivity analysis of parameterized variational inequalities*, Math. Oper. Res., 30 (2005), pp. 109–126.
[26] C. Stegall, *Optimization of functions on certain subsets of Banach spaces*, Math. Ann., 236 (1978), pp. 171–176.

SIAM J. OPTIM.
Vol. 19, No. 3, pp. 1270–1298

ON BIN PACKING WITH CONFLICTS*

LEAH EPSTEIN† AND ASAF LEVIN‡

Abstract. We consider the offline and online versions of a bin packing problem called BIN PACKING WITH CONFLICTS. Given a set of items $V = \{1, 2, \ldots, n\}$ with sizes $s_1, s_2, \ldots, s_n \in [0, 1]$ and a conflict graph $G = (V, E)$, the goal is to find a partition of the items into independent sets of G, where the total size of items in each independent set is at most 1 so that the number of independent sets in the partition is minimized. This problem is clearly a generalization of both the classical (one-dimensional) bin packing problem where $E = \emptyset$ and of the graph coloring problem where $s_i = 0$ for all $i = 1, 2, \ldots, n$. Since coloring problems on general graphs are hard to approximate, following previous work, we study the problem on specific graph classes. For the offline version, we design improved approximation algorithms for perfect graphs and other special classes of graphs: These are a $\frac{5}{2} = 2.5$-approximation algorithm for perfect graphs; a $\frac{7}{3} \approx 2.33333$-approximation algorithm for a subclass of perfect graphs, which contains interval graphs and chordal graphs; and a $\frac{7}{4} = 1.75$-approximation for algorithm bipartite graphs. For the online problem on interval graphs, we design a 4.7-competitive algorithm and show a lower bound of $\frac{155}{36} \approx 4.30556$ on the competitive ratio of any algorithm. To derive the last lower bound, we introduce the first lower bound on the asymptotic competitive ratio of any online bin packing algorithm with known optimal value, which is $\frac{47}{36} \approx 1.30556$.

Key words. bin packing, approximation algorithms, online algorithms

AMS subject classifications. 68Q25, 68W25, 68W40

DOI. 10.1137/060666329

1. Introduction.

We consider the following BIN PACKING WITH CONFLICTS problem (BPC) (see [5, 17] and also the information on the bin packing problem given in [4]). Given a set of items $V = \{1, 2, \ldots, n\}$ with sizes $s_1, s_2, \ldots, s_n \in [0, 1]$ and a conflict graph $G = (V, E)$, the goal is to find a partition of the items into independent sets of G where the total size of each independent set is at most 1 so that the number of independent sets in the partition is minimized. This problem is clearly a generalization of both the classical (one-dimensional) bin packing problem where $E = \emptyset$ and of the graph coloring problem where $s_i = 0$ for all $i = 1, 2, \ldots, n$. In an online environment, items arrive one by one to be packed immediately and irrevocably. A new item is introduced by its size, together with all its incident edges in the current conflict graph (i.e., edges which connect it to previously introduced items).

This problem arises in assigning processes or tasks to processors. In this case we are given a set of tasks, where some pairs of tasks are not allowed to execute on the same processor due to efficiency or fault tolerance reasons. The goal is to assign a minimum number of processors to this set of processes given that the makespan is bounded by some constant (see Jansen [16]). Other applications of this problem arise in the area of database replicas storage, school course time tables construction, scheduling communication systems (see de Werra [6]), and, finally, in load balancing the parallel solution of partial differential equations by two-dimensional domain decomposition (see Irani and Leung [15]). We follow earlier work and consider the

*Received by the editors July 29, 2006; accepted for publication (in revised form) June 25, 2008; published electronically November 19, 2008. An extended abstract of this paper appeared in the Proceedings of the 4th Workshop on Approximation and Online Algorithms (WAOA), 2006.
http://www.siam.org/journals/siopt/19-3/66632.html
†Department of Mathematics, University of Haifa, 31905 Haifa, Israel (lea@math.haifa.ac.il).
‡Department of Statistics, The Hebrew University, 91905 Jerusalem, Israel (levinas@mscc.huji.ac.il).

BPC on subclasses of perfect graphs. This restriction is motivated by the theoretical hardness of approximating graph coloring on general graphs.

In order to analyze our approximation and online algorithms, we use common criteria, which are the approximation ratio (also called performance guarantee) for offline algorithms, and competitive analysis for online algorithms. For an algorithm $\mathcal{A}$, we denote its cost by $\mathcal{A}$ as well. An optimal offline algorithm that knows the complete sequence of items is denoted by OPT. We consider the absolute approximation (respectively, competitive) ratio that is defined as follows. The absolute approximation (respectively, competitive) ratio of $\mathcal{A}$ is the infimum $\mathcal{R}$ such that, for any input, $\mathcal{A} \leq \mathcal{R} \cdot \mathrm{OPT}$. If the absolute approximation (respectively, competitive) ratio of an offline (respectively, online) algorithm is at most ρ, we say that it is a ρ-approximation (respectively, ρ-competitive). The asymptotic approximation (respectively, competitive) ratio of $\mathcal{A}$ is the infimum $\mathcal{R}$ such that, for any input, $\mathcal{A} \leq \mathcal{R} \cdot \mathrm{OPT} + c$ for some c that is independent of the input. For the offline problem, we restrict ourselves to algorithms that run in polynomial time. Our online algorithm is also a polynomial time algorithm (though this property is not always required in the competitive analysis literature). We focus on the absolute criteria and not on the criteria of asymptotic approximation ratio and asymptotic competitive ratio, since coloring problems are typically studied using these absolute criteria. Note that asymptotic criteria are commonly used for bin packing problems. However, the problem studied here is more similar in nature to coloring problems, even though it is an extension of bin packing. Throughout the paper we omit the adjective "absolute," since this is the only criterion studied here. We let $s(X)$ denote the sum of item sizes in a set X, i.e., $s(X) = \sum_{x \in X} s_x$.

Since the BPC problem generalizes the classical coloring problem that is known to be extremely hard to approximate, we follow earlier studies and consider the BPC problem on the class of perfect graphs for which the coloring problem is polynomially solvable (see [27]). The best previously known approximation algorithm for BPC on perfect graphs is the algorithm of Jansen and Öhring [17] with an approximation ratio of 2.7. In section 3.1, we improve this result and present our 2.5-approximation algorithm for BPC on perfect graphs.

Following Jansen and Öhring [17], we consider the class of graphs for which one can solve in polynomial time the PRECOLORING EXTENSION problem defined as follows. Given an undirected graph $G = (V, E)$ and k distinct vertices $v_1, v_2, \ldots, v_k$, the problem is to find a minimum coloring f of G such that $f(v_i) = i$ for $i = 1, 2, \ldots, k$. This problem is reviewed in [14, 23], and it is known to be polynomially solvable for the following graph classes: interval graphs, forests, split graphs, complements of bipartite graphs, cographs, partial K-trees and complements of Meyniel graphs[1] (see [14] for a review of these results), and it is also polynomially solvable for chordal graphs as shown by Marx [24]. However, it is known to be NP-complete for bipartite graphs [14]. We denote by $\mathcal{C}$ the class of graphs G for which one can solve in polynomial time the precoloring extension problem for any induced subgraph of G (including G itself). That is, $\mathcal{C}$ is closed under the operation of induced subgraph extraction. Jansen and Öhring [17] analyzed the following *algorithm with precoloring* for the case where G belongs to $\mathcal{C}$. Denote the set of large items by $L = \{j : s_j > \frac{1}{2}\}$, and denote by $\chi_I(G)$ the minimum number of colors used by an optimal solution for the precoloring extension problem defined by G. Finally, we define the set of precolored vertices to be L. Compute a feasible coloring of G using $\chi_I(G)$ colors, where,

[1] A graph is Meyniel if every cycle of odd length which is at least five, has at least two chords.

for each pair of items in L, they are assigned different colors. For each color class, apply a bin-packing heuristic such as the first-fit-decreasing (FFD) algorithm. They proved that the resulting algorithm is a $\frac{5}{2}$-approximation algorithm. In section 3.2, we improve this result by presenting a $\frac{7}{3}$-approximation algorithm.

For all $\varepsilon > 0$, Jansen and Öhring [17] also presented a $(2+\varepsilon)$-approximation algorithm for BPC on cographs and partial K-trees. Furthermore, they presented a 2-approximation algorithm for bipartite graphs. A d-inductive graph has the property that the vertices can be assigned distinct numbers $1, \ldots, n$ such that each vertex is adjacent to at most d lower-numbered vertices. Jansen [16] showed an asymptotic fully polynomial time approximation scheme for BPC on d-inductive graphs where d is a constant. This result includes the cases of trees, grid graphs, planar graphs, and graphs with constant treewidth. Oh and Son [26] and McCloskey and Shankar [25] considered BPC on graphs that are a union of cliques, but their results are inferior to the 2.7-approximation algorithm of Jansen and Öhring [17]. Even and Shahar [7] considered BPC on unit circular-arc graphs and presented a 3-approximation algorithm for this case. Since our 2.5-approximation algorithm for perfect graphs can be applied also for unit circular-arc graphs (since it only uses a coloring of an induced subgraph using a minimum number of colors, and the coloring problem is solvable in polynomial time on unit circular-arc graphs), we get an improvement of the 3-approximation algorithm of [7].

The hardness of the approximation of BPC follows from the hardness of standard offline bin packing (with respect to the absolute approximation ratio). It is not hard to see that unless $\mathbf{P} = \mathbf{NP}$, no algorithm can have an absolute approximation ratio of less than $\frac{3}{2}$ (due to a simple reduction from the PARTITION problem, see problem SP12 in [10]). Since standard bin packing is a special case of BPC, where the conflict graph is an independent set, we get that, for all graph classes studied in this paper, BPC is **APX**-hard, and unless $\mathbf{P} = \mathbf{NP}$, cannot be approximated within a factor smaller than $\frac{3}{2}$. Note that for bin packing, already the simple FFD algorithm is a $\frac{3}{2}$-approximation [29]. These hardness results hold for the graph classes we consider, since an empty graph is bipartite and perfect.

Our results. In section 2, we describe the methods applied in this paper. We use weights for our analysis. The weights used throughout the paper have the unique and novel property that weights are assigned not only as a function of the size of items, but also as a function of the location of items in an optimal solution or in an approximate solution. We think that this new technical approach can contribute to the analysis of algorithms for other problems as well.

We use these methods in section 2 to give improved and tight bounds on two algorithms designed in [17]. We show that their algorithm for perfect graphs has performance guarantee of approximately 2.691, and their algorithm with precoloring has performance guarantee of approximately 2.423. These tight results follow from our analysis together with lower bounds on the performance guarantee of these algorithms, given in [17]. Note that these bounds and their proofs resemble the analysis of the Harmonic algorithm [21] (the bounds are one unit higher than the upper bounds for Harmonic). However, neither the algorithms of [17] nor our algorithms use a partition into classes as is done in the Harmonic algorithm. Moreover, such a partition in our case would result in an arbitrarily high approximation ratios.

In section 3, we present our improved new algorithms for the offline case of BPC. In section 3.1, we design an improved algorithm for perfect graphs with performance guarantee of 2.5. Our algorithm is also a 2.5-approximation algorithm for BPC on all

graph classes where one can solve the regular coloring problem (i.e., coloring the vertex set of a graph using a minimum number of colors) in polynomial time. In section 3.2, we design an improved algorithm with precoloring with performance guarantee of $\frac{7}{3}$. In section 3.3, we design a $\frac{7}{4}$-approximation algorithm for bipartite graphs.

In section 4 we discuss *online* algorithms for BPC on interval graphs. We design a simple 4.7-competitive algorithm and show a lower bound of $\frac{155}{36} \approx 4.30556$ on the competitive ratio of any online algorithm. We derive the last lower bound by introducing the first nontrivial lower bound for online bin packing with known optimal value, which is $\frac{47}{36} \approx 1.30556$. We also show an $O(\log n)$ competitive algorithm for bipartite graphs, which is best possible. Both algorithms are adaptations of online algorithms for the standard coloring problem; see [20, 22].

2. Weighting functions and the performance of FFD-based algorithms. In this section, we define weighting functions which are a major tool in the analysis of algorithms for bin packing. The weights defined in this section are later adapted and used for the analysis of our improved algorithms.

The idea of such weights is simple. An item receives a weight according to its size and its packing in some fixed solution. The weights of items are typically not equal to their sizes, but are related to them. The weights are assigned in a way that the cost of an algorithm is close to the total sum of weights. In order to complete the analysis, it is usually necessary to consider the total weight that can be packed into a single bin of an optimal solution.

In this paper, we exploit this method in order to achieve improved algorithms for BPC. Though this method was not applied to BPC before, it was widely used for standard bin packing and many variants of bin packing. This technique was used as early as 1971 by Ullman [31] (see also [19, 21, 28]).

Specifically, we use the following theorem.

THEOREM 1. *Consider an algorithm $\mathcal{A}$ for* BPC. *Let w, w' be two weight measures defined on the input items, $w, w' : I \to \mathbb{R}$. Let $W(I)$ and $W'(I)$ denote the sum of the weights of all input items of I, according to w and w', respectively, and assume that, for every input I, $W'(I) \leq W(I)$. Assume that, for every output of the algorithm, the number of bins used by the algorithm $\mathcal{A}$ is upper bounded by $W'(I) + \tau \cdot \mathrm{OPT}(I)$, where $\mathrm{OPT}(I)$ is the cost of an optimal algorithm on the input I. Denote by W_I the supremum amount of weight that can be packed into a bin of the optimal solution, according to measure w. Then the approximation ratio of $\mathcal{A}$ is upper bounded by $W_I + \tau$.*

Note that w, w' are real functions and are not necessarily nonnegative.

Proof. Given an input I, we have $\mathcal{A} \leq W'(I) + \tau\mathrm{OPT}(I)$. Since an optimal algorithm has $\mathrm{OPT}(I)$ bins, with a weight of at most W_I in each one of them, we get the upper bound $W(I) \leq W_I \cdot \mathrm{OPT}(I)$. Using $W'(I) \leq W(I)$, we get $\mathcal{A} \leq (W_I + \tau)\mathrm{OPT}(I)$, and the theorem follows. □

Typically, this theorem is used with $w = w'$, which satisfies all conditions of the theorem. We will also use it in this paper with $w \neq w'$.

In this section, we define a set of weights which depends solely on the size of items. For an item x such that $s_x > \frac{1}{2}$, we define $weight(x) = 1$. We define the interval $\mathcal{I}_1$ by $\mathcal{I}_1 = (\frac{1}{2}, 1]$. For an item x such that $s_x \leq \frac{1}{2}$, let j be an integer such that $s_x \in \mathcal{I}_j = (\frac{1}{j+1}, \frac{1}{j}]$. We define $weight(x) = s_x + \frac{1}{j(j+1)}$. Note that even though this classification to intervals was used before, the weight function is nonstandard. Typically either all items in an interval receive the same weight or are scaled by a common multiplicative factor (see, e.g., [21, 2]). We note that the weight function

does not round up the size of an item to the next unit fraction, but the weight of an item in an interval $(\frac{1}{j+1}, \frac{1}{j}]$ is never lower than the next unit fraction (i.e., it is larger than $\frac{1}{j}$).

We need to use this special weight function in order to make sure that the amount of weight is large enough, even if the input is partitioned into several classes, each of which is packed separately. On the other hand, we must make sure that the weights are not too large so that the bound on the performance guarantee is not increased artificially. A similar (though different) weight function was used before by Galambos and Woeginger [8]. Their weight function can be used to prove Corollary 6 and Theorem 8 but not the other results of this paper. Therefore, we need to modify the weight function of [8] for our needs.

Given this set of weights, we note that, for an item x of size $s_x \in \mathcal{I}_j$ $(j \geq 2)$, the ratio between its weight and its size is bounded as follows:

$$\frac{j+2}{j+1} \leq \frac{weight(x)}{s_x} < \frac{j+1}{j}.$$

For a set of items X, we denote the sum of the weights of all items in X by $W(X)$. That is, $W(X) = \sum_{x \in X} weight(x)$. We next show that any algorithm which first partitions the input into μ color classes and then applies the FFD algorithm on each color class separately, satisfies the following condition on its cost as a function of the total weight and μ.

LEMMA 2. *Consider a subset of items J which forms an independent set and is packed using FFD. Let Y be the number of bins used for this packing. Then we have $Y \leq W(J) + 1$.*

Proof. Note that, for the above weight function, any bin which contains an item of size in $(\frac{1}{2}, 1]$ has a total weight of the items of at least 1. As stated above, the weight of an item in $\mathcal{I}_j = (\frac{1}{j+1}, \frac{1}{j}]$ is at least $\frac{1}{j+1} + \frac{1}{j(j+1)} = \frac{1}{j}$. Therefore, any bin which contains j items of size in the interval $(\frac{1}{j+1}, \frac{1}{j}]$ has a total weight of at least 1. We can remove such bins from the packing and focus on all other bins, which are called *transition bins*. If no bins are left after the removal, we are done.

A transition bin contains only items whose sizes are at most $\frac{1}{2}$. Note that the last bin ever opened may result in a transition bin, and it contains at least one item. Moreover, let a transition bin be of *type* j (for some $j \geq 2$), if the first item ever packed into it has size in $\mathcal{I}_j$. Next, we argue that there can be at most one transition bin of each possible type. Since the items are packed using FFD, transition bins are created in a sorted order, starting with the smallest type (or largest size). If there are two bins of the same type j, this means that during the time between the packing of the first items in these two bins, all packed items were also of size in the interval $\mathcal{I}_j$. Therefore, the first bin must be assigned j such items before the second transition bin of this type is opened, and thus the first bin is not a transition bin. Let k be the largest type of any transition bin ever opened (i.e., the transition bin with the smallest item). Remove from the packing all items of size at most $\frac{1}{k+1}$. This removal may only decrease the total weight. As stated above, the weight of each remaining item in the transition bins is at least a multiplicative factor of $\frac{k+2}{k+1}$ its size.

Let α be the size of the first item in the last transition bin. As the last transition bin is opened, all other bins have a total size of items which is more than $1 - \alpha$. Let $i_1 < \cdots < i_t < i_{t+1} = k$ be the sorted list of the types of transition bins.

It suffices to show that the total weight of all items in the transition bins is at least t (since there are $t + 1$ transition bins). Let $t' = \min\{\lfloor \frac{k+2}{2} \rfloor, t\}$.

CLAIM 3. *The total weight in the last $t'+1$ transition bins is at least t', and each other bin carries a weight of at least 1.*

Proof. If $t = t' = 0$, we are done. Assume therefore that $t' \geq 1$. In the last $t'+1$ transition bins, we get a total weight of at least

$$t'(1-\alpha)\frac{k+2}{k+1} + \alpha + \frac{1}{k(k+1)} = t'\frac{k+2}{k+1} + \frac{1}{k(k+1)} - \alpha\left(t'\frac{k+2}{k+1} - 1\right)$$
$$\geq t'\frac{k+2}{k+1} + \frac{1}{k(k+1)} - \frac{t'\frac{k+2}{k+1} - 1}{k} = t'\frac{k+2}{k+1}\frac{k-1}{k} + \frac{k+2}{k(k+1)}.$$

The inequality holds, since the coefficient multiplied by α is negative and $\alpha \leq \frac{1}{k}$. We need to show that the weight is at least t', i.e., that

$$t'\left(\frac{k^2+k-2}{k^2+k} - 1\right) + \frac{k+2}{k(k+1)} = \frac{-2t'}{k^2+k} + \frac{k+2}{k^2+k} \geq 0.$$

We get that this holds for $t' \leq \frac{k+2}{2}$, which holds due to the definition of t'.

Consider the second part of the claim. It is enough to consider the first $f = t - \lfloor\frac{k+2}{2}\rfloor$ transition bins, in the case where f is positive. These bins are transition bins of types $i_1, \ldots, i_f$. Consider the bin of type i_{f+1}. Note that $i_{f+1} \leq k - \lfloor\frac{k+2}{2}\rfloor$ since no two transition bins are of the same type, and $i_{f+1} \geq 3$ since $i_f \geq 2$. Let β be the size of the first item in the bin of type i_{f+1}. Let $m = i_{f+1}$. Considering only the items of sizes in $(\frac{1}{m}, \frac{1}{2}]$, we have that each bin out of the first f transition bins has a total size of such items of at least $1-\beta$. However, they also have a total size of items in $(\frac{1}{k+1}, \frac{1}{2}]$ of at least $1-\alpha$. Therefore, the weight of items in each such bin is at least

$$(1-\beta)\frac{m+2}{m+1} + (\beta-\alpha)\frac{k+2}{k+1} = \frac{m+2}{m+1} + \beta\left(\frac{1}{k+1} - \frac{1}{m+1}\right) - \alpha\frac{k+2}{k+1}.$$

We will show that this amount is never smaller than 1. This expression is minimized for maximum values of α, β, and thus we need to show that

$$\left(1 - \frac{1}{m}\right)\cdot\frac{m+2}{m+1} + \left(\frac{1}{m} - \frac{1}{k}\right)\cdot\frac{k+2}{k+1} - 1 \geq 0,$$

which is equivalent to $\frac{k-m}{km}\cdot\frac{k+2}{k+1} \geq \frac{2}{m(m+1)}$. Note that $k - m \geq \frac{k+1}{2}$, $m+1 \geq 4$, and thus $\frac{k-m}{km}\frac{k+2}{k+1}\cdot\frac{m(m+1)}{2} \geq \frac{k+2}{k} > 1$. ∎

This completes the proof of Lemma 2. ∎

In what follows, we consider algorithms for an input I of the following structure, called COLORFFD. The set I is partitioned into ν independent sets. Out of these sets, μ sets (where $\mu \leq \nu$) are packed using FFD. Each other independent set J is packed into a single bin and is assigned a total weight of at least 1.

COROLLARY 4. *Let $\mathcal{B}$ be a* COLORFFD *algorithm. Then $\mathcal{B}$ satisfies $\mathcal{B} \leq W(I) + \mu$.*

We now give a tight analysis of the FFD-based algorithm given in [17] for perfect graphs.

The FFD-based algorithm of [17].
Find a coloring of all items with a minimum number of colors. Use FFD to pack each color class.

It was shown in [17] that the performance guarantee of this algorithm is at most 2.7 and at least $1+\Pi_\infty \approx 2.69103$. The value $\Pi_\infty = \Pi_\infty(1)$ is computed using the well-known sequence $\pi_i = \pi_i(1)$, $i \geq 1$, which often occurs in bin packing. Given a positive integer z, let $\pi_0(z) = z$, $\pi_1(z) = z+1$, and, for $i \geq 1$, $\pi_{i+1}(z) = \pi_i(z)(\pi_i(z)-1)+1$. Then $\Pi_\infty(z) = \sum_{i=1}^{\infty} \frac{1}{\pi_i(z)-1}$. An alternative definition is $\pi_0(z) = z, \pi_{i+1}(z) = (\prod_{j=0}^{i} \pi_j(z))+1$ for $i \geq 1$. The equivalence can be shown by induction. For $i = 1$, by this definition, $\pi_1(z) = z+1$ as required. Assume this is true for a given $i \geq 1$. For $i+1$, we have $\pi_{i+1}(z) = \pi_i(z)(\pi_i(z)-1)+1 = (\pi_i(z))^2 - \pi_i(z)+1 = \pi_i(z)(\prod_{j=0}^{i-1} \pi_j(z)+1) - \pi_i(z) + 1 = (\prod_{j=0}^{i} \pi_j(z)) + 1$.

This sequence with $z = 1$ is known as "Sylvester's sequence" and as "Euclid numbers" (see sequence A000058 in [30]). The sequence is often used for the analysis of bin packing algorithms and the design of lower bound for online variants of the problem; see [2, 21, 32].

The relation of this sequence to bin packing is as follows. Each number $\pi_i(z)$ is the smallest integer such that a bin of size $\frac{1}{z}$ that already contains the items of sizes slightly larger than $\frac{1}{\pi_j(z)}$, for $j = 1, \ldots, i-1$, can accommodate an item that is slightly larger than $\frac{1}{\pi_i(z)}$. Indeed, we can show by induction that $\sum_{i=1}^{k} \frac{1}{\pi_i(z)} = \frac{1}{z} - \frac{1}{\pi_{k+1}(z)-1}$ for all $k \geq 0$. This clearly holds for $k = 0$. Assume that this holds for some value of $k \geq 0$, then $\sum_{i=1}^{k+1} \frac{1}{\pi_i(z)} = \frac{1}{z} - \frac{1}{\pi_{k+1}(z)-1} + \frac{1}{\pi_{k+1}(z)} = \frac{1}{z} - \frac{1}{(\pi_{k+1}(z)-1)(\pi_{k+1}(z))} = \frac{1}{z} - \frac{1}{\pi_{k+2}(z)-1}$.

We are now ready to prove a matching upper bound of $1+\Pi_\infty = 2.691$ for this algorithm. In order to do so, we need to find an upper bound on the total weight which can reside in one bin. The proof is similar to those of [2, 21], however, our weights are defined differently, since these proofs do not hold in our case. We assume that the weight of an item x of size in $\mathcal{I}_1$ is $s_x + \frac{1}{2}$, which may only increase the total weight, since we previously assigned weight 1 to these items.

LEMMA 5. *Consider a set of items J, whose total size is $s(J) \leq \frac{1}{z}$. Then $W(J) - s(J) \leq \Pi_\infty(z) - \frac{1}{z}$, and thus $W(J) \leq \Pi_\infty(z)$.*

Proof. We define the "increase," or modified weight, of an item by its weight minus its size, i.e., $weight'(x) = weight(x) - s_x$. Let $W'(X)$, for a set X, be the sum of the increases of items in X. We need to show that $W'(J) \leq \Pi_\infty(z) - \frac{1}{z}$. Specifically, we show that $\Pi_\infty(z) - \frac{1}{z}$ is the supremum total increase of the items of J.

First, consider the sequence $\pi_i(z)$, using its definition, we get that, for any value of k and small enough $\delta > 0$, the following holds: $\sum_{i=1}^{k} (\frac{1}{\pi_i(z)} + \delta) < \frac{1}{z}$. This means that a set of items of sizes $\frac{1}{\pi_i(z)} + \delta$ for $1 \leq i \leq k$ has a total size of at most $\frac{1}{z}$. Taking $k \to \infty$, we get that the total increase in a bin which contains such items tends to $\sum_{i=1}^{\infty} \frac{1}{\pi_i(z)(\pi_i(z)-1)} = \sum_{i=1}^{\infty} \frac{1}{\pi_{i+1}(z)-1} = \Pi_\infty(z) - \frac{1}{z}$. We call a part of a bin of size $\frac{1}{z}$ a *partial bin*. We will next show that a set of the items of these sizes is the worst case contents of a partial bin, which would prove the claim.

Assume by contradiction that there exists a value $\varepsilon > 0$ and a set of items for which the sum of increases is at least $\Pi_\infty(z) - \frac{1}{z} + \varepsilon$. We prove that if the partial bin contains exactly one item of each interval in the set of intervals $\{\mathcal{I}_{\pi_1(z)-1}, \ldots, \mathcal{I}_{\pi_{i-1}(z)-1}\}$, then it also contains exactly one item of the interval $\mathcal{I}_{\pi_i(z)-1}$.

We have $\frac{1}{z} - \sum_{j=1}^{i-1} \frac{1}{\pi_j(z)} = \frac{1}{\pi_i(z)-1}$ for $i \geq 1$. Therefore, the largest interval from which the next item can come from is $\mathcal{I}_{\pi_i(z)-1}$. Assume that there is no such item. Thus, all other item sizes are from the interval $(0, \frac{1}{\pi_i(z)}]$, and therefore the increase of an item is smaller than a multiplicative factor of $\frac{1}{\pi_i(z)}$ of its size. The total increase is therefore smaller than $\sum_{j=1}^{i-1} \frac{1}{\pi_j(z)(\pi_j(z)-1)} + \frac{1}{\pi_i(z)-1} \cdot \frac{1}{\pi_i(z)} = \sum_{j=1}^{i} \frac{1}{\pi_{i+1}(z)-1} < \Pi_\infty(z) - \frac{1}{z}$. This leads to a contradiction, and therefore, in order to get above $\Pi_\infty(z) - \frac{1}{z}$, there must be an item of size in $(\frac{1}{\pi_i(z)}, \frac{1}{\pi_i(z)-1}]$. We have proved that the partial

bin must contain at least one item of $\mathcal{I}_{\pi_i(z)}$. Note that, by the above argument, the largest item, excluding one item of each interval in $\{\mathcal{I}_{\pi_1(z)-1}, \ldots, \mathcal{I}_{\pi_i(z)-1}\}$, is of size at most $\frac{1}{\pi_{i+1}(z)-1}$, and therefore the partial bin must contain exactly one item from $\mathcal{I}_{\pi_i(z)-1}$. By this claim (applied to $i = 1$), there is exactly one item from $\mathcal{I}_{\pi_1(z)-1}$, and, by induction on i, there is exactly one item from each interval $\mathcal{I}_{\pi_i(z)-1}$ for all $i \geq 1$.

Consider the smallest value of i such that $\frac{1}{\pi_{i+1}(z)-1} < \varepsilon$. The total size of all the items excluding the largest i items, is at most $\frac{1}{\pi_{i+1}(z)-1}$. No matter what these items are, they have an increase of at most their size, that is, of at most $\frac{1}{\pi_{i+1}(z)-1}$. This means that the total increase is at most $\Pi_\infty(z) - \frac{1}{z} + \frac{1}{\pi_{i+1}(z)-1} < \Pi_\infty(z) - \frac{1}{z} + \varepsilon$, which leads to a contradiction. □

COROLLARY 6. *The performance guarantee of the* FFD*-based algorithm* $\mathcal{A}$ *of* [17] *for perfect graphs is* $\Pi_\infty + 1 \approx 2.69103$.

Proof. As [17] supplies an example which achieves this bound (asymptotically), we should prove the upper bound. Since the input is colored optimally, the number of independent sets is exactly $\mu = \chi(G) \leq \mathrm{OPT}$. We have $\mathcal{A} \leq W(I) + \mathrm{OPT}$. However, $W(I) \leq \Pi_\infty \cdot \mathrm{OPT}$, since by Lemma 5, any bin (of size 1) may contain a total weight of at most $\Pi_\infty \approx 1.69103$, and thus $\mathcal{A} \leq (\Pi_\infty + 1) \cdot \mathrm{OPT}$. □

We give a formal definition of *the algorithm with precoloring* of [17], which is used for graphs in class $\mathcal{C}$.

The algorithm with precoloring of [17].
Let $L = \{j : s_j > \frac{1}{2}\}$.
Compute a feasible coloring of the conflict graph using the minimum number of colors needed to color it, such that each pair of items in L is assigned two distinct colors.
Use FFD to pack each color class.

In order to analyze the *algorithm with precoloring*, we need to define a set of weights which does not give very high weights to items in $\mathcal{I}_1 = (\frac{1}{2}, 1]$. We define the weight for $s_x \in \mathcal{I}_1$ to be $weight(x) = s_x + \frac{1}{6}$. This unique definition is possible due to the special treatment of items in $\mathcal{I}_1$.

In order to establish a lemma regarding the sum of weights in an independent set, we modify the type of algorithms we allow to use. Once again, the set I is partitioned into ν independent sets. Each independent set has at most one item of size in $\mathcal{I}_1$. Out of these sets $\mu \leq \nu$ are packed using FFD. Assume that each other independent set J is packed into a single bin and is assigned a total weight of at least 1.

LEMMA 7. *An algorithm* $\mathcal{B}$ *as above satisfies* $\mathcal{B} \leq W(I) + \mu$.

Proof. We consider two types of independent sets. The first type of independent sets are sets which are packed into a single bin. For each such set, we note that the total weight of items in it is at least 1, by definition.

The second type of independent sets are all other such sets, i.e., independent sets which require at least two bins. We need to show that the sum of the weights of items in each such independent set J is at least the number of bins in this set, minus 1. That is, if J is packed into λ bins, then the total weight of the items of J is at least $\lambda - 1$.

Therefore, in order to complete the proof, we note that independent sets where no item of size in $\mathcal{I}_1$ exists do not have a change in weights (compared to Lemma 2), and thus the previous proof holds.

Consider therefore an independent set packed using FFD, which has a single item in $\mathcal{I}_1$. We may assume that this set consists of at least two bins, otherwise the total weight clearly satisfies the property. The item of size in $\mathcal{I}_1$ is the first one to be packed

by FFD. In this proof we consider this single bin to be a transition bin as well. Let k be the type of the last transition bin as in the proof of Lemma 2. We again calculate only the weight of these bins and show that their number exceeds their total weight by at most 1. Note that if we show that the number of transition bins is ℓ and the total weight in these bins is at least $\ell - 1$, then we are done. We denote the size of the first item in the last transition bin by α.

We have six cases, which are $k \geq 6$ and $k = 1, 2, 3, 4$, or 5.

- If $k \geq 6$, we have that the total size y of items in the first transition bin is at least $\frac{5}{6}$, and thus the weight is at least $y + \frac{1}{6} \geq 1$ (since the weight of the large item is its size plus $\frac{1}{6}$, and no weight of an item is smaller than its size). The rest of the transition bins can be considered as in the proof of Lemma 2.
- Otherwise, if there is a single transition bin, we are done.
- Otherwise, if $k = 2$ (and thus there are exactly two transition bins), or if $3 \leq k \leq 5$, and there are two transition bins. This means that the first item assigned to the second transition bin could not be assigned to the first one. Thus, the total sum of item sizes in the two bins is strictly larger than 1. The weight of an item is no smaller than its size, so the total weight in the transition bins is at least 1, which is what needs to be proved for two such bins.
- Otherwise, if $k = 3$, there are at most three transition bins. The case of two transition bins was considered, so we may assume that three transition bins exist. As in the proof of Lemma 2, we lower bound the sum of weights in bins that are not the first or last transition bin using $\frac{k+2}{k+1}(1-\alpha)$. Since $\frac{1}{k+1} \leq \alpha \leq \frac{1}{k}$, this value is strictly smaller than 1. The weight of items in the first transition bin is at least $1 - \alpha + \frac{1}{6}$, and in the last one it is at least $\alpha + \frac{1}{k(k+1)}$. For $k = 3$, this gives a total weight of at least $1 - \alpha + \frac{1}{6} + \frac{5}{4}(1 - \alpha) + \alpha + \frac{1}{12} = \frac{5}{2} - \frac{5\alpha}{4} \geq \frac{25}{12} > 2$.
- Otherwise, if $k = 4$, then there are three or four transition bins. Let t denote the number of transition bins. We get a total weight of at least

$$1 - \alpha + \frac{1}{6} + (t-2)\frac{6}{5}(1-\alpha) + \alpha + \frac{1}{20} \geq \frac{6}{5}(t-2)\left(1 - \frac{1}{4}\right) + \frac{73}{60} = \frac{9}{10}(t-2) + \frac{73}{60} > t - 1,$$

  which holds for $t = 3, 4$.
- Finally, we are left with the case $k = 5$. We denote again by t the number of transition bins (which is at least three and at most five) and get a total weight of at least

$$1-\alpha+\frac{1}{6}+(t-2)\frac{7}{6}(1-\alpha)+\alpha+\frac{1}{30} = \frac{7}{6}(t-2)\left(1 - \frac{1}{5}\right)+\frac{6}{5} = \frac{14}{15}(t-2)+\frac{6}{5} \geq t-1,$$

  which holds for $t = 3, 4, 5$. □

We can now show a tight analysis of the FFD-based *algorithm with precoloring* given in [17]. It was shown in [17] that the performance guarantee of this algorithm is at most 2.5 and at least $\Pi_\infty(3) + 2 \approx 2.4231$.

THEOREM 8. *The performance guarantee of the* FFD*-based algorithm with precoloring* $\mathcal{B}$ *of* [17] *is* $\Pi_\infty(3) + 2 \approx 2.4231$.

Proof. Since in this case $\mu = \chi_I(G) \leq \mathrm{OPT}$, it is left to upper bound the amount of weight that can fit into a single bin and show that it is at most $\Pi_\infty(3)+1 \approx 1.4231$.

Given a packed bin in OPT, we may assume that all items have size at most $\frac{1}{2}$. Otherwise, there is a single item of size $y > \frac{1}{2}$, replace this item with two items of size $\frac{y}{2}$. If $\frac{1}{4} < \frac{y}{2} \leq \frac{1}{3}$, then the weight of each of these two items is $\frac{y}{2} + \frac{1}{12}$, and their total weight equals the weight of the original item. If $\frac{1}{3} < \frac{y}{2} \leq \frac{1}{2}$, then the weight of each of these two items is $\frac{y}{2} + \frac{1}{6}$, and their total weight is even larger than the weight of the original item. Hence, we can assume that all items have size at most $\frac{1}{2}$.

Consider the case where the bin contains no items of size in $\mathcal{I}_2$ or just one such item. Since, for smaller items, the ratio between weight and size is at most $\frac{4}{3}$, we conclude that the total weight is at most $\frac{4}{3}$ in the case of zero items and at most $\frac{25}{18} \approx 1.38889$ in the case of one item. If it contains two such items, then since the remainder of the bin is of size smaller than $\frac{1}{3}$, the difference between the total weight of the additional items in the bin and their total size is at most $\Pi_\infty(3) - \frac{1}{3}$, by Lemma 5. The difference between the total weight of the items in $\mathcal{I}_2$ and their total size is $2 \cdot (\frac{1}{6}) = \frac{1}{3}$. Thus, the total weight of all items is at most the total difference, which is at most $\Pi_\infty(3)$, plus the total size, which gives $\Pi_\infty(3) + 1 \approx 1.4231$. Note that the weights of items of size in $(0, \frac{1}{2}]$ are the same as is used for the proof of Lemma 5, and therefore the usage of that lemma here is correct. □

3. Improved algorithms. In the previous section, we showed better bounds for two variants of the problem, based on previously known algorithms from [17]. Though this already gives an improvement over the previously known bounds, the bounds we have shown are tight bounds, and thus further improvement is possible only using new algorithms, which we now design. To analyze these algorithms, we use weighting in a more complex way.

3.1. Perfect conflict graphs. We design an algorithm which uses a preprocessing phase.

Algorithm. MATCHING PREPROCESSING

1. Define the following bipartite graph. One set of vertices consists of all items of size in $\mathcal{I}_1$. The other set of vertices consists of all other items. An edge (a, b) between the vertices of the items of sizes $s_a > \frac{1}{2}$ and $s_b \leq \frac{1}{2}$ occurs if the two following conditions hold:
   (a) $s_a + s_b \leq 1$.
   (b) $(a, b) \notin E(G)$.
   That is, if these two items can be placed in a bin together. If this edge occurs, we give it the cost $c(a, b) = weight(b)$, where $weight(b)$ is defined as above to be $s_b + \frac{1}{j(j+1)}$, for the integer j such that $s_b \in (\frac{1}{j+1}, \frac{1}{j}]$.
2. Find a maximum cost matching in the bipartite graph.
3. Each pair of matched vertices are removed from G and packed into a bin together.
4. Let G' denote the induced subgraph over the items that were not packed in the preprocessing.
5. Compute a feasible coloring of G' using $\chi(G')$ colors.
6. For each color class, apply the FFD algorithm.

We next analyze this algorithm.

THEOREM 9. *The above algorithm is a* $\frac{5}{2} = 2.5$*-approximation algorithm.*

Proof. The outline of the proof is as follows. We assign weights to items according to an optimal packing and call this weight function $weight_1$, or the reduced weights. Afterwards, we take the total weight of all items, according to $weight_1$, and reassign

it to items, to get the weight function $weight_2$, so that the total weight of all items does not grow, and we will be able to apply Corollary 4 to bound the number of bins packed by the FFD algorithm, as the algorithm is a COLORFFD algorithm for these items. We will analyze the algorithm using $weight_2$ and the optimal packing using $weight_1$. Specifically, we will use Theorem 1 with these two weights functions. In what follows, we use the notation *weight* for the regular weight function (as used in the proofs for perfect graphs in section 2).

Fix an optimal packing, OPT. We next define a reduced weight function that is based on the weight function *weight*, which is defined for perfect graphs in section 2, with some modifications. For a bin in OPT with no items of size in $\mathcal{I}_1$, reduced weights are defined to be equal to weights according to *weight*. For an item of size in $\mathcal{I}_1$, the reduced weight is 1, which is equal to its weight according to *weight*. Given a bin in OPT with an item of size in $\mathcal{I}_1$ which contains additional items, pick an item of largest size in the bin among the items in the bin with size at most $\frac{1}{2}$, and give it a reduced weight zero. All other items in the bin receive reduced weights that are equal to their weights.

We next define a valid matching in the bipartite graph, defined by the algorithm MATCHING PREPROCESSING. Consider the set of items whose reduced weight is zero. Match each such item to the item of size in $\mathcal{I}_1$ that is placed together with it in a bin of OPT. Consider the cost of this matching. This cost is the sum of the weights of the items (according to *weight*) whose reduced weight is zero. Thus the cost of this matching is exactly the total reduction in the weights of items, i.e., the difference between the total weight of all items according to *weight* and the total weight according to $weight_1$.

Let ω be the cost of the matching removed by the algorithm. Then, by the optimality of the removed matching, we conclude that $\sum_{x\in I}(weight(x)-weight_1(x)) \leq \omega$. We reassign weights to items so that an item of size in $(0, \frac{1}{2}]$ that was removed in the matching (by the algorithm) receives weight zero, and any other item receives a weight as defined by the function *weight*. This weight function (after the reassignment) is called $weight_2$. We have

$$\sum_{x\in I} weight_2(x) + \omega = \sum_{x\in I} weight(x) \leq \omega + \sum_{x\in I} weight_1(x).$$

Therefore, the total weight according to $weight_2$ is no larger than the total weight according to $weight_1$. By Theorem 1, we may analyze the algorithm using the weights $weight_2$ and analyze OPT using the weights $weight_1$. Clearly, each of the bins removed by the algorithm in the matching has weight of at least 1, since each of these contains an item of unit weight. Therefore, we can use Corollary 4, since the weights of items that are packed using FFD are the same as before.

Finally, we need to find an upper bound on the largest amount of the weight of items that can be packed into a single bin of OPT, according to $weight_1$. Note that, for every item, its weight according to $weight_1$ is no larger than its weight according to *weight*. Using Theorem 8, we can see that if all item sizes are no larger than $\frac{1}{2}$, the total weight of items that can be packed into one bin has a total weight (according to *weight* and thus also according to $weight_1$) smaller than $\frac{3}{2}$. Consider now a bin with an item of size in $\mathcal{I}_1$. If this is the only item in the bin, then the total weight of all items in this bin is simply 1. Otherwise, let $x_1 \geq \cdots \geq x_t$ be the sorted list of other items in the bin, where x_1 is the item which was assigned a zero weight in $weight_1$.

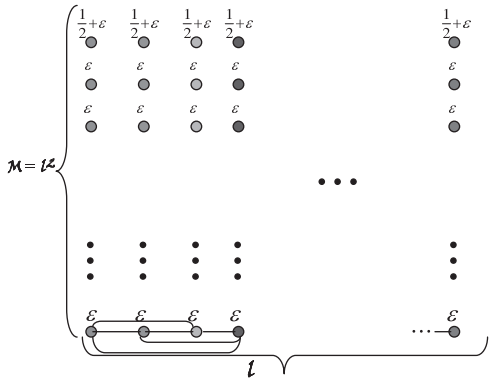


FIG. 1. *The tight example for perfect graphs. The items of each column are packed in one bin by the optimal solution. In the algorithm, the preprocessing step removes the two top items from each column, the remaining items all receive the same color, except for $\ell-1$ items of the last row.*

Let $j \geq 2$ be an integer such that $x_1 \in \mathcal{I}_j$. The total weight, according to $weight_1$, of the large item and items $x_1, \ldots, x_t$ is therefore at most

$$\frac{j+1}{j}\left(\sum_{i=2}^{t} s_{x_i}\right) + 1 \leq 1 + \frac{j+1}{j}\left(1 - \frac{1}{2} - \frac{1}{j+1}\right) = 1 + \frac{j+1}{2j} - \frac{1}{j} = 1 + \frac{j-1}{2j} < \frac{3}{2}.$$

By Corollary 4, the algorithm has an approximation ratio of at most 2.5. □

We next show that our analysis of algorithm MATCHING PREPROCESSING is tight.

PROPOSITION 10. *The approximation ratio of algorithm* MATCHING PREPROCESSING *is at least* 2.5.

Proof. Let M and ℓ be large constants and $\varepsilon = \frac{1}{2M}$. To construct the set of items we do as follows. We use one sequence of ℓ items $a_1, \ldots, a_\ell$ each with size $\frac{1}{2} + \varepsilon$. Furthermore, we have $(M-1)\ell$ additional items $b_{i,j}$, $1 \leq i \leq \ell$, $1 \leq j \leq M-1$, each of size ε. The conflict graph induces a clique with the ℓ items $b_{i,M-1}$ and contains no further edges (see Figure 1).

An optimal solution is given by ℓ independent sets $U_i = \{a_i, b_{i,1}, \ldots, b_{i,M-1}\}$ with total size of exactly one for each set. Thus $\mathrm{OPT} = \ell$.

The preprocessing step finds ℓ pairs, which are $\{a_i, b_{i,1}\}$. Next, a coloring with ℓ colors which is found for the remaining items consists of one independent set, which contains all items $b_{i,j}$ for $1 \leq i \leq \ell$, $2 \leq j \leq M-2$ and additionally contains $b_{1,M-1}$. Each other independent set contains a single item $b_{i,M-1}$ for $2 \leq i \leq \ell$. The number of bins used to color the first independent set is $\lceil \frac{\ell(M-3)+1}{2M} \rceil$, since this is the total size of items. Each other independent set consumes one additional bin, thus in total we get at least $2\ell - 1 + \frac{\ell}{2} - \frac{3\ell}{2M}$ bins. It can be seen that, for $M = \ell^2$ and $\ell >> 1$, the ratio becomes arbitrarily close to $\frac{5}{2}$. □

Remark 11. Algorithm MATCHING PREPROCESSING is a 2.5-approximation algorithm for BPC on any hereditary class of graphs for which one can find in polynomial time a coloring that uses a minimum number of colors.

Remark 11 shows that algorithm MATCHING PREPROCESSING is also a 2.5-approximation algorithm for BPC on unit circular-arc graphs, hence improving the result of [7].

3.2. Conflict graphs that belong to $\mathcal{C}$. In this section, we study an approximation algorithm for the case where the conflict graph G belongs to $\mathcal{C}$. That is, given

an induced subgraph of G, $G' = (V', E')$, and a set of vertices $L' \subseteq V'$, we can find a coloring of G using a minimum number of colors such that each pair of vertices from L' are assigned distinct colors.

We analyze the following algorithm. The weight function $weight$ is defined as in section 2 for items with size at most $\frac{1}{2}$ and for an item x such that $s_x \in \mathcal{I}_1$, $weight(x) = s_x + \frac{1}{6}$. We can use Lemma 7, since our algorithm will follow its conditions.

Algorithm. GREEDY PREPROCESSING

1. **While** there is a set of three items $\{a, b, c\}$ that can fit into one bin (i.e., $s_a + s_b + s_c \leq 1$ and $\{a, b, c\}$ is an independent set of G) such that $weight(a) + weight(b) + weight(c) > 1$ and $s_c \leq s_b \leq s_a \leq \frac{1}{2}$, or two items $\{a, b\}$ that can fit into one bin (i.e., $s_a + s_b \leq 1$ and $\{a, b\}$ is an independent set of G) such that $weight(a) + weight(b) > 1$, **do**, as follows.
   Choose such a set A of maximum total weight. Delete A from G and assign a new bin for the items of A that is dedicated to this set of items.
   Denote by $G' = (V', E')$ the resulting conflict graph induced by the remaining items.
2. Denote the set of large items by $L = \{j \in V' : s_j > \frac{1}{2}\}$, and denote by $\chi_I(G')$ the minimum number of colors used by the optimal solution for the precoloring extension problem defined by G' and the set of precolored vertices L. Compute a feasible coloring of G' using $\chi_I(G')$ colors, where any two items in L are assigned different colors.
3. For each color class, apply the FFD algorithm.

THEOREM 12. *The approximation ratio of the above algorithm is exactly* $\frac{7}{3} \approx 2.33333$.

Proof. Fix an optimal solution OPT. Let $weight$ be the weight function as used in the algorithm. We use Theorem 1 again and apply a method similar to the proof of Theorem 9; however, in the proof here we need only one auxiliary weight function $weight_1$, so the two functions w, w' that are required for the usage of Theorem 1 are equal. We assign weights that are based on the packing of OPT and denote this weight function by $weight_1$ (also called the reduced weights). For an optimal bin which contains no items of size in $\mathcal{I}_1$ and contains no triple of items of total weight strictly larger than 1 with respect to $weight$, we use $weight_1 = weight$ to define the weights of items for all items in the bin. For a bin which contains an item x of size in $\mathcal{I}_1$ but contains no other item y such that $weight(x) + weight(y) > 1$, we again use $weight_1 = weight$ for every item in the bin.

For a bin which contains no items of size in $\mathcal{I}_1$ but contains a triple of items of total weight strictly larger than 1 with respect to $weight$, let a_1, a_2, a_3 be three items with largest weights in the bin ordered according to their weight. Note that $s_{a_1} \in \mathcal{I}_2 \cup \mathcal{I}_3$, since otherwise the sum of weights of the three items cannot exceed 1. We define the reduction value for this bin to be $\Delta = \frac{weight(a_1)+weight(a_2)+weight(a_3)-1}{3}$. The reduction value is used for the definition of weights, according to $weight_1$. For any item b in this bin such that $b \neq a_i$ for $i = 1, 2, 3$, we define $weight_1(b) = weight(b)$. Note that in the preprocessing step, the algorithm removes at least one of a_1, a_2, and a_3, since otherwise if all three items are not removed, then the preprocessing step cannot terminate. Let i' be an index of the item of an item a_i such that $1 \leq i' \leq 3$, and $a_{i'}$ is removed no later than a_j for all $1 \leq j \leq 3$. We define $weight_1(a_{i'}) = weight(a_{i'}) - \Delta$, and, for $i \neq i'$, $weight_1(a_i) = weight(a_i)$.

For a bin which contains an item x of size in $\mathcal{I}_1$ and contains another item y such that $weight(x) + weight(y) > 1$, let y be such an item with maximum weight according

to $weight$. We define the reduction value for this bin to be $\Delta = \frac{weight(x)+weight(y)-1}{3}$. For any item b in the bin for $b \neq x, y$, we define $weight_1(b) = weight(b)$. Note that at least one of x and y is removed in the preprocessing step. If y is removed no later than x, we define $weight_1(y) = weight(y) - \Delta$ and $weight_1(x) = weight(x)$, and otherwise $weight_1(y) = weight(y)$ and $weight_1(x) = weight(x) - \Delta$.

Consider a bin which is removed in the greedy preprocessing step. In order to be able to use Corollary 4, we argue that the total weight of the items in this bin according to $weight_1$ is greater than 1. First note that the total weight of the items according to $weight$ is at least 1. Therefore, if for every item a in this bin $weight_1(a) = weight(a)$, we get that the sum of weights in this bin is strictly larger than 1. We will show that a possible reduction in the weights (i.e., a possible difference between $weight_1$ and $weight$) does not decrease the sum of weights below 1. Let A be the set of (two or three items) in this bin and $\Gamma = \frac{\left(\sum_{a\in A} weight(a)\right)-1}{3}$. For an item $a \in A$, we have $weight_1(a) < weight(a)$ if the following conditions hold. Consider the bin to which a belongs in OPT. Then a value $\Delta(a)$ was computed for this bin such that $weight_1(a) = weight(a) - \Delta(a)$. We get that a is removed in the preprocessing, and at the time of the removal of a, it belongs to a set of items $\mathbf{A}$ of largest weight that is valid for removal in the preprocessing step. Moreover, no item of $\mathbf{A}$ has been already removed at the time that a is being removed. This means that, in the greedy process, we have $\Gamma \geq \Delta(a)$. Thus we have $\sum_{a\in A} weight_1(a) = \sum_{a\in A}[weight(a) - \Delta(a)] \geq \sum_{a\in A} weight(a) - 3\Gamma = 1$. Therefore, each of the bins that were removed by the algorithm in the greedy preprocessing step has weight of at least 1. Therefore, we can use Lemma 7, since the weights of items that are packed using FFD are the same as in section 2.

Finally, we need to analyze the largest amount of weight that can be packed into a single bin of OPT. This analysis is done with respect to $weight_1$. Consider the set of items A in a given bin of OPT.

If all items in A have size at most $\frac{1}{3}$, then for all $a \in A$, $weight_1(a) \leq \frac{4}{3} \cdot s_a$, and thus the total weight of the items in A is at most $\frac{4}{3}$. This covers both the case where there is no reduction in the weight of items in $weight_1$ compared to $weight$ and the case where there is such a reduction for some items.

Next, assume that A has an item x of size in $\mathcal{I}_2$, but all weights in this bin were assigned according to $weight$ (i.e., for all $a \in A$, $weight_1(a) = weight(a)$). This can happen in two cases.

- If A contains an additional item y of size in $\mathcal{I}_2$, then $A = \{x, y\}$. This is so as a third item in the bin would imply a triple whose total weight is strictly more than 1, and hence we will have $weight_1(x) \neq weight(x)$. Therefore, in this case where $A = \{x, y\}$, we get a total weight of
$$s_x + \frac{1}{6} + s_y + \frac{1}{6} \leq 1 + \frac{1}{3} \leq \frac{4}{3}.$$
- Otherwise, for all $y \in A \setminus \{x\}$, we conclude that $s_y \in (0, \frac{1}{3}]$.
  - If all $y \in A \setminus \{x\}$ actually have size in $(0, \frac{1}{4}]$, then $weight_1(y) \leq \frac{5}{4} \cdot s_y$, and the total size of all items in $A \setminus \{x\}$ is at most $1 - s_x$. Together this gives a total weight of at most
$$s_x + \frac{1}{6} + \frac{5}{4} \cdot (1 - s_x) = \frac{17}{12} - \frac{s_x}{4}\ .$$
This value is maximized when s_x is minimized, and therefore the total weight of the items in A is at most $\frac{16}{12} = \frac{4}{3}$.

– Finally, if there is $y \in A$ such that $s_y \in \mathcal{I}_3$, then we conclude that all items of $A \setminus \{x, y\}$ have size in $(0, \frac{1}{6}]$, and thus their weights are at most $\frac{7}{6}$ times their sizes. This is so because a third item of size in $(\frac{1}{6}, \frac{1}{4}]$ in the bin would imply a triple whose total weight is at least $\frac{1}{2}+\frac{1}{3}+\frac{1}{5} > 1$. This triple will force $weight_1(x) < weight(x)$, contradicting our assumption. Therefore, in this case the total weight of the items in A is at most

$$s_x+\frac{1}{6}+s_y+\frac{1}{12}+(1-s_x-s_y)\cdot\frac{7}{6} = \frac{17}{12}-\frac{s_x+s_y}{6} \le \frac{17}{12}-\frac{7}{72} = \frac{95}{72} < \frac{4}{3}.$$

Suppose that A has an item y of size in $\mathcal{I}_1$.

- If $A = \{y\}$, then the total weight is at most $\frac{7}{6}$.
- Otherwise, let $A = \{y, x_1, x_2, \ldots, x_t\}$, where $s_{x_1} \ge \cdots \ge s_{x_t}$ is the sorted list of other items in the bin. Then x_1 is an item of largest weight according to *weight* among $A \setminus \{y\}$. Let $j \ge 2$ be an integer such that $x_1 \in \mathcal{I}_j$. Let

$$\Delta = \frac{weight(y)+weight(x_1)-1}{3} = \frac{s_y+\frac{1}{6}+s_{x_1}+\frac{1}{j(j+1)}-1}{3}$$

be the reduction value of this bin. We have $weight(y)+weight(x_1) = 3\Delta+1$. The total weight (according to $weight_1$) of the items $y, x_1, \ldots, x_t$ is therefore at most

$$\begin{aligned}
&\frac{j+1}{j}\left(\sum_{i=2}^{t} s_{x_i}\right) + weight(y) + weight(x_1) - \Delta \\
&\le \frac{j+1}{j}(1-s_y-s_{x_1}) + 1 + \frac{2}{3}\left(s_y+\frac{1}{6}+s_{x_1}+\frac{1}{j(j+1)}-1\right) \\
&= -\frac{(j+3)(s_y+s_x)}{3j} + \frac{j+1}{j} + \frac{4}{9} + \frac{2}{3j(j+1)}.
\end{aligned}$$

We use $s_y \ge \frac{1}{2}$ and $s_{x_1} \ge \frac{1}{j+1}$ and get total weight of at most

$$-\frac{3(j+3)^2}{18j(j+1)} + \frac{18(j+1)^2+8j(j+1)+12}{18j(j+1)} = \frac{23j^2+26j+3}{18j(j+1)} = \frac{23}{18}+\frac{1}{6j}\,.$$

If $j \ge 3$, we are done. However, if $j = 2$, then all items but y and x_1 are of size strictly smaller than $\frac{1}{6}$, and thus their weights are at most $\frac{7}{6}$ times their sizes. We get a weight of at most

$$\begin{aligned}
\frac{7}{6}(1-s_y-s_{x_1}) + 1 + \frac{2}{3}\left(s_y+\frac{1}{6}+s_{x_1}+\frac{1}{6}-1\right) &= -\frac{s_{x_1}+s_y}{2}+\frac{31}{18} \\
&\le -\frac{5}{12}+\frac{31}{18} = \frac{47}{36} < \frac{4}{3}\,.
\end{aligned}$$

It is left to consider the case where all items in A are no larger than $\frac{1}{2}$, but $weight_1$ is not equivalent to *weight* for this set of items. Such a bin must contain at least one item of size in $\mathcal{I}_2$ (otherwise, this case is already covered). There are two types of such bins. One option is that A has a single item y of size in $\mathcal{I}_2$. The other option is that A has two items y_1, y_2 of size in $\mathcal{I}_2$.

- Consider the first option and assume first that A contains one item of $\mathcal{I}_2$ and does not contain an item of size in $\mathcal{I}_3$. In this case, we can show that the total

weight according to *weight* of the items in A is at most $\frac{4}{3}$. This is so because, in this case, the total weight is at most $s_y + \frac{1}{6} + \frac{5}{4} \cdot (1 - s_y) = \frac{17}{12} - \frac{s_y}{4} \leq \frac{4}{3}$, where the last inequality holds, since $s_y > \frac{1}{3}$.

- Consider the first option and assume that A contains an item of size in $\mathcal{I}_3$. Let y_1, y_2 be the items of sizes in $\mathcal{I}_2$,$\mathcal{I}_3$ (respectively) and assume that $A = \{y_1, y_2, x_1, \ldots, x_t\}$, where $s_{x_1} \geq \cdots \geq s_{x_t}$, so x_1 is an item of largest weight according to *weight* in $A \backslash \{y_1, y_2\}$. Let $j \geq 3$ be an integer such that $s_{x_1} \in \mathcal{I}_j$. Let

$$\begin{aligned}\Delta &= \frac{weight(y_1) + weight(y_2) + weight(x_1) - 1}{3} \\ &= \frac{s_{y_1} + \frac{1}{6} + s_{y_2} + \frac{1}{12} + s_{x_1} + \frac{1}{j(j+1)} - 1}{3}\end{aligned}$$

be the reduction value of this bin. We have $weight(y_1) + weight(y_2) + weight(x_1) = 3\Delta + 1$. The total weight (according to $weight_1$) of the items in A is therefore at most

$$\begin{aligned}&\frac{j+1}{j}\left(\sum_{i=2}^{t} s_{x_i}\right) + weight(y_1) + weight(y_2) + weight(x_1) - \Delta \\ &\leq \frac{j+1}{j}(1 - s_{y_1} - s_{y_2} - s_{x_1}) + 1 \\ &\quad + \frac{2}{3}\left(s_{y_1} + \frac{1}{6} + s_{y_2} + \frac{1}{12} + s_{x_1} + \frac{1}{j(j+1)} - 1\right) \\ &= -\frac{(j+3)(s_{y_1} + s_{y_2} + s_x)}{3j} + \frac{j+1}{j} + \frac{1}{2} + \frac{2}{3j(j+1)}.\end{aligned}$$

We use $s_{y_1} \geq \frac{1}{3}$, $s_{y_2} \geq \frac{1}{4}$, and $s_{x_1} \geq \frac{1}{j+1}$ and get a total weight of at most

$$\begin{aligned}&-\frac{(j+3)(7j+19)}{36j(j+1)} + \frac{36(j+1)^2 + 18j(j+1) + 24}{36j(j+1)} \\ &= -\frac{7j^2 + 40j + 57}{36j(j+1)} + \frac{36j^2 + 72j + 36 + 18j^2 + 18j + 24}{36j(j+1)} = \frac{47}{36} + \frac{1}{12j} \leq \frac{4}{3}.\end{aligned}$$

- Consider the second option. Assume that $A = \{y_1, y_2, x_1, \ldots, x_t\}$, where $s_{x_1} \geq \cdots \geq s_{x_t}$, so x_1 is an item of largest weight according to *weight* in $A \setminus \{y_1, y_2\}$. Let $j \geq 3$ be an integer such that $s_{x_1} \in \mathcal{I}_j$. Let

$$\begin{aligned}\Delta &= \frac{weight(y_1) + weight(y_2) + weight(x_1) - 1}{3} \\ &= \frac{s_{y_1} + \frac{1}{6} + s_{y_2} + \frac{1}{6} + s_{x_1} + \frac{1}{j(j+1)} - 1}{3}\end{aligned}$$

be the reduction value of this bin. We have $weight(y_1) + weight(y_2) + weight(x_1) = 3\Delta + 1$. The total weight (according to $weight_1$) of the items

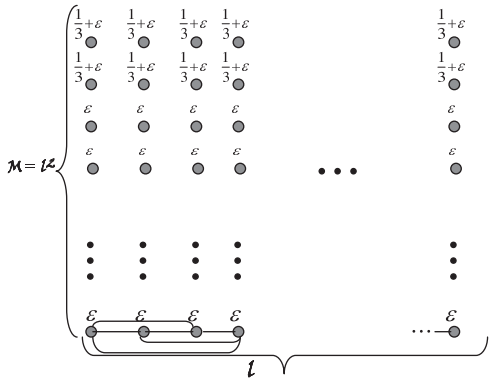


FIG. 2. *The tight example for the class $\mathcal{C}$. The items of each column are packed in one bin by the optimal solution. In the algorithm, the preprocessing step removes the three top items from each column, and the remaining items all receive the same color, except for $\ell - 1$ items of the last row.*

in A is therefore at most

$$\begin{aligned}
&\frac{j+1}{j}\left(\sum_{i=2}^{t} s_{x_i}\right) + weight(y_1) + weight(y_2) + weight(x_1) - \Delta \\
&\leq \frac{j+1}{j}(1 - s_{y_1} - s_{y_2} - s_{x_1}) + 1 \\
&\quad + \frac{2}{3}\left(s_{y_1} + \frac{1}{6} + s_{y_2} + \frac{1}{6} + s_{x_1} + \frac{1}{j(j+1)} - 1\right) \\
&= -\frac{(j+3)(s_{y_1} + s_{y_2} + s_x)}{3j} + \frac{j+1}{j} + \frac{5}{9} + \frac{2}{3j(j+1)}.
\end{aligned}$$

We use $s_{y_1}, s_{y_2} \geq \frac{1}{3}$, and $s_{x_1} \geq \frac{1}{j+1}$ and get a total weight of at most

$$\begin{aligned}
&-\frac{(j+3)(2j+5)}{9j(j+1)} + \frac{9(j+1)^2 + 5j(j+1) + 6}{9j(j+1)} = -\frac{2j^2 + 11j + 15}{9j(j+1)} \\
&+ \frac{9j^2 + 18j + 9 + 5j^2 + 5j + 6}{9j(j+1)} = \frac{4}{3}\,.
\end{aligned}$$

We next show that our analysis of algorithm GREEDY PREPROCESSING is tight. Let M and ℓ be large constants and $\varepsilon = \frac{1}{3M}$. To construct the set of items, we do as follows. We use two sequences of ℓ items each, $a_1, \ldots, a_\ell, b_1, \ldots, b_\ell$, where each of these items has size $\frac{1}{3} + \varepsilon$. Furthermore, we have $(M-2)\ell$ additional items $c_{i,j}$, $1 \leq i \leq \ell$, $1 \leq j \leq M-2$, each of size ε. The conflict graph induces a clique on the ℓ items $c_{i,M-2}$ and contains no further edges (see Figure 2).

An optimal solution is given by ℓ independent sets $U_i = \{a_i, b_i, c_{i,1}, \ldots, c_{i,M-2}\}$ with total size of exactly one for each set. Thus $\mathrm{OPT} = \ell$.

The preprocessing step finds ℓ sets of triples, which are $\{a_i, b_i, c_{i,1}\}$. Next, a coloring with ℓ colors which is found for the remaining items consists of one independent set which contains all items $c_{i,j}$ for $1 \leq i \leq \ell$, $2 \leq j \leq M-3$, and additionally contains $c_{1,M-2}$. Each other independent set contains a single item $c_{i,M-2}$ for $2 \leq i \leq \ell$. The number of bins used to color the first independent set is $\lceil \frac{\ell(M-4)+1}{3M} \rceil$, since $\frac{\ell(M-4)+1}{3M}$ is the total size of these items. Each other independent set consumes one additional

bin, thus in total we get at least $2\ell - 1 + \frac{\ell}{3} - \frac{4\ell}{3M}$ bins. It can be seen that, for $M = \ell^2$ and $\ell >> 1$, the ratio becomes arbitrarily close to $\frac{7}{3}$. □

Consider now alternative algorithms which remove triples in a different way in the first step of the algorithm. That is, instead of picking a triple or pair greedily at each time, it applies a different heuristic. The example actually shows that even if this step is processed optimally (using an exponential time algorithm), that is, if a set of triples and pairs with total maximum weight is removed, still the performance cannot be improved. Therefore, we keep this step simple and do not apply advanced methods (such as local search; see [1]) that allow the removal of a set of triples and pairs of a larger weight.

3.3. Bipartite graphs. In this section, we can assume that $\mathrm{OPT} \geq 2$ and that the conflict graph contains at least one edge. This holds since $\mathrm{OPT} = 1$ means that the conflict graph is empty, and the total size of the items is at most 1. Each one of these properties can be easily checked in polynomial time, and if both conditions hold, the algorithm returns an optimal solution. Moreover, if the conflict graph is empty, the problem reduces to standard bin packing, for which FFD has an approximation ratio of $\frac{3}{2}$ [29]. Since the algorithm presented in this section has an approximation ratio larger than $\frac{3}{2}$, we can assume that the conflict graph contains at least one edge.

In [17], the following simple algorithm was analyzed for graphs which have a nonempty conflict graph.

The algorithm Two-Set (TS) of [17].

Find a coloring of the conflict graph with two colors. Pack each color class using a simple heuristic (Next-Fit, First-Fit (FF), or FFD).

It was shown in [17] that TS is a 2-approximation for all of these suggested heuristics. (Using FFD instead of Next-Fit does not give a better approximation ratio.)

We design an algorithm which gives special treatment to some of the problematic cases and thus get a $\frac{7}{4}$-approximation.

We start with an analysis of the algorithm TS (with FFD), as a function of the value OPT. Let A and B denote the sets of the items of the two colors. Let $\ell(A)$ and $\ell(B)$ denote the numbers of bins packed by FFD for each of the two sets, and let $\mathrm{OPT}(X)$ denote the cost of an optimal solution for a set X. Clearly, we have $s(X) \leq \mathrm{OPT}(X) \leq \mathrm{OPT}$ for $X = A, B$, and also $\mathrm{OPT} \geq s(A) + s(B)$.

As stated above, Simchi-Levi [29] proved that, for any input Y, the solution of FFD on this output satisfies $\mathrm{FFD}(Y) \leq \frac{3}{2}\mathrm{OPT}(Y)$. Therefore, if the size of one of the sets (without loss of generality, the set A) is small enough, namely, this set fits into one bin $s(A) \leq 1$, we get $TS \leq \mathrm{FFD}(B) + 1 \leq \frac{3}{2}\mathrm{OPT} + 1$.

Otherwise, if for both sets, the output of FFD created at least one bin where the smallest item that opens a new bin is in the interval $(0, \frac{1}{3}]$. Then, for each set A and B, all bins but the last one are occupied by more than $\frac{2}{3}$, and the sum of items in the two last bins together is more than 1. We get for $X = A, B$, $s(X) > \frac{2}{3}(\ell(X) - 2) + 1$. Thus $TS \leq \ell(A) + \ell(B) < \frac{3}{2}\mathrm{OPT} + 1$.

Suppose next that both sets A and B do not have a bin opened by an item with size in the interval $(0, \frac{1}{3}]$. Then, we remove all items smaller than $\frac{1}{3}$ from the input. Clearly, the output does not change. Each bin contains an item of size in $(\frac{1}{2}, 1]$ (and possibly one smaller item as well) or two items in the interval $(\frac{1}{3}, \frac{1}{2}]$, except possibly the last bin for each set, that may contain a single item of this last interval. Let Z denote the number of the items of size in $(\frac{1}{2}, 1]$ in $A \cup B$, and let V denote the number of items from $A \cup B$ with size in the interval $(\frac{1}{3}, \frac{1}{2}]$. Therefore,

$TS \le Z + \frac{V-2}{2} + 2 = Z + \frac{V}{2} + 1$. However, for any packing and thus for an optimal one, we have that each bin contains at most one item with size larger than $\frac{1}{2}$, and at most two items with size larger than $\frac{1}{3}$, thus we have $\mathrm{OPT} \ge Z$ and $\mathrm{OPT} \ge \frac{Z+V}{2}$. We get $TS \le \frac{Z+V}{2} + \frac{Z}{2} + 1 \le \frac{3}{2}\mathrm{OPT} + 1$.

We are left with the case where (without loss of generality) the set A contains a bin opened by an item in $(0, \frac{1}{3}]$, and B does not. If A does not contain a bin opened by an item of size in $(0, \frac{1}{4}]$, we can remove all items smaller than $\frac{1}{4}$ from the input and get the same output. Let Z denote again the number of items in $(\frac{1}{2}, 1]$ and V denote the number of items in $(\frac{1}{4}, \frac{1}{2}]$. We now argue that $V \le 3(\mathrm{OPT} - Z) + Z = 3\mathrm{OPT} - 2Z$. This last inequality holds, since a bin with an item larger than $\frac{1}{2}$ can contain at most one item larger than $\frac{1}{4}$, and any other bin can contain at most three such items. Therefore, $TS \le Z + \frac{V-2}{2} + 2 \le Z + \frac{3}{2}\mathrm{OPT} - Z + 1 = \frac{3}{2}\mathrm{OPT} + 1$.

Finally, we need to consider the case where A contains at least one bin opened by an item of size in $(0, \frac{1}{4}]$, and B does not have a bin opened by an item whose size is at most $\frac{1}{3}$. Thus all bins of A but the last one are occupied by more than $\frac{3}{4}$. We get $s(A) > \frac{3}{4}(\ell(A) - 2) + 1$ and $s(B) > \frac{1}{2}\ell(B)$. The last inequality holds for any Any-Fit type algorithm and for FFD in particular. Moreover, note that the packing of B is an optimal one. This can be proved using simple exchange arguments (see [29]). Thus we have $\ell(B) \le \mathrm{OPT}$. We get $\mathrm{OPT} \ge s(A) + s(B) > \frac{3}{4}\ell(A) + \frac{1}{2}\ell(B) - \frac{1}{2}$. Thus $TS < \frac{4}{3}\mathrm{OPT} + \frac{2}{3} + \frac{1}{3}\mathrm{OPT} = \frac{5}{3}\mathrm{OPT} + \frac{2}{3}$. Since both OPT and TS are integers and our last inequality is a strict inequality, we get $TS \le \frac{5}{3}\mathrm{OPT} + \frac{1}{3}$.

We can prove the following lemma.

LEMMA 13. *If* $\mathrm{OPT} \ge 3$ *then the algorithm above satisfies* $TS \le \frac{7}{4}\mathrm{OPT}$*, and this bound is tight when* $\mathrm{OPT} = 4$*.*

Proof. We obtained two bounds, and since TS is integer, we conclude that $TS \le \max\{\lfloor \frac{3}{2}\mathrm{OPT} + 1 \rfloor, \lfloor \frac{5}{3}\mathrm{OPT} + \frac{1}{3} \rfloor\}$. If $\mathrm{OPT} \ge 4$, we get $\frac{3}{2}\mathrm{OPT} + 1 \le \frac{7}{4}\mathrm{OPT}$ and $\lfloor \frac{5}{3}\mathrm{OPT} + \frac{1}{3} \rfloor \le \frac{7}{4}\mathrm{OPT}$. For $\mathrm{OPT} = 3$, we get $TS \le 5 = \frac{5}{3}\mathrm{OPT}$.

To see that this bound is tight, consider the following example. Let $\varepsilon > 0$ be a small number and define the following set of item sizes $A = \{\frac{1}{4} + \varepsilon, \frac{1}{4} + \varepsilon, \frac{1}{4} + \varepsilon, \frac{1}{4} + \varepsilon, \frac{1}{4} - 2\varepsilon, \frac{1}{4} - 2\varepsilon, \frac{1}{4} - 2\varepsilon, \frac{1}{4} - 2\varepsilon\}$, and $B = \{\frac{1}{2} + \varepsilon, \frac{1}{2} + \varepsilon, \frac{1}{2} + \varepsilon, \frac{1}{2} + \varepsilon\}$. Assume that the conflict graph has a single edge between one item of size $\frac{1}{2} + \varepsilon$ and one of the items of size $\frac{1}{4} + \varepsilon$. Then an optimal solution has four bins, each of which has one item of size $\frac{1}{2} + \varepsilon$, one item of size $\frac{1}{4} + \varepsilon$, and one item of size $\frac{1}{4} - 2\varepsilon$. Clearly, the two items which have a conflict do not share a bin. However, assume that the coloring into two colors partitions the items into A and B. Then, $\mathrm{FFD}(A) = 3$ and $\mathrm{FFD}(B) = 4$. Therefore, for this example, $\mathrm{OPT} = 4$, and the algorithm returns a solution that uses seven bins. ∎

Note that if we were considering the asymptotic approximation ratio rather than the absolute approximation ratio, this analysis of TS proves that its asymptotic approximation ratio is at most $\frac{3}{2}$.

As we can see, the only case which is left is $\mathrm{OPT} = 2$, which requires special treatment. This case can be identified by a solution returned by TS of cost 4. Clearly, such solutions can be achieved also for $\mathrm{OPT} = 3$ and $\mathrm{OPT} = 4$. We define an algorithm and prove that it succeeds if $\mathrm{OPT} = 2$. Thus, if it fails, then $\mathrm{OPT} \ge 3$, which means that the original solution already does not violate the approximation ratio $\frac{7}{4}$ which we would like to prove. We call this algorithm MODIFIED TWO-SET (MTS).

If $\mathrm{OPT} = 2$, this means that it is possible to color the input using two colors and pack each independent set into a single bin. If the conflict graph is connected, there is a unique way to color the items, and thus this optimal packing can be achieved.

However, a bipartite disconnected graph has more than one possible coloring with two colors, since the roles of the two colors in each connected component can be swapped. The first step of MTS is to color each connected component using two colors. Let z be the number of components and denote the items of component i by V_i. For each $1 \le i \le z$, this gives two sets A_i and B_i, such that $A_i \cup B_i = V_i$, $A_i \cap B_i = \emptyset$, and $s(A_i) \ge s(B_i)$. Each set in $\{A_i, B_i\}$ contains the vertices of V_i that share one color. Define $p_i = s(A_i) - s(B_i)$ and assume that the values p_i are sorted such that $p_1 \ge p_2 \ge \cdots \ge p_z$. Let $q_i = s(B_i) = s(A_i) - p_i$. Use the sizes p_i to define a scheduling problem on two machines. Run LPT (Longest Processing Time First) on this input. This means that two empty sets of indices $\mathcal{V}_1$ and $\mathcal{V}_2$ are initialized. Then starting from $i = 1$, assign i for $i = 1, \ldots, z$ to the set ($\mathcal{V}_1$ or $\mathcal{V}_2$) whose total sum (of the values p_j, where j is a member of the set) is minimal. Graham [12] defined and analyzed this algorithm for an arbitrary number of machines (subsets). It is not difficult to see that when the algorithm terminates, $|\sum_{i\in\mathcal{V}_1} p_i - \sum_{i\in\mathcal{V}_2} p_i| \le p_1$ holds. For $1 \le i \le z$, we define a coloring using two colors (which are defined by the sets C and D) as follows. If $i \in \mathcal{V}_1$, then MTS assigns the items in A_i to C and the items in B_i to D. Otherwise, it assigns the items in B_i to C and the items in A_i to D. This assignment means that $s(C) = \sum_{i\in\mathcal{V}_1} p_i + \sum_{i=1}^{z} q_i$ and $s(D) = \sum_{i\in\mathcal{V}_2} p_i + \sum_{i=1}^{z} q_i$. Thus we have $|s(C) - s(D)| = |\sum_{i\in\mathcal{V}_1} p_i - \sum_{i\in\mathcal{V}_2} p_i| \le p_1$. Assume (without loss of generality) that $s(C) \ge s(D)$. Since $\mathrm{OPT} = 2$, $s(C) + s(D) \le 2$. Thus $s(D) \le 1$, and all of the items assigned to D fit into a single bin. Let i_1 be the maximum index in $\mathcal{V}_1$. Now remove all items of A_{i_1} (where $s(A_{i_1}) = p_{i_1} + q_{i_1}$) from C. We get a total sum of less than $s(D) \le 1$, and thus the remaining items of C fit into one additional bin. Finally, the items of A_{i_1} need to be packed. If indeed $\mathrm{OPT} = 2$, then the items of A_{i_1} are packed into a single bin in any optimal solution and thus can be packed into a third bin. This gives a total of three bins.

We summarize the action of MODIFIED TWO-SET as follows.

Algorithm. MODIFIED TWO-SET (MTS)

Run TS on the input. If the output consists of four bins, apply the following algorithm, which assumes $\mathrm{OPT} = 2$. If the algorithm does not fail (that is, each one of the three bins it creates is valid, in terms of total size), give its output. Otherwise, give the output of TS as output.

1. Color each connected component of the conflict graph using two colors. Denote the vertices of connected component $1 \le i \le z$ by V_i. Assume that the two independent sets resulting from connected component V_i are A_i,B_i, where $s(A_i) \ge s(B_i)$ and that the sets are sorted so that the values $p_i = s(A_i) - s(B_i)$ are nonincreasing.
2. Apply the algorithm LPT on the values p_i to partition the (indices of the) sets V_i into two subsets $\mathcal{V}_1$ and $\mathcal{V}_2$, where $\sum_{i\in\mathcal{V}_1} p_i \ge \sum_{i\in\mathcal{V}_2} p_i$. Let i_1 be the maximum index in $\mathcal{V}_1$, then $\sum_{i\in\mathcal{V}_1} p_i - \sum_{i\in\mathcal{V}_2} p_i \le p_{i_1}$.
3. Pack all items of $\bigcup_{i\in\mathcal{V}_1} A_i \cup \bigcup_{i\in\mathcal{V}_2} B_i$, except for items of A_{i_1} into one bin.
4. Pack all items of $\bigcup_{i\in\mathcal{V}_2} A_i \cup \bigcup_{i\in\mathcal{V}_1} B_i$ into one bin.
5. Pack all items of A_{i_1} into one bin.

We proved the following theorem.

THEOREM 14. *Algorithm* MTS *has an approximation ratio of exactly* $\frac{7}{4}$.

Proof. We showed that if $\mathrm{OPT} = 2$, the process above succeeds to pack the input into three bins. Otherwise, the theorem follows from Lemma 13. ∎

4. Online algorithms. In this section, we discuss online algorithms for interval graphs.

In the online problem, items arrive one by one. When an item arrives, the following information is revealed to the algorithm: the size of the item, the existing edges in the conflict graph between the new item, and the previous items. The algorithm has to pack the new item before the next item arrives. Once an item is packed in a bin, its location cannot be changed. For interval graphs, an item arrives together with its realization, that is, the coordinates of the interval on the real line, which corresponds to this item.

For many classes of graphs, the online coloring problem is hard to approximate. Note that online coloring is a special case of BPC, where all item sizes are zero.

Consider, e.g., the problem on trees. Gyárfás and Lehel [13] proved a deterministic lower bound of $\Omega(\log n)$ on the online coloring of bipartite graphs on n vertices, which holds already for trees. Lovász, Saks, and Trotter [22] showed an online coloring algorithm which colors such a graph (which is 2 colorable) using $O(\log n)$ colors. This immediately implies an online coloring algorithm for BPC on bipartite graphs, which is optimal up to a constant multiplicative factor on the competitive ratio. This algorithm $\mathcal{A}$ uses the algorithm of [22] to color the conflict graph using C colors. Then the items of each color class are packed independently using some reasonable algorithm for bin packing, e.g., Next-Fit. We get that for each color class i, which contains ℓ_i bins, the total size of items S_i of color class i is more than $\frac{\ell_i - 1}{2}$ (since no two consecutive bins can be combined). We get that

$$\mathcal{A} \leq \sum_{i=1}^{C} \ell_i < \sum_{i=1}^{C} (2S_i + 1) \leq 2\mathrm{OPT} + C \leq O(\log n)\mathrm{OPT} \ .$$

Since the same can be applied for any graph class for which no constant competitive algorithm exists, we focus on a graph class for which such an algorithm exists, namely, interval graphs. Kierstead and Trotter [20] constructed an online coloring algorithm for interval graphs, which uses at most $3\omega - 2$ colors, where ω is the maximum clique size of the graph. They also presented a matching lower bound of $3\omega - 2$ on the number of colors in a coloring of an arbitrary online coloring algorithm.

The main idea of the algorithm of [20] is the creation of "levels." At the time of the arrival of an interval, it is classified into a level as follows. Denote by A_k the union of the sets of intervals which currently belong to all levels $1, \ldots, k$. Intervals are classified so that the largest cardinality clique in A_k is of size k. Thus, A_1 is simply a set of nonintersecting intervals. On the arrival of an interval, the algorithm finds the smallest k such that the new interval can join level k, without violating the rule above. It can be shown that each level can be colored using two colors by an offline algorithm. Since the algorithm defined here is online, such a coloring cannot be found, in general. However, it is shown in [20] that at most three colors are required for each such level, and a coloring using three colors can be found by applying FF on each level (with disjoint sets of colors). Moreover, the first level can always be colored using a single color, and ω is equal exactly to the number of levels. Thus a total number of colors, which is at most $3(\omega - 1) + 1 = 3\omega - 2$, is used.

Note that the chromatic number of interval graphs equals to the size of a maximum clique, which is equivalent in the case of interval graphs to the largest number of intervals that intersect any point (see [18, 11]). The technique above implies a 5-competitive algorithm. We can show that using FF instead of Next-Fit for coloring each class slightly improves this bound.

A combined algorithm for interval graphs.
Use the algorithm of [20] to color the arriving intervals and FF to pack the items of each color.

THEOREM 15. *The combined algorithm has a competitive ratio of* 4.7.

Proof. The proof is similar to the proof of [17] for perfect graphs. We can use the well-known weight function $\hat{w}$ defined for FF already in [9] (see also page 219 in [3]). This weight function is defined as follows:

$$\hat{w}(r) = \begin{cases} \frac{6r}{5} & \text{if } 0 \le r < \frac{1}{6}, \\ \frac{9r}{5} - \frac{1}{10} & \text{if } \frac{1}{6} \le r < \frac{1}{3}, \\ \frac{6r}{5} + \frac{1}{10} & \text{if } \frac{1}{3} \le r < \frac{1}{2}, \\ 1 & \text{if } \frac{1}{2} \le r \le 1. \end{cases}$$

It was shown that, for a set of items J on which FF is applied, we have $\ell(J) \le W(J) + 1$, where $W(J)$ denotes the total weight of items in J and $\ell(J)$ is the number of bins in the packing of J by FF. On the other hand, if we remove the incompatibility constraint (i.e., assume that the conflict graph has no edges) and let OPT' denote an optimal solution for this instance, we have $W(I) \le 1.7\mathrm{OPT}' \le 1.7\mathrm{OPT}$. Let I_j denote the items colored by the algorithm of [20] by color j, and let C denote the number of colors which is used. This algorithm colors the items with at most $3\chi(G) - 2 \le 3\mathrm{OPT} - 2$ colors; therefore $C < 3\mathrm{OPT}$, and thus we get $\mathcal{A} \le \sum_{j=1}^{C} \ell(I_j) < \sum_{j=1}^{C} (W(I_j) + 1) = W(I) + C \le 4.7\mathrm{OPT}$.

Based on the examples of [17, 19] and [20], we can show a sequence of instances whose competitive ratio is arbitrarily close to 4.7. The tight example simply combines the two previously known tight examples. The example of [20] is used as a black box. Given an interval $[x, y]$, we use the fact that it is possible to construct an example, which is a sequence of intervals contained in $[x, y]$, where the size of the largest clique is ω, for a given value of ω, and for which the algorithm of [20] uses the $3\omega - 2$ colors $1, 2, \ldots, 3\omega - 2$.

We fix a value of $\mathrm{OPT} = k + 2$ and a value of $\varepsilon >> \delta > 0$ that is small enough, and we let $\ell = \frac{k}{10}$. The construction is composed of two phases. In the first phase, we repeat the construction of [20] with a maximum clique size $k+1$, where all items that correspond to the vertices which the intervals represent have zero size. We repeat the construction to have $3k$ copies of this construction that use the same set of $3k + 1$ colors. Since the intervals are colored by the algorithm of [20], all structures are all colored in the same way. We call these $3k + 1$ colors $1, 2, \ldots, 3k + 1$.

At the end of the first phase, we call the subintervals of the real line such that each one of them contains a copy of the construction of [20] that uses the color set $\{1, 2, \ldots, 3k + 1\}$, $[x_i, y_i]$ for $1 \le i \le 3k$. Each future interval will completely contain an interval $[x_i, y_i]$ and will not have any overlap with other intervals from the first phase or the second phase.

We next present $3k$ disjoint intervals one by one (i.e., this set of intervals is an independent set of G). Each of these intervals contains exactly one interval $[x_i, y_i]$ of the first phase, and therefore these intervals cannot be colored using colors $1, 2, \ldots, 3k+1$, by the algorithm.

- The first k intervals, which are denoted by $a_{i,p}$ for $i = 1, \ldots, 10$ and $p = 1, \ldots, \ell$, have sizes according to the following: $a_{i,p}$ has size $\frac{1}{6} + \frac{\varepsilon}{3^p} - \delta$ for $1 \le i \le 5$, and otherwise, it has size $\frac{1}{6} - \frac{\varepsilon}{3^{p+1}} - \delta$. These k intervals are

introduced according to the following order:

$$a_{1,1}, a_{2,1}, a_{3,1}, a_{6,1}, a_{7,1}, a_{4,1}, a_{5,1}, a_{8,1}, a_{9,1}, a_{10,1}, a_{1,2}, \ldots .$$

- The next k intervals are denoted by $b_{i,p}$ for $i = 1, \ldots, 10$ and $p = 1, \ldots, \ell$. Their sizes are defined according to the following: the size of $b_{i,p}$ for $1 \le i \le 5$ is $\frac{1}{3} + \frac{\varepsilon}{3^{p-1}} - \delta$, and otherwise, the size of $b_{i,p}$ is $\frac{1}{3} - \frac{\varepsilon}{3^p} - \delta$. These k intervals are introduced in the following order: $b_{1,1}, b_{6,1}, b_{2,1}, b_{7,1}, b_{3,1}, b_{8,1}, b_{4,1}, b_{9,1}, b_{5,1}, b_{10,1}, b_{1,2}, \ldots$.
- The last k intervals are denoted by c_i for $i = 1, 2, \ldots, k$, and each of these has size of $\frac{1}{2} + \delta$.

Note that the coloring algorithm of [20] will color all of the intervals of the second phase using color $3k + 2$, and therefore we need to compute the number of bins that the FF algorithm uses in order to pack all of these items. Applying FF will open ℓ bins of the following type: $\{a_{1,p}, a_{2,p}, a_{3,p}, a_{6,p}, a_{7,p}\}$ for $p = 1, 2, \ldots, \ell$. It will use another ℓ bins of the following type: $\{a_{4,p}, a_{5,p}, a_{8,p}, a_{9,p}, a_{10,p}\}$. There will be another 5ℓ bins each containing $\{b_{i,p}, b_{i+5,p}\}$ for $i = 1, 2, \ldots, 5$ and $p = 1, 2, \ldots, \ell$. Last, the algorithm will pack each of the items c_i separately using another 10ℓ bins. The total number of bins used to pack the second phase intervals is $17\ell = 1.7k$. By adding the $3k + 1$ colors that are used to color the first phase intervals, we get a solution whose cost is $4.7k + 1$ (see Figure 3).

It remains to show that the optimal solution costs $k + 2$. We first show a packing of the second phase intervals using exactly $k + 2$ bins. We use 5ℓ bins each containing $\{a_{i,p}, b_{5+i,p}, c_{5(p-1)+i}\}$ for $i = 1, \ldots, 5$ and $p = 1, \ldots, \ell$. We use another $5\ell - 10$ bins each containing $\{a_{5+i,p-2}, b_{i,p}, c_{5(p+\ell-3)+i}\}$. We use another five bins containing $\{c_{10(\ell-1)+i}, b_{i,1}\}$ for $i = 1, \ldots, 5$, and another five bins containing $\{c_{10(\ell-1)+5+i}, b_{i,2}\}$ for $i = 1, \ldots, 5$. We have another two bins where the first one consists of $\{a_{6,\ell-1}, a_{7,\ell-1}, a_{8,\ell-1}, a_{9,\ell-1}, a_{10,\ell-1}\}$ and the second bin consists of $\{a_{6,\ell}, a_{7,\ell}, a_{8,\ell}, a_{9,\ell}, a_{10,\ell}\}$. Then, for each one of the $3k$ constructions of the first phase, there is a single color out of the list of $k + 2$ colors (or bins) that is used in the coloring of the second phase interval that contains this construction. Therefore, we have a set of $k + 1$ colors that can be used to color the intervals of the first phase construction; these are all the colors except the one that is used for the interval of the second phase. It is possible to complete the coloring of the construction using a set of $k + 1$ colors, as the maximum clique size in each such construction is $k + 1$. Therefore, we are able to color the entire set of intervals using $k + 2$ colors, each of them containing items of total size at most one. Hence, $\mathrm{OPT} = k + 2$. □

We can show that an algorithm of much smaller competitive ratio does not exist.

THEOREM 16. *The competitive ratio of any online algorithm for* BPC *on interval graphs has a competitive ratio of at least* $\frac{155}{36} \approx 4.30556$.

In order to prove this theorem, we prove the following two lemmas.

The proof of the next lemma uses a construction similar to the lower bound given in [20]. In our construction, we prove a lower bound of $3 + c$ on the competitive ratio. We assume that we know the optimal value $\mathrm{OPT} = k$, and thus we are allowed to use a set of at most $(3 + c)k$ bins. The construction is composed of two phases. In the first phase, we introduce intervals such that the corresponding items have size zero. We call the intervals that correspond to these items zero sized intervals, where the size of an interval is unrelated to its length. The *size of an interval* is simply the size of the item that is associated with the vertex in the conflict graph that this interval represents. The maximum cardinality clique among these intervals will be $k - 1$, and

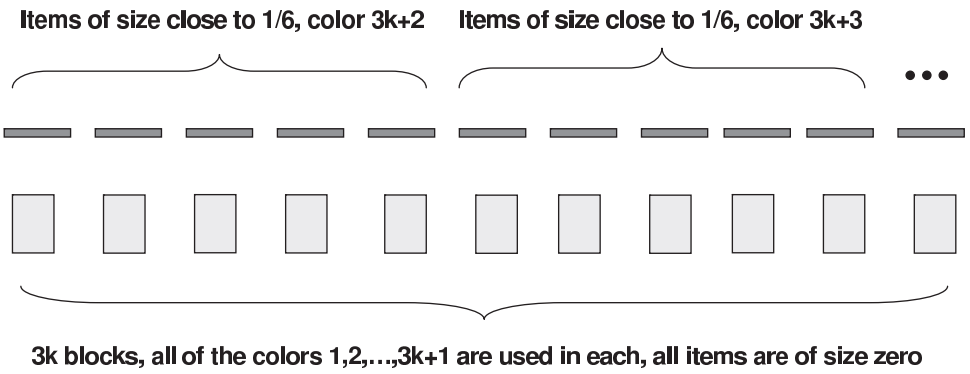


FIG. 3. *A schematic illustration of the behavior of the algorithm in the example of the proof of Theorem* 15. *The large rectangles represent the first part of the input. The thin rectangles represent the additional intervals. Only the first set of such intervals is shown.*

the algorithm is forced to use at least $3k-5$ bins, denoted by $1, 2, \ldots, 3k-5$. This part of the construction follows a similar framework as the lower bound in [20]. The second phase of our construction is similar to the second phase of the tight example in the proof of Theorem 15.

LEMMA 17. *Let c be a lower bound on the asymptotic competitive ratio of any online algorithm for standard bin packing, which knows the value* OPT *in advance. Then the competitive ratio for any online algorithm for* BPC *on interval graphs is at least $3+c$.*

Proof. Our construction is composed of two phases. During the first phase, we *shrink* some parts of the line into single points. We next define the operation of shrinking an interval $[a,b]$ into a point p. Such an operation is performed only if the interval $[a,b]$ and each previously presented interval $[x,y]$ satisfy that either $[x,y] \subseteq [a,b]$ or $[x,y]$ and $[a,b]$ are disjoint. The shrinking operation is a mapping of the real line onto itself such that the points in $[a,b]$ are mapped to $p=a$, a point q such that $q<a$ is mapped to q, and a point $q'>b$ is mapped to $q'-(b-a)$. We note that at the time of the shrinking, each existing interval does not contain p as a (strictly) inner point. The shrinking operation does not affect the sizes of intervals which are not shrunk.

Therefore, every interval presented in the past which is contained in $[a,b]$ is also shrunk into p, and thus such a point inherits a list of colors that such intervals received. These colors cannot be assigned to any interval that contains the point p.

If at some time during the construction, an algorithm uses more than $U = 4\left\lceil \frac{(3+c)k}{4} \right\rceil$ bins, the construction is stopped. Therefore, we assume that the algorithm is initially given U bins, or equivalently, a palette of U colors. As soon as all of these colors are used, the proof is complete, and no further phases are presented. This is just one stopping condition; we may stop the first phase of the sequence earlier as well, after the algorithm used $3k-5$ colors from the set of colors $\{1, 2, \ldots, U\}$. In this case, a second phase is used.

The first phase is composed of at most $k-1$ steps that we define as follows. At the first step of the first phase, we introduce S disjoint (unit length and zero sized) intervals, where $S = U^{3k}X$ and X is fixed later. Since the algorithm is using at most U colors, this means that there exists a set of at least $\frac{S}{U}$ intervals that share the exact same color $\mathbf{c}$. We shrink all intervals into single points. Later steps result in additional points.

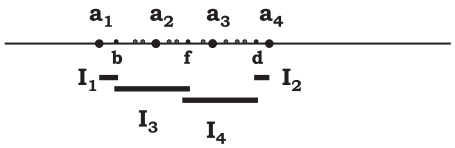


FIG. 4. *The set of intervals in the case that I_1 and I_2 receive the same color.*

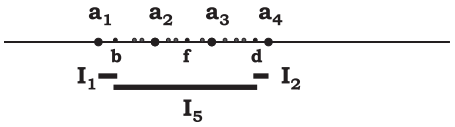


FIG. 5. *The set of intervals in the case that I_1 and I_2 receive distinct colors.*

We now define step $j > 1$ of the first phase. The steps are constructed in a way that, in the beginning of step $j \geq 1$, there is a set of at least $U^{3(k-j)+2}X$ points that contain a given subset of the U colors (which clearly holds after the first step). These points are called *points of interest*. After the first step, all points of interest contain one fixed color.

Note that at all times, there may exist some other points containing some subsets of colors. All of these points are called *void points*.

At the beginning of a step $j > 1$, we partition $U^{3(k-j+1)+2}X$ of the points of interest into consecutive sets of four, $X\frac{U^{3(k-j+1)+2}}{4}$ sets in total. All other points of interest, if they exist, that do not participate in this become void points.

We next define additional intervals, increasing the size of the largest cardinality clique (with respect to the number of intervals, i.e., ignoring sizes) by exactly 1. Given a set of four points of interest a_1, a_2, a_3, a_4 (listed from left to right), let b be a point between a_1 and a_2, which is not a void point, and let d be a point between a_3 and a_4, which is not a void point. Let f be a point between a_2 and a_3, which is not a void point. We introduce the zero sized intervals $I_1 = [a_1, b]$ and $I_2 = [d, a_4]$.

If they both receive the same color, we introduce the zero sized intervals $I_3 = [b, f]$ and $I_4 = [f, d]$. The interval I_3 intersects with a_2 and with I_1. The second interval I_4 intersects I_3, a_3, and I_2; therefore two new colors must be used. In total, three new colors were used (See Figure 4).

If I_1 and I_2 receive distinct colors, we introduce the zero sized interval $I_5 = [b, d]$. Interval I_5 intersects with I_1, I_2, a_2, and a_3, and thus gets a new color. In total, three new colors were used (See Figure 5).

We shrink every such interval $[a_1, a_4]$ into a single point. Each of the new shrunk points received three new colors.

Note that throughout all of the steps of the first phase, we use at most U colors (otherwise, this phase is stopped), and each new shrunk point receives exactly three new colors, since four intervals are introduced only if the first two received the same color, and otherwise, three intervals of three different colors are introduced. However, even though the initial sets of colors are the same for all points of interest, the sets of the three new colors may differ for different sets of four points of interest from the previous step. Since the number of choices of three colors out of at most U colors is $\frac{U^3}{6}$, there are less than $\frac{U^3}{6}$ options to choose a set of three new colors.

We can choose at least $\frac{(XU^{3(k-j+1)+2})/4}{U^3/6} > XU^{3(k-j)+2}$ points having the exact same set of used colors. The points containing these exact sets of colors become the

points of interest of the next step, and the others become void points of the next step. Points that are void points of previous steps and are not contained in shrunk intervals remain void points. Note that the points where the new intervals intersect are points with no previous intervals, and therefore the clique size increases by exactly 1.

The result of each step is therefore an increase of 3 in the number of colors that each point of interest contains, whereas the cardinality of the largest clique increased by 1.

At the end of the $k-1$th step, the first phase where we have a set of zero sized intervals ends, and the online algorithm used at least $3k-5$ colors to color these intervals, since the first step creates points of interest with one color, and each additional step increases the number of colors attributed to points of interest by 3. At this time, the number of points of interest, where each point contains the same set of $3k-5$ colors, is at least XU^2.

Assume that c is a lower bound on the asymptotic competitive ratio of any online algorithm for standard bin packing, which knows the value $\mathrm{OPT} = k$ in advance. This means that, for any $\varepsilon > 0$, there exists a sequence of at most $f(k)$ items (defined by their sizes) such that $\mathrm{OPT} = k$, and the algorithm is forced to use at least $(c-\varepsilon)k$ bins. For example, in the proof of Lemma 18, $f(k) = 3k$. (Note that even if the lower bound construction uses an infinite number of items, we can always use a subsequence of the construction of finite length, which gives a lower bound of $c-\varepsilon$.)

We let $X = f(k) \cdot \binom{U}{3k-5}$. In this way, we create at least $f(k)$ disjoint subintervals of the real line (each of them can be obtained from unshrinking a point of interest), where each contains a set of zero sized intervals such that at least one interval of each range is colored with colors $\{1, 2, \ldots, 3k-5\}$. In the second phase, we will introduce disjoint intervals which contain these subintervals of the real line obtained in the first phase. That is, in the conflict graph, a vertex of an interval of the second phase would be adjacent to a set of vertices such that, for each color in $1, 2, \ldots, 3k-5$, at least one neighbor of the new vertex has this color. Therefore, all colors used in the second phase would be of a larger index.

In the second phase, we consider the lower bound instance of the bin packing problem where OPT is known to the algorithm. If the lower bound construction asks to present an item of size s_i, we present an interval with size s_i that overlaps exactly one subinterval of the real line defined by the first phase (and therefore it cannot be colored with a color from $\{1, 2, \ldots, 3k-5\}$), and it does not intersect any preceding interval of the second phase. In this way, all intervals of the second phase are colored with colors greater than $3k-5$, and since they cannot be packed by the online algorithm using less than $(c-\varepsilon)k$ bins, they use colors $3k-4, \ldots, (3+c-\varepsilon)k-5$ (this is without loss of generality after renaming the colors).

To prove the claim, it suffices to show that $\mathrm{OPT} = k$. To see this, note that each of the first phase construction can be colored using $k-1$ colors (as the maximum clique size in it is $k-1$). Therefore, we consider the optimal solution for the bin packing instance that uses k bins. Then we traverse the first phase constructions one by one and allocate the intervals in the first phase constructions to $k-1$ colors among the existing k colors so that the overlapping interval of the second phase (if there is such one) has a different color. In this way, the total size of items that are allocated to a color is at most one, and we obtain a coloring using k colors that satisfies the conflicts constraints such that the total size of each color class is at most one. ∎

LEMMA 18. *Any online algorithm for standard bin packing, which knows the value* OPT *in advance, has a competitive ratio of at least* $\frac{47}{36} \approx 1.30556$.

Proof. We use a construction similar to the lower bound given by Yao in [33] (see also [32]). The difference is that since we commit on a given value of OPT in advance, we need to pad the sequence with items of size 1 in cases where we would otherwise simply stop the sequence.

Let N be a large integer which is divisible by 6. The input consists of one of the following three inputs:

1. N items of size 0.15, followed by $\frac{5}{6}N$ items of size 1.
2. N items of size 0.15, followed by N items of size 0.34, followed by $\frac{N}{2}$ items of size 1.
3. N items of size 0.15, followed by N items of size 0.34, followed by N items of size 0.51.

It is not difficult to verify that, in all three cases, $\mathrm{OPT} = N$. We use the following variables. For $i = 1, \ldots, 6$, X_i denotes the number of bins with exactly i items of size 0.15, after only these items have arrived. For $i = 0, \ldots, 6$, $0 \le j \le 2$, $i + j > 0$, $X_{i,j}$ denotes the number of bins with exactly i items of size 0.15 and j items of size 0.34, after these two sets of items have arrived. Clearly, if $i \ge 3$, then $X_{i,2} = 0$ and if $i \ge 5$, then $X_{i,j} = 0$ for $j \ne 0$. For convenience, we also let $X_{0,0} = 0$. We define $X_0 = X_{0,1} + X_{0,2}$ to be the number of bins with only (one or two) items of size 0.34. Moreover, we have for $1 \le i \le 6$, $X_i = X_{i,0} + X_{i,1} + X_{i,2}$.

The following equalities must hold due to the amounts of items: $\sum_{i=1}^{6} iX_i = N$ and $\sum_{i=0}^{6} \sum_{j=0}^{2} jX_{i,j} = N$.

Let $\mathcal{R}$ be the competitive ratio of an algorithm. We can compute the cost of the algorithm for each of the three inputs. This cost is at most $\mathcal{R} \cdot N$. The costs are $\sum_{i=1}^{6} X_i + \frac{5}{6}N$, $\sum_{i=0}^{6} X_i + \frac{N}{2}$, since, in these cases, the algorithm must put the large items into new bins, and $X_{0,2} + X_{1,2} + X_{2,1} + X_{2,2} + X_{3,1} + X_4 + X_5 + X_6 + N$. This is true, since the following bins can accommodate an item of size 0.51: bins with no items of size 0.34 and at most three items of size 0.15, bins with one item of each of these sizes, and bins with only one item which is of size 0.34.

We have three inequalities, which we multiply by the coefficients $1, 2, 3$, respectively, and get the following:

$$\begin{aligned} &2X_{0,1} + 5X_{0,2} + 3X_{1,0} + 3X_{1,1} + 6X_{1,2} + 3X_{2,0} + 6X_{2,1} + 6X_{2,2} \\ &+ 3X_{3,0} + 6X_{3,1} + 6X_{4,0} + 6X_{4,1} + 6X_{5,0} + 6X_{6,0} + \frac{29}{6}N \le 6\mathcal{R}N. \end{aligned}$$

We have established the following two equalities: $\sum_{i=1}^{6} iX_i = N$ and $\sum_{i=0}^{6} \sum_{j=0}^{2} jX_{i,j} = N$, which we multiply by the coefficients $1, 2$, respectively. Hence, we get the following:

$$\begin{aligned} &2X_{0,1} + 4X_{0,2} + X_{1,0} + 3X_{1,1} + 5X_{1,2} + 2X_{2,0} + 4X_{2,1} + 6X_{2,2} \\ &+ 3X_{3,0} + 5X_{3,1} + 4X_{4,0} + 6X_{4,1} + 5X_{5,0} + 6X_{6,0} = 3N. \end{aligned}$$

Since all variables are nonnegative, we substitute and get $\frac{29}{6}N + 3N \le 6\mathcal{R}N$, and thus $\mathcal{R} \ge \frac{47}{36}$. ∎

5. Conclusion. We have improved the upper bounds for BPC on perfect graphs, interval graphs (and a few related classes), and bipartite graphs. Most of our results follow from the adaptation of weighting systems to enable an analysis of algorithms for BPC, and new algorithms which carefully remove small subgraphs of items which cause problematic instances. There is still a gap between the inapproximability which

follows from bin packing and the upper bounds. An open problem would be to close this gap.

Another open question is the following. As in [17], we used the absolute approximation ratio to analyze the performance of our algorithms. It can be seen that, using the asymptotic approximation ratio, we can achieve a slightly better upper bound for bipartite graphs. It is unclear whether the same is true for other graph classes, i.e., whether the asymptotic approximation ratio for BPC is strictly lower than the absolute one for some cases.

REFERENCES

[1] E. Arkin and R. Hassin, *On local search for weighted packing problems*, Math. Oper. Res., 23 (1998), pp. 640–648.

[2] B. S. Baker and E. G. Coffman, Jr., *A tight asymptotic bound for Next-Fit-Decreasing bin-packing*, SIAM J. Algebr. Discrete Methods, 2 (1981), pp. 147–152.

[3] A. Borodin and R. El-Yaniv, *Online Computation and Competitive Analysis*, Cambridge University Press, London, 1998.

[4] P. Crescenzi, V. Kann, M. M. Halldórsson, M. Karpinski, and G. J. Woeginger, *A Compendium of NP Optimization Problems*, http://www.nada.kth.se/~viggo/problemlist/compendium.html.

[5] J. Csirik and J. Y.-T. Leung, *Variants of classical one-dimensional bin packing*, in Handbook of Approximation Algorithms and Metaheuristics, T. F. Gonzalez, ed., Chapman & Hall/CRC, 2007, pp. (33-1)–(33-13).

[6] D. de Werra, *An introduction to timetabling*, European J. Oper. Res., 19 (1985), pp. 151–162.

[7] G. Even and S. Shahar, *Scheduling of a smart antenna: Capacitated coloring of unit circular-arc graphs*, in Proceedings of the Third Workshop on Combinatorial and Algorithmic Aspects of Networking (CAAN2006), 2006, pp. 58–71.

[8] G. Galambos and G. J. Woeginger, *Repacking helps in bounded space online bin packing*, Computing, 49 (1993), pp. 329–338.

[9] M. R. Garey, R. L. Graham, D. S. Johnson, and A. C. C. Yao, *Resource constrained scheduling as generalized bin packing*, J. Combin. Theory Ser. A, 21 (1976), pp. 257–298.

[10] M. R. Garey and D. S. Johnson, *Computers and Intractability*, W. H. Freeman and Company, New York, 1979.

[11] M. C. Golumbic, *Algorithmic Graph Theory and Perfect Graphs*, Academic Press, New York, 1980.

[12] R. L. Graham, *Bounds on multiprocessing timing anomalies*, SIAM J. Appl. Math., 17 (1969), pp. 263–269.

[13] A. Gyárfás and J. Lehel, *On-line and first fit colorings of graphs*, J. Graph Theory, 12 (1988), pp. 217–227.

[14] M. Hujter and Z. Tuza, *Precoloring extension,* III: *Classes of perfect graphs*, Combin. Probab. Comput., 5 (1996), pp. 35–56.

[15] S. Irani and V. J. Leung, *Scheduling with conflicts on bipartite and interval graphs*, J. Sched., 6 (2003), pp. 287–307.

[16] K. Jansen, *An approximation scheme for bin packing with conflicts*, J. Comb. Optim., 3 (1999), pp. 363–377.

[17] K. Jansen and S. Öhring, *Approximation algorithms for time constrained scheduling*, Inform. and Comput., 132 (1997), pp. 85–108.

[18] T. R. Jensen and B. Toft, *Graph Coloring Problems*, Wiley, New York, 1995.

[19] D. S. Johnson, A. Demers, J. D. Ullman, M. R. Garey, and R. L. Graham, *Worst-case performance bounds for simple one-dimensional packing algorithms*, SIAM J. Comput., 3 (1974), pp. 299–325.

[20] H. A. Kierstead and W. T. Trotter, *An extremal problem in recursive combinatorics*, Congr. Numer., 33 (1981), pp. 143–153.

[21] C. C. Lee and D. T. Lee, *A simple online bin packing algorithm*, J. ACM, 32 (1985), pp. 562–572.

[22] L. Lovász, M. Saks, and W. T. Trotter, *An on-line graph coloring algorithm with sublinear performance ratio*, Discrete Math., 75 (1989), pp. 319–325.

[23] D. Marx, *Precoloring Extension*, http://www.cs.bme.hu/~dmarx/prext.html.

[24] D. Marx, *Precoloring Extension on Chordal Graphs*, manuscript, 2004.

[25] B. McCloskey and A. Shankar, *Approaches to Bin Packing with Clique-Graph Conflicts*, Technical report UCB/CSD-05-1378, EECS Department, University of California, Berkeley, 2005.
[26] Y. Oh and S. H. Son, *On a Constrained Bin-packing Problem*, Technical report CS-95-14, Department of Computer Science, University of Virginia, Charlottesville, VA, 1995.
[27] A. Schrijver, *Combinatorial Optimization Polyhedra and Efficiency*, Springer-Verlag, Berlin, 2003.
[28] S. S. Seiden, *On the online bin packing problem*, J. ACM, 49 (2002), pp. 640–671.
[29] D. Simchi-Levi, *New worst-case results for the bin-packing problem*, Naval Res. Logist., 41 (1994), pp. 579–585.
[30] N. J. A. Sloane, *On-line Encyclopedia of Integer Sequences*, 1996–2005, http://www.research.att.com/∼njas/sequences/Seis.html.
[31] J. D. Ullman, *The performance of a Memory Allocation Algorithm*, Technical report 100, Princeton University, Princeton, NJ, 1971.
[32] A. van Vliet, *An improved lower bound for online bin packing algorithms*, Inform. Process. Lett., 43 (1992), pp. 277–284.
[33] A. C. C. Yao, *New algorithms for bin packing*, J. ACM, 27 (1980), pp. 207–227.

SIAM J. OPTIM.
Vol. 19, No. 3, pp. 1299–1325

DYNAMIC CONTROL OF INFEASIBILITY IN EQUALITY CONSTRAINED OPTIMIZATION*

ROBERTO H. BIELSCHOWSKY† AND FRANCISCO A. M. GOMES‡

Abstract. This paper describes a new algorithm for solving nonlinear programming problems with equality constraints. The method introduces the idea of using trust cylinders to keep the infeasibility under control. Each time the trust cylinder is violated, a restoration step is called and the infeasibility level is reduced. The radius of the trust cylinder has a nonincreasing update scheme, so eventually a feasible (and optimal) point is obtained. Global and local convergence of the algorithm are analyzed, as well as its numerical performance. The results suggest that the algorithm is promising.

Key words. nonlinear programming, constrained optimization, large-scale optimization

AMS subject classifications. 65K05, 90C30

DOI. 10.1137/070679557

1. Introduction.

We consider the equality constrained optimization problem

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && h(x) = 0, \end{aligned} \tag{1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}^m$ are C^2 functions.

Usual algorithms for solving problem (1.1) alternate normal (or "vertical") steps towards the feasible set $\mathcal{H}_0 = \{x : h(x) = 0\}$ with tangential (or "horizontal") steps towards the dual manifold $\nabla\mathcal{L} = \{x : \nabla L(x, \lambda) = 0\}$, where L is the Lagrangian function. Generally, these steps are obtained from some quadratic model for (1.1). This feature is shared, for example, by the trust region methods proposed by Biegler, Nocedal, and Schmid [4], Byrd, Gilbert, and Nocedal [9], Byrd, Hribar, and Nocedal [10], Dennis and Vicente [15], El-Alem [17], Gomes, Maciel, and Martínez [22], and Lalee, Nocedal, and Plantenga [23].

In this paper, we propose an algorithm that uses normal and tangential trust region models in a more flexible way. Our bet is that, rather than taking one normal and one tangential step per iteration, we might do better if, at some iterations, $\nabla\mathcal{L}$ is pursued with some priority, so several successive horizontal steps are taken before one vertical step is computed. On the other hand, we believe that, in some cases, it is preferable to move closer and closer to $\mathcal{H}_0$, so we systematically force the vertical step. To allow this to occur, we introduce a single mechanism based on what we call *trust cylinders*.

Another feature that distinguishes our method from most of the nonlinear optimization algorithms recently proposed is that, using trust cylinders, we need not rely on a filter (see, for example, [13, Chap. 15]) or on a merit function (see [22]) to obtain global convergence. Instead, we accept the horizontal step if it sufficiently decreases

*Received by the editors January 8, 2007; accepted for publication (in revised form) July 2, 2008; published electronically November 19, 2008.
http://www.siam.org/journals/siopt/19-3/67955.html
†Departamento de Matemática, Universidade Federal do Rio Grande do Norte, Natal, RN, Brazil (rhbiel@ccet.ufrn.br).
‡Departamento de Matemática Aplicada, IMECC, Universidade Estadual de Campinas, CP 6065, 13081-970, Campinas, SP, Brazil (chico@ime.unicamp.br).

the Lagrangian function, subject to the condition of staying (dynamically) close to feasibility, in a sense that will be explained in what follows.

Algorithms that generate feasible iterates, without solving $h(x) = 0$ explicitly, go back to the early 1960s, with methods usually classified either as generalized reduced gradient (see [44, 1, 2]) or as projected gradient (PG) [37, 38]. Variations of the PG method, including some strategies to relax feasibility in a controlled way, began to appear at the end of the '60s with the suggestive denomination of sequential gradient-restoration algorithm [29, 30]. See also [32, 35, 36]. More recently, Martínez introduced a new class of algorithms, called *inexact restoration methods* [24, 25, 26, 27, 28], that also controls infeasibility at each iteration.

Our approach has the flavor of a PG algorithm and could be characterized as a relaxed feasible point method, with a *dynamic control of infeasibility* (DCI). We look for a compromise between allowing a large enough horizontal step, in a direction approximately tangent to the restrictions $h(x) = 0$, and keeping infeasibility under control. The main idea is to force each iterate x^k to remain in a trust cylinder defined by

$$\mathcal{C}^k = \{x \in \mathbb{R}^n : \|h(x)\| \leq \rho^k\},$$

where $\|.\|$ denotes the ℓ_2 norm.

The dynamic control of infeasibility is kept defining the "radii" ρ^k of the trust cylinders in such a way that

$$\rho^k = O(\|g_p(x^k)\|), \tag{1.2}$$

where $g_p(x)$ stands for the projected gradient, i.e., the orthogonal projection of the gradient $g(x) = \nabla f(x)$ onto the null space of $\nabla h(x)$, the Jacobian of h. In our case of interest, $g_p(x)$ will be calculated at regular points of h, where $\nabla h(x)$ has full rank. In this situation, the least squares multiplier estimates, $\lambda_{LS}(x)$, are given by

$$\lambda_{LS}(x) = \operatorname{argmin}\{\|\nabla h(x)^T\lambda + g(x)\|\} = -(\nabla h(x)\nabla h(x)^T)^{-1}\nabla h(x)g(x), \tag{1.3}$$

and the resulting projected gradient is

$$g_p(x) = g(x) + \nabla h(x)^T\lambda_{LS}(x). \tag{1.4}$$

Given x^{k-1}, the kth iteration begins with a *restoration step*, if necessary, in order to obtain a point $x_c = x_c^k$ and a radius $\rho = \rho^k$ such that

$$\|h(x_c)\| \leq \rho = O(\|g_p(x_c)\|) \tag{1.5}$$

and

$$\|x_c - x^{k-1}\| = O(\|h(x^{k-1})\|). \tag{1.6}$$

A radius $\rho = \rho^k$ satisfying (1.2) may be defined as $\rho = \nu\, n_p(x_c)\, \rho_{max}$, where

$$n_p(x_c) = \frac{\|g_p(x_c)\|}{\|g(x_c)\| + 1} \tag{1.7}$$

and $10^{-4} \leq \nu \leq 1$ and $\rho_{max} > 0$ are constants.

Given x_c in $\mathcal{C}^k$, the second part of the kth iteration looks for a *horizontal step*, δ_t, that provides a sufficient decrease for a quadratic approximation of f and guarantees

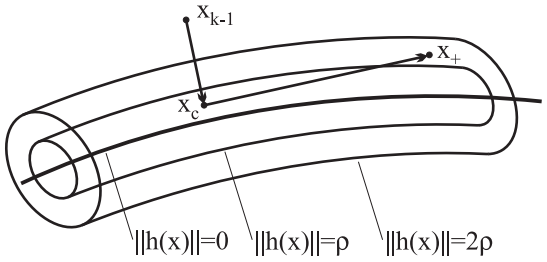


FIG. 1.1. *The step and the trust cylinders.* x_c *satisfies* $\|h(x_c)\| < \rho$, *while* x_+ *satisfies* $\|h(x_+)\| < 2\rho$.

that $x_+ = x_c + \delta_t$ remains in a bigger trust cylinder of radius 2ρ. An optional second order correction δ_{soc} may also be used to reduce the infeasibility, so $x^k = x_c + \delta_t + \delta_{soc}$.

Figure 1.1 sketches the vertical and the horizontal steps of a typical iteration.

An advantage of staying close to the feasible set is that a "good horizontal step" in a level set given by $h(x) = c$ is likely to be close to a "good horizontal step" in the feasible set given by $h(x) = 0$ if c is relatively small.

The parameter $\rho_{max} = \rho^k_{max}$ is nonincreasing and is responsible for the trustability of the trust cylinders. It is decreased every time there is evidence that the reduction of the Lagrangian function obtained in the horizontal step was menaced by a significant increase in the restoration step.

In the next section, we formalize the DCI algorithm. In section 3, a global convergence result for the algorithm is presented, followed by the local convergence theory introduced in section 4. Section 5 contains some preliminary numerical results. Finally, some conclusions and lines for future work are included in section 6.

2. The DCI algorithm. In this section, we depict a typical iteration of our main algorithm. As usual, we use the Lagrangian function, defined as

$$L(x, \lambda) = f(x) + \lambda^T h(x),$$

to evaluate the algorithm behavior. In fact, the control of the trust cylinder radius is also based on the variation of the Lagrangian at x^k_c, given by

$$\Delta L^k_c = L^k_c - L^{k-1}_c,$$

where $L^k_c = L(x^k_c, \lambda^k)$. Since our algorithm divides the step into two components, one vertical and one horizontal, this variation is also split according to

$$\Delta L^k_c = \Delta L^{k-1}_H + \Delta L^k_V, \tag{2.1}$$

where

$$\Delta L^{k-1}_H = L(x^{k-1}, \lambda^{k-1}) - L(x^{k-1}_c, \lambda^{k-1}), \tag{2.2}$$

$$\Delta L^k_V = L(x^k_c, \lambda^k) - L(x^{k-1}, \lambda^{k-1}). \tag{2.3}$$

In the vertical step of the algorithm, we seek a point x_c that satisfies (1.5) and (1.6). Under usual regularity assumptions for $\nabla h(x)$ along the iterations, we can achieve (1.5) and (1.6) with almost every algorithm available for the least squares problem

$$\text{minimize} \quad \|h(x)\|^2. \tag{2.4}$$

Careful line searches in the Cauchy, Gauss–Newton, or Newton directions, or in some combination of them, can be used, for example, to solve problem (2.4). As a matter of fact, algorithms for (2.4) that take steps in the form $d = -M\nabla h(x)^T h(x)$, where M represents a family of uniformly bounded and positive definite matrices, will produce a "sufficiently" fast convergence to a feasible point, and that is what we need to guarantee (1.6). In the implementation of the algorithm, we will give preference to a trust region method. Namely, our restoration step successively solves the linearized least squares problem

$$\begin{aligned} &\text{minimize} && \|h(x) + Ad\|^2 \\ &\text{subject to} && \|d\|_\infty \le \Delta_{VS}, \end{aligned} \tag{2.5}$$

where A is an approximation of $\nabla h(x)$ and $\Delta_{VS} > 0$ is a trust region radius adequately updated in the vertical subproblems.

In the horizontal step we solve the quadratic programming problem

$$\begin{aligned} &\text{minimize} && q(\delta) = g(x_c)^T\delta + \frac{1}{2}\delta^T B\delta \\ &\text{subject to} && \nabla h(x_c)\delta = 0, \\ & && \|\delta\|_\infty \le \Delta, \end{aligned} \tag{2.6}$$

where B is a symmetric approximation of the Hessian of the Lagrangian and $\Delta > 0$ is the trust region radius.

We suppose that, at the beginning of the kth iteration, the previous approximate solution, x^{k-1}, and the Lagrange multipliers estimate, λ^{k-1}, are available. In addition, we also suppose that the following are known: the upper limit for the trust cylinder radius, ρ_{max}; the Lagrangian function at some previous iteration j, $L_{ref} = L(x_c^j, \lambda^j)$; the horizontal variation of the Lagrangian, ΔL_H^{k-1}; and the trust region radii, $\Delta_{VS} \ge \Delta_{min}$ and $\Delta \ge \Delta_{min}$.

ALGORITHM 2.1. *The kth iteration of the DCI method.*

1. *Vertical step:*
1.1. $x_c = x^{k-1}$.
1.2. Choose an approximate value for ρ.
1.3. REPEAT
1.3.1. IF $\|h(x_c)\| > \rho$
1.3.1.1. Find x_c, such that $\|h(x_c)\| \le \rho$.
1.3.2. $g_p \leftarrow g_p(x_c)$; $n_p \leftarrow \|g_p(x_c)\|/(\|g(x_c)\| + 1)$.
1.3.3. Choose $\rho \in [10^{-4} n_p \rho_{max}, n_p \rho_{max}]$.
1.4. UNTIL $\|h(x_c)\| \le \rho$.
1.5. Compute λ_+.
2. *Convergence test:*
2.1. IF $(n_p = 0)$,
2.1.1. QUIT (x_c is a stationary point).
3. *ρ_{max} update:*
3.1. $\Delta L_V^k \leftarrow L(x_c, \lambda_+) - L(x^{k-1}, \lambda^{k-1})$.
3.2. IF $\Delta L_V^k \ge \frac{1}{2}[L_{ref} - L(x^{k-1}, \lambda^{k-1})]$,
3.2.1. $\rho_{max} \leftarrow \rho_{max}/2$.
3.3. IF $\Delta L_V^k > -\frac{1}{2}\Delta L_H^{k-1}$,
3.3.1. $L_{ref} \leftarrow L(x_c, \lambda_+)$.

4. *Horizontal step:*
4.1. REPEAT
4.1.1. Compute the Cauchy step δ_{CP}, solution of
$$\text{minimize} \quad q(\mu g_p)$$
$$\text{subject to} \quad \|\mu g_p\| \le \Delta,\ \mu \in [0,\infty).$$
4.1.2. Compute a trial step δ_t such that
$q(\delta_t) \le q(\delta_{CP})$,
$\|\delta_t\| \le \Delta$, and
$\nabla h(x_c)\,\delta_t = 0$.
4.1.3. Optionally, compute a second order correction δ_{soc}.
4.1.4. $\delta_+ \leftarrow \delta_t + \delta_{soc}$; $x_+ \leftarrow x_c + \delta_+$.
4.1.5. $\Delta L_H^k \leftarrow L(x_+,\lambda_+) - L(x_c,\lambda_+)$; $r \leftarrow \Delta L_H^k / q(\delta_t)$.
4.1.6. IF $(\|h(x_+)\| > 2\rho)$ OR $(r < \eta_1)$,
4.1.6.1. $\Delta \leftarrow \alpha_R \Delta$;
4.1.7. ELSE IF $r > \eta_2$,
4.1.7.1. $\Delta \leftarrow \alpha_I \Delta$.
4.2. UNTIL $(\|h(x_+)\| \le 2\rho)$ AND $(r \ge \eta_1)$.
5. *Approximate solution update:*
5.1. $x^k \leftarrow x_+$; $\lambda^k \leftarrow \lambda_+$; $k \leftarrow k+1$.
5.2. Choose $\Delta \ge \Delta_{min}$.

In Algorithm 2.1, we suppose that the restoration step 1.3.1.1 will always succeed. Obviously, this may not occur, since problem (1.1) may be infeasible. Therefore, some termination criterion needs to be defined to prevent the algorithm from getting stuck on this step.

In step 1.3.2 of the algorithm and in the next section, we assume that n_p is computed according to (1.7). As a matter of fact, all we need for our convergence theory is that $n_p = O(\|g_p(x_c)\|)$, but we decided to use an explicit formula for n_p in order to keep the text more readable. Another choice we made was to define λ_+ as the vector of least squares multipliers (1.3) computed at x_c^k, although other update schemes would work as well.

Most of the constants used in Algorithm 2.1 are explicitly shown above, so the reader does not need to guess the meaning of several obscure Greek letters. We do prefer to write $\|h(x_+)\| > 2\rho$ instead of $\|h(x_+)\| > \zeta\rho$, for example, to make clear that, in steps 4.1.6 and 4.2, we are considering a larger trust cylinder. Naturally, the algorithm will also work if we use $\zeta = 3$, although this modification will slightly affect the proofs of some lemmas presented in the next section. Only Δ_{min}[1] and the four constants that control the behavior of the trust region method used to compute the horizontal step were not specified. These parameters must satisfy $0 < \eta_1 \le 1/2$, $\eta_1 \le \eta_2 < 1$, $0 < \alpha_R < 1$, $\alpha_I \ge 1$, and $\Delta_{min} > 0$. Possible values are $\eta_1 = 10^{-3}$, $\eta_2 = 0.7$, $\alpha_R = 0.25$, $\alpha_I = 2.5$, and $\Delta_{min} = 10^{-5}$.

The following relations, easily derived from steps 1.3 and 4.2 of Algorithm 2.1, will be used frequently in the next two sections and are presented here for convenience:

$$\rho^k \le \rho_{max}^k \|g_p(x_c^k)\|, \tag{2.7}$$

[1]The parameter Δ_{min} plays no role in the global convergence theory of Algorithm 2.1. It is also unnecessary for the local convergence theory if we use the true Lagrangian second order polynomial as the quadratic model $q(\delta)$ or if B is a good approximation of $\nabla^2_{xx}L(x,\lambda)$ in the plane that contains the Cauchy and the quasi-Newton directions, as will become clear in section 4. However, our numerical tests indicate that keeping $\Delta \ge 10^{-5}$ may slightly improve the performance of the algorithm, so we decided to include step 5.2.

$$\rho_{max}^k \leq 10^4 \rho^k \frac{1}{n_p} = 10^4 \rho^k \frac{\|g(x_c^k)\| + 1}{\|g_p(x_c^k)\|}, \tag{2.8}$$

$$\|h(x_c^k)\| \leq \|h(x^{k-1})\| \leq 2\rho^{k-1}. \tag{2.9}$$

The global convergence of DCI will be guaranteed, under reasonable assumptions, by a typical *sufficient decrease* argument for the Lagrangian function evaluated at x_c^k. The variation of the Lagrangian between two successive iterations is given by (2.1). The idea is to prevent the decrease of the Lagrangian obtained at the horizontal step from being destroyed by the restoration. For that, ρ_{max}^k is decreased in step 3 of DCI every time ΔL_V^k is larger than a fraction of the difference between the Lagrangian at the current iteration and a reference value L_{ref} fixed in some previous iteration j. If the increase in ΔL_V^k significantly menaces the decrease in the Lagrangian obtained since iteration j, ρ_{max} is divided by 2 and L_{ref} is updated. L_{ref} is also updated every time $\Delta L_V^k > -\frac{1}{2}\Delta L_H^{k-1}$. The main argument to guarantee global convergence establishes, under suitable assumptions, the existence of *enough normal space*, dynamically calibrated for horizontal steps of reasonable size, in the sense that ρ_{max}^k remains bounded away from zero unless $\liminf(\|g_p(x_c^k)\|) = 0$.

3. Global convergence. The global convergence analysis of the DCI algorithm is based on the following hypotheses.

H1 (differentiability): f and h are C^2.

H2 (compacity): The generated sequences $\{{x_c}^k\}$ and $\{x^k\}$, the Hessian approximations B^k, and the multipliers $\{\lambda^k\}$ remain uniformly bounded.

H3 (regularity and restoration): The restoration never fails, and $Z = \{{x_c}^k\}$ remains far from the singular set of h, in the sense that h is regular in the closure of Z. Equivalently, $\{\|\nabla h({x_c}^k)^T \nabla h({x_c}^k)^{-1}\|\}$ remains uniformly bounded. Also, if the generated sequence $\{{x_c}^k\}$ is infinite, it satisfies

$$\|{x_c}^{k+1} - x^k\| = O(\|h(x^k)\|). \tag{3.1}$$

H4 (second order correction): $\|{\delta_{soc}}^k\| = O(\|{\delta_t}^k\|^2)$.

Supposing that H1 holds, we can assure that the remaining hypotheses will hold if, for example, the feasible set $\mathcal{H}_0$ is compact and regular (i.e., $\nabla h(x)$ is of maximal rank on $\mathcal{H}_0$) and x^0 is feasible. In this case, if we choose an initial ρ_{max}^0 sufficiently small, we can keep $\nabla h(x)$ with maximal rank and assure (3.1) using standard algorithms for restoration, such as the Gauss–Newton method. We could also replace the compacity property of $\mathcal{H}_0$ by adequate properties on f, such as requiring f to satisfy $\lim_{\|x\|\to\infty} f(x) = \infty$. In such situations, H2–H4 can be guaranteed by construction.

From now on we assume that the sequences $\{x_c^k\}$ and $\{x^k\}$ generated by DCI satisfy H1–H4. In addition, when we say that a number is a constant, we mean that it can be used for all k and is associated with these specific sequences generated by DCI.

Denoting by δ_V^k the vertical step and by δ_H^k the horizontal step in the kth iteration, we have

$$\delta_V^k = {x_c}^k - x^{k-1} \quad \text{and} \quad \delta_H^k = x^k - {x_c}^k = {\delta_t}^k + \delta_{soc}^k.$$

Hypotheses H1–H4 allow us to choose a constant $\delta_{max} > 0$ such that, for all k,

$$\|\delta_t^k\| + \|\delta_{soc}^k\| + \|\delta_V^k\| \leq \delta_{max}. \tag{3.2}$$

We can also define a second positive constant ξ_0 such that, for all k, if $\|x-x_c^k\| \le \delta_{max}$, then

$$\|\nabla h_j(x)\| \le \xi_0, \quad j = 1, \ldots, m, \tag{3.3}$$

$$\|\nabla^2 h_j(x)\| \le \xi_0, \quad j = 1, \ldots, m, \tag{3.4}$$

$$\|\nabla f(x)\| \le \xi_0, \tag{3.5}$$

$$\|\nabla^2 f(x)\| \le \xi_0, \tag{3.6}$$

$$\|\nabla h(x_c^k)^T \nabla h(x_c^k)\| \le \xi_0, \tag{3.7}$$

$$\|B^k\| \le \xi_0, \tag{3.8}$$

$$\|\lambda^k\| \le \xi_0, \tag{3.9}$$

$$\|\delta_{soc}^k\| \le \xi_0 \|\delta_t^k\|^2. \tag{3.10}$$

To simplify our notation, we suppose that the constant ξ_0 is large enough so that (3.1) can be rewritten as

$$\|{x_c}^{k+1} - x^k\| \le \xi_0 \|h(x^k)\|. \tag{3.11}$$

The main result of this section, presented in Theorem 3.6, is based on five lemmas. The first gives an upper limit for the increase in the infeasibility produced by the horizontal step.

LEMMA 3.1. *The trial iterate x_+ generated in step* 4.1.4 *of Algorithm* 2.1 *satisfies*

$$\|h(x_+) - h({x_c}^k)\| \le \overline{\xi_0} \|\delta_t\|^2. \tag{3.12}$$

Proof. Since $x_+ = {x_c}^k + \delta_+ = {x_c}^k + \delta_t + \delta_{soc}$, with $\|\delta_+\| \le \delta_{max}$, we can use a Taylor expansion, together with relations (3.4), (3.3), (3.10), and (3.2) and the fact that $\nabla h({x_c}^k){\delta_t}^k = 0$, to show that, for every $j = 1, \ldots, m$,

$$\begin{aligned} |h_j(x_+) - h_j(x_c)| &\le |\nabla h_j(x_c)^T(\delta_t + \delta_{soc})| + \frac{\xi_0}{2}\|\delta_t + \delta_{soc}\|^2 \\ &= |\nabla h_j(x_c)^T \delta_{soc}| + \frac{\xi_0}{2}\|\delta_t + \delta_{soc}\|^2 \\ &\le \xi_0^2 \|\delta_t\|^2 + \xi_0(\|\delta_t\|^2 + \|\delta_{soc}\|^2) \\ &\le (\xi_0^2 + \xi_0 + \xi_0^2 \delta_{max})\|\delta_t\|^2. \end{aligned}$$

Setting $\overline{\xi}_0 = \sqrt{m}(\xi_0^2 + \xi_0 + \xi_0^2 \delta_{max})$, we get the desired result. ∎

The second lemma establishes that, under H1–H4, each iteration succeeds and the Lagrangian is sufficiently decreased.

LEMMA 3.2. *If x_c^k is not a stationary point for* (1.1)*, then x_+ is eventually accepted in step* 4 *of DCI. Moreover, we can define positive constants ξ_1, ξ_2, and ξ_3 such that, for all k,*

$$\begin{aligned} \Delta L_H^k &= L(x^k, \lambda^k) - L(x_c^k, \lambda^k) \\ &\le -\xi_1 \|g_p(x_c^k)\| \min\left\{\xi_2 \|g_p(x_c^k)\|,\, \xi_3 \sqrt{\rho^k}\right\}. \end{aligned} \tag{3.13}$$

Proof. To simplify the notation we will omit here the superscript k. Suppose that x_c is not stationary for (1.1). Let $x_+ = x_c + \delta_+ = x_c + \delta_t + \delta_{soc}$ be a candidate obtained in step 4 of the kth iteration of the DCI algorithm, and let λ_+ be the corresponding multiplier.

Using a Taylor expansion, Lemma 3.1, and relations (3.6), (3.9), (3.5), (3.2), (3.10), and (3.8), we obtain

$$\begin{aligned}\Delta L_H^+ &= L(x_+,\lambda_+) - L(x_c,\lambda_+) = f(x_+) - f(x_c) + \lambda_+^T(h(x_+) - h(x_c)) \\ &\le g(x_c)^T\delta_t + g(x_c)^T\delta_{soc} + \frac{\xi_0}{2}\|\delta_t + \delta_{soc}\|^2 + \xi_0\overline{\xi}_0\|\delta_t\|^2 \\ &\le q(\delta_t) - \frac{1}{2}\delta_t^T B\delta_t + \overline{\xi}_1\|\delta_t\|^2 \\ &\le q(\delta_t) + \overline{\xi}_2\|\delta_t\|^2, \qquad (3.14)\end{aligned}$$

where $\overline{\xi}_1 = \xi_0^2 + \xi_0 + \xi_0^2\delta_{max} + \xi_0\overline{\xi}_0$ and $\overline{\xi}_2 = \xi_0/2 + \overline{\xi}_1$.

Because δ_{CP}, defined in step 4.1.1 of DCI, is a Cauchy step tangent to the constraints, we have (see, for example, [13])

$$\|\delta_{CP}\| \ge \min\left\{\frac{\|g_p(x_c)\|}{\|B\|}, \Delta\right\} \ge \min\left\{\frac{\|g_p(x_c)\|}{\xi_0}, \Delta\right\}$$

and

$$q(\delta_{CP}) \le \frac{1}{2}g(x_c)^T\delta_{CP} \le -\frac{1}{2}\|g_p(x_c)\| \min\left\{\frac{\|g_p(x_c)\|}{\xi_0}, \Delta\right\}. \qquad (3.15)$$

To prove the first part of the lemma, we will show that x_+ is always accepted whenever $\Delta \le \overline{\Delta}$, where

$$\overline{\Delta} = \min\left\{\frac{\|g_p(x_c)\|}{4\overline{\xi}_2}, \sqrt{\frac{\rho}{\overline{\xi}_0}}\right\}. \qquad (3.16)$$

We start by noting that, since $\xi_0 < 4\overline{\xi}_2$,

$$\overline{\Delta} \le \frac{\|g_p(x_c)\|}{\xi_0}. \qquad (3.17)$$

Based on the fact that $\Delta \le \overline{\Delta}$ and on (3.17), the upper limit of $q(\delta_{CP})$ given by (3.15) can be simplified to

$$q(\delta_{CP}) \le -\frac{1}{2}\|g_p(x_c)\|\Delta. \qquad (3.18)$$

Combining the conditions $\|\delta_t\| \le \Delta$ and $q(\delta_t) \le q(\delta_{CP})$, stated in step 4.1.2 of the DCI algorithm, with (3.16) and (3.18), we obtain

$$\overline{\xi}_2\|\delta_t\|^2 \le \overline{\xi}_2\Delta^2 \le \overline{\xi}_2\overline{\Delta}\Delta \le \frac{1}{4}\|g_p(x_c)\|\Delta \le -\frac{1}{2}q(\delta_{CP}) \le -\frac{1}{2}q(\delta_t). \qquad (3.19)$$

Now, from (3.14) and (3.19), we get

$$\Delta L_H^+ \le \frac{1}{2}q(\delta_t) < 0, \qquad (3.20)$$

which implies that

$$r = \frac{\Delta L_H^+}{q(\delta_t)} \ge \frac{1}{2} \ge \eta_1. \qquad (3.21)$$

Since $\Delta \leq \overline{\Delta}$ and $\|h(x_c)\| \leq \rho$, we can use (3.12) and (3.16) to guarantee that

$$\|h(x_+)\| \leq \rho + \overline{\xi}_0 \|\delta_t\|^2 \leq \rho + \overline{\xi}_0 \overline{\Delta}^2 \leq 2\rho.$$

Therefore, both conditions stated in step 4.2 of the algorithm are satisfied, and x_+ is accepted.

To prove the second part of the lemma, let us recall that, each time the step is rejected, Δ is multiplied by α_R, which means that we can assume the accepted trust region radius satisfies $\Delta \geq \alpha_R \overline{\Delta}$, where $0 < \alpha_R < 1$. Combining this with (3.20), the condition $q(\delta_t) \leq q(\delta_{CP})$, (3.15), (3.17), and (3.16), we get

$$\begin{aligned} \Delta L_H^+ \leq \frac{1}{2} q(\delta_t) \leq \frac{1}{2} q(\delta_{CP}) &\leq -\frac{1}{4} \|g_p(x_c)\| \min\left\{ \frac{\|g_p(x_c)\|}{\xi_0}, \Delta \right\} \\ &\leq -\frac{1}{4} \|g_p(x_c)\| \alpha_R \overline{\Delta}. \end{aligned} \tag{3.22}$$

Defining $\xi_1 = \alpha_R/4$, $\xi_2 = 1/(4\overline{\xi}_2)$, and $\xi_3 = 1/\sqrt{\overline{\xi}_0}$, we obtain (3.13) from (3.16) and (3.22). ∎

Our third lemma defines an upper limit for the (possibly positive) vertical variation of the Lagrangian, ΔL_V^{k+1}.

LEMMA 3.3. *There exists a positive constant ξ_4 such that*

$$\Delta L_V^{k+1} \leq \xi_4 \rho_{max}^k \|g_p(x_c^k)\|. \tag{3.23}$$

Proof. Using a Taylor expansion, (3.5), (3.9), (3.11), (2.9), and (2.7), we get, for the vertical variation,

$$\begin{aligned} \Delta L_V^{k+1} &= L(x_c^{k+1}, \lambda^{k+1}) - L(x^k, \lambda^k) \\ &= f(x_c^{k+1}) - f(x^k) + {\lambda^{k+1}}^T h(x_c^{k+1}) - {\lambda^k}^T h(x^k) \\ &\leq \xi_0 \|x_c^{k+1} - x^k\| + \xi_0 \|h(x_c^{k+1})\| + \xi_0 \|h(x^k)\| \\ &\leq (\xi_0^2 + 2\xi_0) \|h(x^k)\| \leq 2(\xi_0^2 + 2\xi_0)\rho^k \leq \xi_4 \rho_{max}^k \|g_p(x_c^k)\|. \end{aligned}$$

Therefore, defining $\xi_4 = 2(\xi_0^2 + 2\xi_0)$, we obtain the desired result. ∎

Our fourth lemma establishes that, between successive iterations without changes in ρ_{max}, the Lagrangian decreases proportionally to the descent in the corresponding horizontal steps.

LEMMA 3.4. *If $\rho_{max}^{k+1} = \rho_{max}^{k+2} = \cdots = \rho_{max}^{k+j}$, for $j \geq 1$, then*

$$L_c^{k+j} - L_c^k = \sum_{i=k+1}^{k+j} \Delta L_c^i \leq \frac{1}{4} \sum_{i=k}^{k+j-1} \Delta L_H^i + r^k, \tag{3.24}$$

where $r^k = \frac{1}{2}[L_{ref}^k - L_c^k]$.

Proof. Let us suppose that L_{ref} does not change between iterations $k+1$ and $k + j_1 - 1$, where $0 < j_1 \leq j + 1$. In this case, by (2.1) and the criterion defined in step 3.3 of the algorithm, we have

$$L_c^{k+j_1-1} - L_c^k = \sum_{i=k+1}^{k+j_1-1} (\Delta L_V^i + \Delta L_H^{i-1}) \leq \frac{1}{2} \sum_{i=k}^{k+j_1-2} \Delta L_H^i. \tag{3.25}$$

On the other hand, if L_{ref} changes at iteration $k+j_1$, then the condition stated in step 3.3 of DCI is satisfied. In this case, using the hypothesis that ρ_{max} stays unchanged at iteration $k+j_1$ (so the inequality in step 3.2 is not satisfied) and the fact that $\Delta L_H^k \leq 0$, for all k, we have

$$\begin{aligned} L_c^{k+j_1} - L_c^k &= \Delta L_V^{k+j_1} + L(x^{k+j_1-1}, \lambda^{k+j_1-1}) - L_{ref}^k + [L_{ref}^k - L_c^k] \\ &\leq \frac{1}{2}(L(x^{k+j_1-1}, \lambda^{k+j_1-1}) - L_{ref}^k) + [L_{ref}^k - L_c^k] \\ &= \frac{1}{2}(\Delta L_H^{k+j_1-1} + L_c^{k+j_1-1} - L_c^k) + \frac{1}{2}[L_{ref}^k - L_c^k] \\ &\leq \frac{1}{4}\sum_{i=k}^{k+j_1-1} \Delta L_H^i + r^k. \end{aligned} \tag{3.26}$$

If $j_1 \geq j$, then (3.25) and (3.26) imply (3.24).

On the other hand, if L_{ref} is updated at iterations $k+j_1, \ldots, k+j_s$, where $0 < j_1 < j_2 < \cdots < j_s \leq j$, then $r^{k+j_1} = r^{k+j_2} = \cdots = r^{k+j_s} = 0$. Therefore, applying the same procedure described above several times and defining $j_0 = 0$, we obtain

$$L_c^{k+j} - L_c^k = \sum_{i=1}^{s}[L_c^{k+j_i} - L_c^{k+j_{i-1}}] + L_c^{k+j} - L_c^{k+j_s} \leq \frac{1}{4}\sum_{i=k}^{k+j-1} \Delta L_H^i + r^k. \qquad \square$$

Our fifth lemma establishes the existence of *enough normal space* in the trust cylinders $\mathcal{C}^k$ to guarantee that the Lagrangian can be sufficiently decreased. The idea supporting this lemma is that (3.13) guarantees, asymptotically, that $\|\Delta L_H^k\|$ is bigger than a fraction of $\sqrt{\rho^k}$, while, on the other hand, $\|\Delta L_V^k\| = O(\rho^k)$ (see the proof of Lemma 3.3). This means that a restoration cannot, asymptotically, destroy the decrease in the Lagrangian achieved at the horizontal step, and this prevents further ρ_{max} updates.

LEMMA 3.5. *If DCI generates an infinite sequence $\{x^k\}$, then the following hold:*

(i) *There are positive constants ξ_5 and ξ_6 such that, whenever*

$$\rho_{max}^k < \min\{\xi_5 \|g_p(x_c^k)\|, \xi_6\}, \tag{3.27}$$

ρ_{max} does not change at iteration $k+1$.

(ii) *Furthermore, if $\liminf \|g_p(x_c^k)\| > 0$, then there exists $k_0 > 0$ such that, for every $k \geq k_0$,*

$$\rho_{max}^k = \rho_{max}^{k_0}. \tag{3.28}$$

(iii) *If the horizontal step and the vector of Lagrange multipliers satisfy*

$$\|x^k - x_c^k\| = O(\|g_p(x_c^k)\|), \tag{3.29}$$

$$\|\lambda^k - \lambda_{LS}(x_c^k)\| = O(\|g_p(x_c^k)\|), \tag{3.30}$$

then (3.28) is satisfied, regardless of the value of $\liminf \|g_p(x_c^k)\|$. In other words, ρ_{max}^k remains bounded away from zero.

Proof. Let us consider the first part of the lemma. To prove that ρ_{max} does not change at iteration $k+1$, we just need to show that $\Delta L_V^{k+1} < -\Delta L_H^k/2$ (see step 3.2 of Algorithm 2.1). From Lemma 3.2, this result is attained whenever

$$\Delta L_V^k < \frac{\xi_1 \xi_2}{2}\|g_p(x_c^k)\|^2 \tag{3.31}$$

and

$$\Delta L_V^k < \frac{\xi_1\xi_3}{2}\sqrt{\rho^k}\|g_p(x_c^k)\|. \tag{3.32}$$

Condition (3.31) can be easily obtained from Lemma 3.3 and (3.27), taking $\xi_5 = \xi_5^a \equiv \xi_1\xi_2/(2\xi_4)$. To obtain (3.32), we need a few more steps. First we use (2.8) and (3.5) to write

$$\sqrt{\rho_{max}^k} \le 10^2\sqrt{\rho^k}\frac{(\xi_0+1)^{1/2}}{\|g_p(x_c^k)\|^{1/2}}. \tag{3.33}$$

Then, taking the square root from both sides of (3.27) and combining the result with (3.33), we get

$$\rho_{max} \le \sqrt{\xi_5}10^{-2}\sqrt{\xi_0+1}\sqrt{\rho^k}. \tag{3.34}$$

Now, defining $\xi_5 = \xi_5^b \equiv 10^{-4}\xi_1^2\xi_2^2/[4\xi_4^2(\xi_0+1)]$ and using Lemma 3.3 and (3.27), we obtain (3.32). The desired result follows from taking $\xi_5 = \min\{\xi_5^a, \xi_5^b\}$.

In order to prove item (ii), let us define $b = \liminf(\|g_p(x_c^k)\|)$ and choose an index $\overline{k}_0$ such that $\|g_p(x_c^k)\| > b/2$ for $k \ge \overline{k}_0$. Then, as we proved above, $\rho_{max}^k \ge \min\{\rho_{max}^{\overline{k}_0}, \xi_5 b/2, \xi_6\}$ for $k > k_0$. Thus, ρ_{max} will never be decreased after a certain iteration k_0, as claimed.

To prove the third part of the lemma, we begin observing that (1.3)–(1.4) and H1–H3 imply that $\lambda_{LS}(x)$ and $g_p(x)$ are well defined and of class C^1 in a compact neighborhood of $\overline{Z}$, the closure of $Z = \{x_c^k\}$. Therefore, $\lambda_{LS}(x)$ and $g_p(x)$ are Lipschitz continuous on the iterates in the sense that

$$\|\lambda_{LS}(x_c^{k+1}) - \lambda_{LS}(x_c^k)\| = O(\|x_c^{k+1} - x_c^k\|) \tag{3.35}$$

and

$$\|g_p(x_c^{k+1}) - g_p(x_c^k)\| = O(\|x_c^{k+1} - x_c^k\|). \tag{3.36}$$

From (3.1), (2.9), (2.7), and (3.29) we get

$$\|x_c^{k+1} - x_c^k\| \le \|x_c^{k+1} - x^k\| + \|x^k - x_c^k\| = O(\|g_p(x_c^k)\|), \tag{3.37}$$

and from (3.36) and (3.37) we obtain

$$\|g_p(x_c^{k+1})\| = O(\|g_p(x_c^k)\|). \tag{3.38}$$

Noticing that

$$L(x_c^{k+1}, \lambda^{k+1}) = L(x_c^{k+1}, \lambda_{LS}(x_c^{k+1})) + [\lambda^{k+1} - \lambda_{LS}(x_c^{k+1})]^T h(x_c^{k+1})$$

and

$$\begin{aligned} L(x^k, \lambda^k) = L(x^k, \lambda_{LS}(x_c^{k+1})) &- [\lambda_{LS}(x_c^{k+1}) - \lambda_{LS}(x_c^k)]^T h(x^k) \\ &-[\lambda_{LS}(x_c^k) - \lambda^k]^T h(x^k), \end{aligned}$$

we get the following decomposition of ΔL_V^{k+1} into a sum of four terms:

$$\begin{aligned}\Delta L_V^{k+1} &= L(x_c^{k+1},\lambda^{k+1}) - L(x^k,\lambda^k)\\ &= [L(x_c^{k+1},\lambda_{LS}(x_c^{k+1})) - L(x^k,\lambda_{LS}(x_c^{k+1}))]\\ &\quad +[\lambda^{k+1}-\lambda_{LS}(x_c^{k+1})]^T h(x_c^{k+1})\\ &\quad +[\lambda_{LS}(x_c^{k+1})-\lambda_{LS}(x_c^k)]^T h(x^k)\\ &\quad +[\lambda_{LS}(x_c^k)-\lambda^k]^T h(x^k).\end{aligned} \tag{3.39}$$

Using a Taylor expansion, (1.4), hypothesis H2, (2.9), (3.1), (3.4)–(3.7), and (3.38), we get

$$\begin{aligned}&L(x_c^{k+1},\lambda_{LS}(x_c^{k+1})) - L(x^k,\lambda_{LS}(x_c^{k+1}))\\ &= g_p(x_c^{k+1})^T(x_c^{k+1}-x^k) + O(\|x_c^{k+1}-x^k\|^2)\\ &= O(\|g_p(x_c^k)\|\rho^k + {\rho^k}^2).\end{aligned} \tag{3.40}$$

Since (2.7) implies that ${\rho^k}^2 \le \rho_{max}^k \|g_p(x_c^k)\|\rho^k$, (3.40) ensures that the first term in the right-hand side of (3.39) is $O(\|g_p(x_c^k)\|\rho^k)$.

From (3.30) and (3.38), we deduce that $\|\lambda^{k+1}-\lambda_{LS}(x_c^{k+1})\|$ and $\|\lambda_{LS}(x_c^k)-\lambda^k\|$ are $O(\|g_p(x_c^k)\|)$. From (3.35) and (3.37), we also obtain $\|\lambda_{LS}(x_c^{k+1})-\lambda_{LS}(x_c^k)\| = O(\|g_p(x_c^k)\|)$. Finally, (2.9) ensures that $\|h(x_c^{k+1})\| \le \|h(x^k)\| \le 2\rho^k$. This implies that the remaining three terms in the right-hand side of (3.39) are also $O(\|g_p(x_c^k)\|\rho^k)$. Together with (2.7), this ensures that there exists $\xi_7 > 0$ such that

$$\Delta L_V^{k+1} \le \xi_7 \rho_{max}^k \|g_p(x_c^k)\|^2. \tag{3.41}$$

Let $\overline{\rho}_{max}$ be defined by

$$\overline{\rho}_{max} = \min\left\{\frac{\xi_1\xi_2}{2\xi_7},\ \frac{10^{-4}}{4\xi_0(\xi_0+1)}\left(\frac{\xi_1\xi_3}{\xi_7}\right)^2\right\}. \tag{3.42}$$

With arguments entirely similar to those used to show (3.31)–(3.32), we can prove, from (3.41) and (3.42), that, if $\rho_{max}^{k_0} < \overline{\rho}_{max}$ and $k \ge k_0$, then $\Delta L_V^{k+1} < -\frac{1}{2}\Delta L_H^k$. Therefore, ρ_{max}^k does not change after k_0. ∎

We say that a point x is stationary for (1.1), i.e., it satisfies the KKT conditions for the problem, if $h(x) = 0$ and $g_p(x) = 0$. The next theorem states that, under H1–H4, the sequence $\{{x_c}^k\}$ generated by the DCI algorithm has stationary points for (1.1) in its accumulation set. Some additional conditions are defined to ensure that every accumulation point is stationary for (1.1).

THEOREM 3.6. *Under H1–H4, either DCI stops at a stationary point for* (1.1)*, in a finite number of iterations, or generates a sequence with stationary points in its accumulation set. In addition, if we impose the horizontal step and the Lagrange multipliers to satisfy* (3.29) *and* (3.30)*, then every accumulation point of* ${x_c}^k$ *is stationary for* (1.1).

Proof. Let us suppose, by contradiction, that $\liminf(\|g_p(x_c^k)\|) = 2b > 0$, and let $\overline{k}_0$ be such that $\|g_p(x_c^k)\|) > b$ for $k \ge \overline{k}_0$. In this case, item (ii) from Lemma 3.5 allows us to choose $k_0 \ge \overline{k}_0$ such that, for every $k \ge k_0$, $\rho_{max}^k = \rho_{max}^{k_0}$. Together with (2.8) and (3.5), this implies that $\rho^k \ge 10^{-4}\rho_{max}^{k_0} b/[2(\xi_0+1)]$.

Now, using (3.24) and (3.13), we can assure that, for $k > k_0$,

$$L(x_c^k, \lambda^k) - L(x_c^{k_0}, \lambda^{k_0}) = \sum_{i=k_0+1}^{k} \Delta L_c^i \leq \frac{1}{4} \sum_{i=k_0}^{k-1} \Delta L_H^i + r^{k_0}$$
$$\leq -(k - k_0)\theta + r^{k_0} \to -\infty,$$

where

$$\theta = \frac{1}{4}\xi_1 b \min\left\{\xi_2 b,\ 10^{-2}\xi_3\sqrt{\frac{b\rho_{max}^{k_0}}{\xi_0 + 1}}\right\} > 0. \tag{3.43}$$

This contradicts H1–H2, imposing $\liminf(\|g_p(x_c^k)\|) = 0$.

For the second part of the theorem, let us assume that (3.29) and (3.30) apply. In this case, Lemma 3.5 ensures that $\rho_{max}^k = \rho_{max}^{k_0}$ for some k_0 and every $k \geq k_0$.

Suppose, by contradiction, that $\|g_p(x_c)^{k_\ell})\| \geq b > 0$ for an infinite subsequence $\{k_\ell\}$. Let n_k be the number of iterations between k_0 and k for some index $k \in \{k_\ell\}$. In this case, using (3.24) and (3.13) again and taking $n_k \to \infty$, we have

$$L(x_c^k, \lambda^k) - L(x_c^{k_0}, \lambda^{k_0}) = \sum_{i=k_0+1}^{k} \Delta L_c^i \leq \frac{1}{4} \sum_{i=k_0}^{k-1} \Delta L_H^i + r^{k_0}$$
$$\leq -n_k\theta + r^{k_0} \to -\infty, \tag{3.44}$$

where θ is given by (3.43). This also contradicts H1–H2, implying that $\|g_p(x_c^{k_\ell})\| \to 0$ for every subsequence of x_c^k. ∎

Theorem 3.6 can equally be proved if we admit inexact solutions for the subproblems associated with Algorithm 2.1, using fairly loose conditions on the residues for accepting the step. For instance, we could relax the condition $\nabla h(x_c)\,\delta_t = 0$ or admit inexact computations of $g_p(x)$ and the solution of the quadratic subproblem (2.4). Although this modification can be interesting for large-scale problems and would not change the proofs significantly, we preferred not to present it in this article, since its details might look rather messy on a first reading. We also believe that the second order correction would play a very interesting role if inexact methods were used.

4. Local convergence. Let $N(M)$ represent the null space of M. Also let $\{x_c^k\}$ and $\{x^k\}$ be sequences generated by Algorithm 2.1, converging to x^*, a "good minimizer" of problem (1.1). By "good minimizer" we mean that $\nabla h(x^*)$ has full row rank, $\nabla f(x^*) = -\nabla h(x^*)^T\lambda^*$ with $\lambda^* = \lambda_{LS}(x^*)$, and there is a constant $\mu_1 > 0$ such that, for $y \in N(\nabla h(x^*))$,

$$\mu_1\|y\|^2 \leq y^T\nabla_{xx}^2 L(x^*, \lambda^*)y. \tag{4.1}$$

In this section, we will restrict our attention to a neighborhood V^* of x^*, where, due to the fact that h is C^2 and $\nabla h(x^*)$ has full row rank, the orthogonal projector onto $N(\nabla h(x))$, i.e., $P(x) = I - \nabla h(x)^T(\nabla h(x)\nabla h(x)^T)^{-1}\nabla h(x)$, is Lipschitz continuous. Sometimes we will use the term δ_c to represent the "full" step taken by the algorithm, i.e., $\delta_c = x_c^{k+1} - x_c^k = \delta_H^k + \delta_V^{k+1}$.

Besides considering hypotheses H1–H4, our analysis of the local convergence of x_c^k and x^k will be based on four additional local assumptions. The first three of these assumptions are used in the proof of Lemma 4.1 and are described below.

A1: $\lambda^k - \lambda_{LS}(x_c^k) = O(\|g_p(x_c^k)\|)$.

A2: B^k is asymptotically uniformly positive definite in the tangent space to the restrictions, which means that, in some neighborhood of x^*, we can redefine μ_1 so that

$$\mu_1\|y\|^2 \leq y^T B^k y \leq \mu_2\|y\|^2 \tag{4.2}$$

for $y \in N(\nabla h(x_c^k))$, where μ_2 is just the constant ξ_0 defined in (3.8).

A3: Let δ_{HN}^k be the minimizer of the quadratic model (2.6) without the trust region constraint. We assume that, whenever δ_{HN}^k is within the trust region ($\|\delta_{HN}^k\| \leq \Delta$), it is the first horizontal step tried by the algorithm. In addition, we also suppose that it satisfies

$$P(x_c^k)(B^k - \nabla_{xx}^2 L(x^*, \lambda^*))\delta_{HN}^k = o(\|\delta_{HN}^k\|).$$

Assumption A1 is not a stringent condition. Usual estimates for the Lagrange multipliers (see, for example, [41]) satisfy $\|\lambda^k - \lambda^*\| = O(\|x_c^k - x^*\|)$, so A1 can be guaranteed by our Lemma 4.2, along with (1.5) and (1.2).

Assumptions A2 and A3 are essentially equivalent to standard conditions for superlinear convergence in two steps of SQP quasi-Newton methods, such as those established by Powell in [34]. These assumptions are satisfied, for example, if we define $B^k = \nabla_{xx} L(x_c^k, \lambda^k)$. In a future paper, we intend to incorporate in our analysis the use of secant reduced Hessian approximation schemes, as well as the inexact solution of the subproblems involved, in such a way that A2 and A3 are satisfied.

From now on, we also suppose that $\delta_{soc} = 0$. This is done only to simplify the exposition. In fact, the arguments presented below are still valid if we consider $\delta_{soc} = O(\|\delta_t\|^2)$.

Since $\nabla h(x)$ and $\nabla_{xx}^2 L(x, \lambda)$ are continuous and $\nabla h(x^*)$ has full row rank, our assumptions imply that there is a constant $\mu_3 > 0$ and a neighborhood V^* of x^* such that, for x, $x_c^k \in V^*$,

$$\mu_3\|\lambda\| \leq \|\nabla h(x)^T \lambda\| \quad \text{for } \lambda \in \mathbb{R}^m, \tag{4.3}$$

$$P(x_c^k)(B^k - \nabla_{xx}^2 L(x_c^k, \lambda_{LS}(x_c^k)))\delta_{HN}^k = o(\|\delta_{HN}^k\|), \quad \text{and} \tag{4.4}$$

$$P(x_c^k)(B^k - \nabla_{xx}^2 L(x_c^k, \lambda^k))\delta_{HN}^k = o(\|\delta_{HN}^k\|). \tag{4.5}$$

Let Z^k be a matrix whose columns form an orthonormal basis for the null space $N(\nabla h(x_c^k))$. We can define the global minimizer of the quadratic model in the tangent space as $\delta_{HN}^k = Z^k \nu^k \in N(\nabla h(x_c^k))$. This point clearly satisfies

$$(Z^k)^T(B^k \delta_{HN}^k + \nabla_x f(x_c^k)) = (Z^k)^T B^k Z^k \nu^k + (Z^k)^T g_p(x_c^k) = 0. \tag{4.6}$$

From (4.2) and the fact that $(Z^k)^T Z^k = I$, matrix $((Z^k)^T B^k Z^k)^{-1}$ satisfies, in the neighborhood V^* and for all $u \in R^{n-m}$,

$$\frac{1}{\mu_2}\|u\|^2 \leq u^T((Z^k)^T B^k Z^k)^{-1} u \leq \frac{1}{\mu_1}\|u\|^2. \tag{4.7}$$

In the next lemma we will prove the eventual acceptance, in Algorithm 2.1, of

$$\delta_{HN}^k = -Z^k \nu^k = -Z^k((Z^k)^T B^k Z^k)^{-1}(Z^k)^T g_p(x_c^k). \tag{4.8}$$

LEMMA 4.1. *δ_{HN}^k is accepted by Algorithm* 2.1 *for k sufficiently large.*

Proof. Combining (4.7) and (4.8), we have that

$$\|\delta_{HN}^k\| \leq \frac{1}{\mu_1}\|g_p(x_c^k)\| \tag{4.9}$$

for $x_c^k \in V^*$. Because the trust region radius satisfies $\Delta \geq \Delta_{min}$ at the beginning of each iteration, assumption A3 and (4.9) imply that, in a suitable V^*, δ_H^+ will be the first horizontal step tried by Algorithm 2.1.

For $\nu \in \mathbb{R}^{n-m}$, the *reduced* polynomial

$$\overline{q}(\nu) = q(Z^k\nu) = ((Z^k)^T g_p(x_c^k))^T\nu + \nu^T((Z^k)^T B^k Z^k)\nu$$

has degree 2, with positive definite quadratic form. Therefore, its minimum, $\overline{q}(\nu^k)$, satisfies (see [13])

$$\begin{aligned} q(\delta_{HN}^k) &= q(Z^k\nu^k) = \overline{q}(\nu^k) \\ &= -\frac{1}{2}((Z^k)^T g_p(x_c^k))^T((Z^k)^T B^k Z^k)^{-1}(Z^k)^T g_p(x_c^k) \\ &\leq -\frac{1}{2\mu_2}\|g_p(x_c^k)\|^2, \end{aligned} \tag{4.10}$$

where the last inequality comes from (4.7).

Now, using a Taylor expansion, the fact that $\delta_{HN}^k = P(x_c^k)\delta_{HN}^k$, (4.5), and (4.9), we get

$$\begin{aligned} \Delta L_H^+ &= L(x_c^k + \delta_{HN}^k, \lambda^k) - L(x_c^k, \lambda^k) \\ &= g_p(x_c^k)^T\delta_{HN}^k + \frac{1}{2}\delta_{HN}^{kT}\nabla_{xx}{}^2 L(x_c^k)\delta_{HN}^k + o(\|\delta_{HN}^k\|^2) \\ &= g_p(x_c^k)^T\delta_{HN}^k + \frac{1}{2}\delta_{HN}^{kT}B^k\delta_{HN}^k + o(\|\delta_{HN}^k\|^2) \\ &= q(\delta_{HN}^k) + o(\|g_p(x_c^k)\|^2). \end{aligned} \tag{4.11}$$

It follows from (4.10)–(4.11) that

$$|r| = \left|\frac{\Delta L_H^+}{q(\delta_{HN}^k)}\right| = 1 + \frac{o(\|g_p(x_c^k)\|^2)}{\|g_p(x_c^k)\|^2/(2\mu_2)} \to 1,$$

so one of the acceptance conditions stated in step 4.2 of Algorithm 2.1 is satisfied for k sufficiently large.

Let us now prove that the other acceptance condition, $\|h(x_c^k + \delta_{HN}^k)\| \leq 2\rho^k$, also holds. From (4.9), assumption A1, and Lemma 3.5, there exists k_0 sufficiently large so that $\rho_{max}^k = \rho_{max}^{k_0} > 0$ for $k \geq k_0$. Therefore, (2.8) and (3.5) guarantee that, for $k \geq k_0$, $\|g_p(x_c^k)\| \leq \beta\rho^k$, where $\beta = 10^4(1+\xi_0)/\rho_{max}^{k_0}$. Together with (3.12) and (4.9), this implies that, for k sufficiently large,

$$\begin{aligned} \|h(x_c^k + \delta_{HN}^k)\| &\leq \|h(x_c^k)\| + \overline{\xi}_0\|\delta_{HN}^k\|^2 \leq \|h(x_c^k)\| + \frac{\overline{\xi}_0}{{\mu_1}^2}\|g_p(x_c^k)\|^2 \\ &\leq \rho^k\left(1 + \beta\frac{\xi_0}{{\mu_1}^2}\|g_p(x_c^k)\|\right). \end{aligned}$$

Since, for k sufficiently large, $\beta\frac{\xi_0}{{\mu_1}^2}\|g_p(x_c^k)\| < 1$, the step δ_{HN}^k will eventually be accepted. ∎

This lemma is based on the fact that $\Delta \geq \Delta_{min}$ at the beginning of an iteration. This condition can be removed if we replace A3 by the following more restrictive assumption:

A3′: Let δ_+ be obtained as a positive linear combination of δ_{CP} and δ_{HN}. Also let δ_+ satisfy

$$P(x_c^k)(B^k - \nabla^2_{xx} L(x^*, \lambda^*))\delta_+ = o(\|\delta_+\|).$$

In this case, we can also prove that δ_+ is accepted whenever x_c^k and x^k are in a suitable neighborhood V^* of x^*. Therefore, there exists k_1 such that Δ^k is not reduced for $k > k_1$, so we can restrict our attention to the case where $\delta_H^k = \delta_{HN}^k$.

Notice that the dynamic control of the infeasibility might force us to compute more than one single vertical step δ_V^+, starting from x^k, if $\|g_p(x^k + \delta_V^+)\|$ is too small.

At the beginning of iteration $k+1$, we have $x_c = x^k$, while the vertical step ends at $x_c = x_c^{k+1}$. In order to avoid unnecessary updates of $\nabla h(x_c^k)$, we state our fourth local assumption:

A4: Each nonzero vertical step $\delta_V^{k+1} = x_c^{k+1} - x^k$ is computed by taking one or more steps in the form

$$\delta_V^+ = -A^T(AA^T)^{-1}h(x_c), \tag{4.12}$$

where A satisfies

$$\|A - \nabla h(x_c)\| = O(\|g_p(x_c^k)\|). \tag{4.13}$$

Vector δ_V^+ given by (4.12) is the usual Gauss–Newton step for solving $h(x) = 0$, with an approximation A for the Jacobian $\nabla h(x_c)$. Using a Taylor expansion, (4.3), (4.12), (4.13), and the continuity of $\nabla h(x)$, it is easy to show that, if $x_c^{k+1} \neq x^k$, then the first vertical step δ_V^+ of iteration $k+1$ satisfies

$$\|\delta_V^+\| = O(\|h(x^k)\|) \quad \text{and} \tag{4.14}$$

$$\|h(x_c^{k+1})\| \leq \|h(x^k + \delta_V^+)\| = o(\|h(x^k)\|). \tag{4.15}$$

As $g_p(x^k)$ becomes small, it is natural to force a restoration after each horizontal step. This can be done, for instance, by choosing ρ^k slightly smaller than $\|h(x^k)\|$. In [11] and [12], Coleman and Conn analyze algorithms that alternate a horizontal step with a single vertical restoration step. Under local assumptions similar to those presented here, these so called horizontal-vertical algorithms are superlinear convergent in two steps. Coleman and Conn also point out in [11] that the restoration step adopted by their methods differs from the usual SQP vertical step, since it is based on $x^k + \delta_H$, while, in the SQP framework, x^k is used to define the vertical subproblem.

One difference between these horizontal-vertical algorithms and ours is that we admit more than one vertical step like (4.12) at each iteration, as mentioned above. Our main result on the local behavior of the algorithm is based on the following lemma that, in a sense, expresses analytically the "good" structure we have in the neighborhood of a "good" KKT point. This approach is similar to the one used by Powell in [34], although his focus was restricted to each sequence generated by an SQP algorithm.

It is well known that the function $\phi(x) = \|h(x)\| + \|g_p(x)\|$ can be used to measure how close $x \in V^*$ is to x^*. However, we need a stronger result. We want to say that, in a vicinity of x^*, $\phi(x)$ is *equivalent* to $\|x - x^*\|$, in the sense that $\|x - x^*\| = \Theta(\phi(x))$, i.e., $\|x - x^*\| = O(\phi(x))$ and $\phi(x) = O(\|x - x^*\|)$.

LEMMA 4.2. *There is a neighborhood V^* of x^*, where*

$$\|x - x^*\| = \Theta(\|h(x)\| + \|g_p(x)\|). \tag{4.16}$$

Proof. We just have to show that

$$\|x - x^*\| = O(\|h(x)\|) + O(\|g_p(x)\|). \tag{4.17}$$

The converse follows trivially from the fact that $h(x)$ and $g_p(x)$ are Lipschitz continuous in V^*.

In [18], Fletcher shows that the SQP method has quadratic convergence to a good minimizer x^*. The same argument can be used to prove that, if δ_x is an SQP step from $x \in V^*$ to $x_+ = x + \delta_x$, where V^* is a suitable neighborhood of x^*, then we have $x_+ - x^* = O(\|x - x^*\|^2)$. It is also easy to show (see equations (10.1.11)–(10.1.13) in [18]) that δ_x satisfies $\delta_x = x_+ - x = O(\|g_p(x)\|) + O(\|h(x)\|)$. From these two relations, it follows that $x - x^* = (x_+ - x^*) - (x_+ - x) = O(\|x_+ - x\|) = O(\|g_p(x)\|) + O(\|h(x)\|)$. ∎

Byrd [7] and Yuan [45] give examples showing that we cannot expect superlinear convergence in one step for x_c^k. However, Byrd [8] points to the possibility of obtaining superlinear convergence in one step for x^k. To understand why this happens, notice that a vertical step that moves from x^k to x_c^{k+1} approaches the feasible set in a "superlinear" way. After that, the horizontal step superlinearly pushes x_c^{k+1} towards the dual manifold $\mathcal{L}^* = \{x \in V^* : g_p(x) = 0\}$, with δ_H^{k+1} tangent to the feasible directions. Therefore, this horizontal step does not destroy the "vertical superlinear approximation." On the other hand, if we start at x_c^k, the superlinear convergence in a single step cannot be guaranteed since the vertical step δ_V^{k+1} usually is not tangent to $\mathcal{L}^*$ and, for this reason, δ_V^{k+1} can partly spoil the good approach to $\mathcal{L}^*$ obtained by δ_H^k.

To close this section, we present our main theorem, showing that the algorithm is 2-step superlinearly convergent. Besides, convergence in one step can also be obtained if we call a restoration at each iteration.

THEOREM 4.3. *Under* H1–H4 *and* A1–A4, *x^k and x_c^k are 2-step superlinearly convergent to x^*. If a restoration is computed at each x^k, then x^k converges superlinearly to x^*.*

Proof. Since $\|x^{k+1} - x^*\| \le \|x^{k+1} - x_c^{k+1}\| + \|x_c^{k+1} - x^*\|$, (4.9) and (4.16) imply that

$$\|x^{k+1} - x^*\| = O(\|x_c^{k+1} - x^*\|). \tag{4.18}$$

In addition, observing that $\|x_c^{k+1} - x^*\| \le \|x^k - x_c^{k+1}\| + \|x^k - x^*\|$ and using (3.1) and (4.16), we have

$$\|x_c^{k+1} - x^*\| = O(\|x^k - x^*\|). \tag{4.19}$$

In order to prove the 2-step superlinear convergence, we shall use the following relations:

$$g_p(x^k) = o(\|x_c^k - x^*\|), \tag{4.20}$$

$$g_p(x_c^{k+1}) = o(\|x_c^{k-1} - x^*\|), \tag{4.21}$$

$$h(x^k) = o(\|x_c^{k-1} - x^*\|), \text{ and} \tag{4.22}$$

$$h(x_c^{k+1}) = o(\|x_c^k - x^*\|). \tag{4.23}$$

Let us show that these relations are valid, starting with (4.20). Using a Taylor expansion along with (1.4), we have, for $x^k \in V^*$,

$$\begin{aligned} \|g_p(x^k)\| &= \|P(x^k)\Gamma^k\| + o(\|\delta_H^k\|) \\ &\le \|(P(x^k) - P(x_c^k))\Gamma^k\| + \|P(x_c^k)\Gamma^k\| + o(\|\delta_H^k\|), \end{aligned} \tag{4.24}$$

where $\Gamma^k = g_p(x_c^k) + \nabla^2_{xx} L(x_c^k, \lambda_{LS}(x_c^k))\delta_H^k$.

The continuity of $P(x)$ in V^* and (4.9) give us

$$\|(P(x^k) - P(x_c^k))\Gamma^k\| = o(\|\Gamma^k\|) = o(\|g_p(x_c^k)\|). \tag{4.25}$$

In addition, (4.4), (4.8), and (4.9) imply that

$$\|P(x_c^k)\Gamma^k\| = \|P(x_c^k)(g_p(x_c^k) + B^k\delta_H^k)\| + o(\|\delta_H^k\|) = o(\|g_p(x_c^k)\|). \tag{4.26}$$

Substituting (4.25) and (4.26) into (4.24) and also considering (4.16), we get (4.20).

To prove (4.23), we need to consider separately two situations. First, let k_i be an infinite subsequence at which no vertical step was made, i.e., $x_c^{k_i+1} = x^{k_i}$. In this case, the dynamic control of the infeasibility, together with (4.20), implies that

$$\|h(x_c^{k_i+1})\| = O(\|g_p(x_c^{k_i+1})\|) = O(\|g_p(x^{k_i})\|) = o(\|x_c^{k_i} - x^*\|). \tag{4.27}$$

Now let us consider an infinite subsequence of iterations k_j at which at least one vertical step δ_V^+ satisfying A4 was made. In this case, (4.15) and (4.16) imply that

$$\|h(x_c^{k_j+1})\| \le \|h(x_c^{k_j} + \delta_V^+)\| = o(\|h(x^{k_j})\|) = o(\|x^{k_j} - x^*\|). \tag{4.28}$$

Equation (4.23) follows directly from (4.18), (4.27), and (4.28).

Combining (3.12), (4.9), (4.16), (4.18), (4.19), and (4.23), we can write

$$\begin{aligned} \|h(x^k)\| &= \|h(x_c^k)\| + O(\|{\delta_H}^k\|^2) = h(x_c^k) + O(\|g_p(x_c^k)\|^2) \\ &= \|h(x_c^k)\| + O(\|x_c^k - x^*\|^2) = o(\|x_c^{k-1} - x^*\|), \end{aligned} \tag{4.29}$$

so (4.22) is proved.

Finally, to obtain (4.21), we use a Taylor expansion, (3.1), (4.18), (4.19), (4.20), and (4.22), so

$$\begin{aligned} \|g_p(x_c^{k+1})\| &= \|g_p(x^k)\| + O(\|x^k - x_c^{k+1}\|) \\ &= \|g_p(x^k)\| + O(\|h(x^k)\|) = o(\|x_c^{k-1} - x^*\|). \end{aligned}$$

The 2-step superlinear convergence of x_c^k and x^k follows from (4.16) and (4.18)–(4.23), since these equations imply that

$$\|x^{k+1} - x^*\| = O(\|g_p(x^{k+1})\| + \|h(x^{k+1})\|) = o(\|x^{k-1} - x^*\|) \text{ and} \tag{4.30}$$

$$\|x_c^{k+1} - x^*\| = O(\|g_p(x_c^{k+1})\| + \|h(x_c^{k+1})\|) = o(\|x_c^{k-1} - x^*\|). \tag{4.31}$$

In order to conclude the proof, let us assume a nonzero restoration step is done at each iteration. Then, (3.12) and (4.15), together with (4.9), (4.16), and (4.19), allow us to improve (4.22), obtaining

$$\|h(x^k)\| = \|h(x_c^k)\| + O(\|g_p(x_c^k)\|^2) = o(\|x^{k-1} - x^*\|). \tag{4.32}$$

Substituting (4.20) and (4.32) into (4.30) and also considering (4.19), we get

$$\begin{aligned} \|x^{k+1} - x^*\| &= O(\|g_p(x^{k+1})\| + \|h(x^{k+1})\|) \\ &= o(\|x_c^{k+1} - x^*\| + \|x^k - x^*\|) = o(\|x^k - x^*\|), \end{aligned}$$

so the desired superlinear convergence of x^k to x^* is attained. ∎

5. Numerical experience. The success of an algorithm is based not only on its theoretical convergence results, but also on its practical behavior. In this section, we present one possible implementation for the DCI algorithm, along with the numerical results obtained by applying it to some problems from the CUTEr collection [21].

We do not claim to have implemented the ultimate version of the algorithm. On the contrary, our implementation is quite simple and should be improved in order to compete with modern commercial codes. Our only purpose is to show that the algorithm can successfully solve medium-sized equality constrained problems. Some hints on how to improve the code are given in the next section.

5.1. A practical implementation of the algorithm. We begin the detailed description of the algorithm by explaining how the vertical and horizontal steps can be implemented. After that, we discuss how to solve the linear systems that appear when computing these steps. Finally, we present a second order correction used to reduce the infeasibility after applying the horizontal step.

5.1.1. Vertical step. Whenever $\|h(x_c)\| > \rho$ at the beginning of an iteration, we need to reduce the infeasibility. Unfortunately, this test is tricky to perform, since ρ depends on $n_p(x_c)$ and this term, in turn, depends on the matrix $\nabla h(x_c)$. Naturally, it would not be wise to compute $\nabla h(x_c)$ just before calling the restoration, as we will need to update this matrix after this step. For this reason, in step 1.2 of Algorithm 2.1, we define an approximate value for ρ, replacing n_p by

$$n_p^a = \frac{|\Delta L_H|}{|f(x^{k-1}) - f(x_c^{k-1})| + \|\delta_t^{k-1}\|}.$$

The restoration is done by applying Powell's *dogleg* method [33] to the constrained linear least squares problem (2.5), replacing x by x_c. Again, the solution of this problem depends on $\nabla h(x_c)$. Therefore, the first time we try to solve (2.5), we use $A = \nabla h(x_c^{k-1})$. If the infeasibility is not sufficiently reduced, we define $A = \nabla h(x_c)$ and recompute the step.

To find an approximate solution for the trust region problem, the dogleg method uses a path consisting of two line segments. The first connects the origin to the Cauchy point, defined as

$$s_{CS} = -\gamma A(x_c)^T h(x_c),$$

where

$$\gamma = \min\left\{ \frac{\Delta_{VS}}{\|A(x_c)^T h(x_c)\|}, \frac{\|A(x_c)^T h(x_c)\|^2}{\|A(x_c)A(x_c)^T h(x_c)\|^2} \right\}.$$

The second line runs from the Cauchy point to the Newton point

$$s_{NS} = -A(x_c)^T (A(x_c)A(x_c)^T)^{-1} h(x_c). \tag{5.1}$$

If $\|s_{NS}\| \leq \Delta_{VS}$, then the Newton point is the solution of the problem. Otherwise, the point of intersection of the dogleg path and the trust region boundary is chosen.

The trust region radius Δ_{VS} used to compute the vertical step is updated using rules similar to those defined for the horizontal step.

Let P_{red} denote the predicted reduction and A_{red} the actual reduction of the infeasibility. The step is rejected if $A_{red}/P_{red} < 10^{-3}$. In this case, Δ_{VS} is divided by 4. On the other hand, if $A_{red}/P_{red} \geq 0.5$, we double Δ_{VS}.

Sometimes it is necessary to apply the dogleg method several times in order to obtain the desired level of infeasibility. To avoid recomputing A frequently, we try to take a new step using the same matrix whenever the dogleg method is able to reduce $\|h(x_c)\|$ by at least 10%. This expedient is used up to four times in a row, after which A is recalculated.

After the restoration, $\nabla h(x_c)$ is available, and we need to choose ρ satisfying the conditions stated in step 1.3.3 of Algorithm 2.1. These conditions are quite loose, so a good scheme for defining the trust cylinder radius can be devised, taking into account some problem characteristics and the values of ρ_{max} and n_p. In our implementation, however, a naive rule was used. If the approximate ρ computed in step 1.2 satisfies $10^{-4} n_p \rho_{max} \leq \rho \leq n_p \rho_{max}$, we keep this value. Otherwise, we simply define

$$\rho = \min\{n_p \rho_{max},\, 0.75\rho_{max}\}.$$

The reduction obtained by the dogleg method may be small depending on the curvature of h. When this happens, we abandon the constrained linear least squares problem and try to apply the Moré and Thuente line search algorithm [31] to the unconstrained nonlinear least squares problem

$$\text{minimize} \quad \|h(x)\|^2, \tag{5.2}$$

using a BFGS approximation for the second order part of the Hessian of the objective function [14].

Since this last approach is more time-consuming than the dogleg method, it is applied only if $\|h(x_c)\|/\|h(x_{k-1})\| < 0.95$ for three successive dogleg steps. Fortunately, this is unlikely to occur, as the dogleg method usually works well.

5.1.2. Horizontal step. The horizontal step of the method consists in solving the quadratic programming problem (2.6). If Z is a matrix that spans the null space of $\nabla h(x_c)$, then it is possible to rewrite (2.6) as the constrained nonlinear programming problem

$$\begin{aligned} &\text{minimize} \quad g(x_c)^T Zv + \frac{1}{2} v^T Z^T B Z v \\ &\text{subject to} \quad \|Zv\|_\infty \leq \Delta, \end{aligned} \tag{5.3}$$

where δ was replaced by Zv.

One should notice that B need not be positive definite; thus we cannot use the dogleg method to solve (5.3), as we did in the vertical step. Instead, we use the Steihaug–Toint method [40, 42], which is an extension of the conjugate gradient (CG) method for nonconvex problems.

Since computing the product of Z times a vector several times would be too costly, we write the Steihaug–Toint algorithm using δ directly, as described by Lalee, Nocedal, and Plantenga in [23].

The method starts by computing the Cauchy step defined in step 4.1.1 of Algorithm 2.1. If this point falls inside the trust region, it is improved by applying successive CG iterations until $q(\delta) \leq 0.01 q(\delta_{CP})$, a direction of negative curvature is found, or the trust region boundary is violated. In the last two cases, a point on the boundary of the trust region is chosen.

5.1.3. Linear systems. In the core of both the vertical and the horizontal steps, we have linear systems involving AA^T. Such systems need to be solved when we compute

- the Newton step (5.1) in the dogleg method;
- the Lagrange multipliers (1.3) and, consequently, the projected gradient (1.4);
- the second order correction (see (5.4));
- the projection of the residual vector onto $N(A)$ in the Steihaug–Toint method.

Two routines are provided for solving these systems. One is based on the sparse Cholesky decomposition of AA^T. The second uses the CG method to generate an approximate solution.

If we choose to work with the Cholesky decomposition, the approximate minimum degree algorithm of Amestoy, Davis, and Duff [3] is used to reorder the rows and columns of AA^T, so the fill-in created during the factorization is minimized. For the CG method, a band preconditioner has been implemented to accelerate the method.

As an attractive alternative, we could use the augmented system approach to solve such systems, since it reduces the fill-in produced by dense rows in A and keeps the condition number of the matrix under control. Direct methods for solving symmetric indefinite augmented systems are presented, for example, in [6, 39], while iterative approaches are introduced in [19, 5], just to cite a few references. We plan to include one or more of these algorithms in our code in the near future.

5.1.4. Second order correction. In DCI, a second order correction can be used to reduce the infeasibility after the horizontal step, as the acceptance of this compound step is more likely to happen. Clearly, $\delta_{soc} = 0$ would be a possibility for the second order correction term. In fact, any $\delta_{soc} = O(\|\delta_t\|^2)$ is acceptable for global convergence purposes. The nonzero natural candidate corresponds to

$$\begin{aligned} \delta_{soc} &= \operatorname{argmin}\{\|\nabla h(x_c)\,\delta + (h(x_c+\delta_t) - h(x_c))\|\} \\ &= -\nabla h(x_c)^T(\nabla h(x_c)\nabla h(x_c)^T)^{-1}(h(x_c+\delta_t) - h(x_c)). \end{aligned} \tag{5.4}$$

If $g_p(x_c) = g(x_c) + \nabla h(x_c)^T\lambda_{LS} = \operatorname{argmin}\{\|\nabla h(x_c)^T\lambda + g(x_c)\|\}$ is obtained from the Cholesky factorization of $\nabla h(x_c)\nabla h(x_c)^T$, the second order correction turns out to be computationally cheap. On the other hand, if we use iterative methods to compute $g_p(x_c)$, it looks reasonable to relax the convergence to g_p so that we can save some time for computing δ_{soc}.

The second order correction is called if, after computing the horizontal step, we have

$$\|h(x_c+\delta_t)\| > \min\{2\rho,\ 2\|h(x_c)\| + 0.5\rho\}$$

or

$$\|h(x_c)\| \le 10^{-5} \ \text{ and } \ \|h(x_c+\delta_t)\| > \max\{10^{-5},\ 2\|h(x_c)\|\}.$$

If the second order correction is refused, it is not calculated again at the same global iteration of Algorithm 2.1.

5.2. Algorithm performance. To analyze the behavior of the algorithm just described, we used a set of 53 medium-sized equality constrained problems extracted from the CUTEr collection [21]. The selected problems are presented in Table 5.1. The number of variables of the problem is given by n, while m is the number of constraints.

Originally, all of the equality constrained problems of the CUTEr library were selected to compose the test set. However, at this moment, the DCI algorithm is not

TABLE 5.1
Selected medium-sized problems from the CUTEr collection.

Problem	n	m	Problem	n	m
AUG2D	20200	10000	HAGER1	10001	5000
AUG2DC	20200	10000	HAGER2	10001	5000
AUG3D	27543	8000	HAGER3	10001	5000
AUG3DC	27543	8000	LCH	3000	1
CATENA	3003	1000	LUKVLE1	10000	9998
CATENARY	501	166	LUKVLE10	10000	9998
CHAIN	6402	3201	LUKVLE11	9998	6664
DTOC1L	14995	9990	LUKVLE13	9998	6664
DTOC1NA	7495	4990	LUKVLE14	998	664
DTOC1NB	7495	4990	LUKVLE15	997	747
DTOC1NC	7495	4990	LUKVLE16	9997	7497
DTOC1ND	7495	4990	LUKVLE3	10000	2
DTOC2	5998	3996	LUKVLE4	10000	4999
DTOC3	14999	9998	LUKVLE5	10002	9996
DTOC4	14999	9998	LUKVLE6	9999	4999
DTOC5	9999	4999	LUKVLE7	10000	4
DTOC6	10001	5000	LUKVLE8	10000	9998
EIGENA2	2550	1275	LUKVLE9	10000	6
EIGENACO	1640	820	OPTCTRL3	4502	3000
EIGENB2	2550	1275	ORTHRDM2	4003	2000
EIGENBCO	1640	820	ORTHRDS2	1003	500
EIGENC2	2652	1326	ORTHREGA	2053	1024
EIGENCCO	1722	861	ORTHREGC	1005	500
ELEC	600	200	ORTHREGD	1003	500
GRIDNETB	13284	6724	ORTHRGDM	2003	1000
GRIDNETE	13284	6724	ORTHRGDS	1003	500
GRIDNETH	13284	6724			

prepared to handle singular Jacobian matrices, so some of the problems needed to be excluded from the list.

The DCI algorithm was implemented in Fortran 77, and the executable program was generated using the ifort 9.0 compiler, under the Fedora 4 Linux operating system. To evaluate the performance of the new method, it was compared with two freely available nonlinear programming solvers. The first is Lancelot-B, the well-known algorithm distributed along with the GALAHAD library [20]. The second is Ipopt (version 3.3.3) [43], an interior point method that also tackles equality constrained problems quite well. Both codes include a nice interface for solving CUTEr problems.

The tests were performed on a Dell Optiplex GX280 computer, using an Intel Pentium 4 540 processor, with a clock speed of 3.2GHz, 1MB of cache memory, a 800MHz front side bus, and the Intel 915G chipset. Exact first and second derivatives were computed by all of the methods.

The DCI algorithm was designed to declare convergence when both $\|h(x)\| < \epsilon_h$ and $n_p < \epsilon_p$, as well as when $\rho_{max} < \epsilon_r$. However, since Lancelot-B uses the infinity norm in its convergence criteria, we decided to change the first criterion, stopping the algorithm when $\|h(x)\|_\infty < \epsilon_h$ and one of $\|g_p\|_\infty < \epsilon_g$ or $n_p < \epsilon_p$ occurs. Besides, it also terminates if $\|\delta_t\| < \epsilon_d \|x\|$ for 10 successive iterations or if the restoration fails to obtain a feasible point. The constants $\epsilon_h = 10^{-5}$, $\epsilon_g = 10^{-5}$, $\epsilon_p = 10^{-7}$, $\epsilon_r = 10^{-7}$, and $\epsilon_d = 10^{-8}$ were adopted, so the stopping tolerances are compatible with those used in Lancelot-B. The Ipopt stopping tolerances were changed accordingly. Default values were used for the remaining Ipopt parameters. The default settings were also used in Lancelot-B, except for the maximum number of iterations, which was increased to 10000.

Other parameters used in the DCI algorithm are

$$\rho^0_{max} = \max\{10^{-5},\, 5.1\|h(x^0)\|,\, 50\, n_p(x^0)\}, \tag{5.5}$$
$$\Delta^0 = \Delta^0_{VS} = \max\{10\|x^0\|,\, 10^5\},$$

and $\Delta_{min} = 10^{-5}$. For all of the problems presented here, we used the Cholesky decomposition to compute the solution of $(AA^T)s = b$, although, for many of them, it would be preferable to use the preconditioned CG method.

The comparison of the methods was done using the performance profiles defined by Dolan and Moré [16]. To draw the performance profiles for a set S of solvers on a set P of problems, we need to compute, for each problem $p \in P$ and each solver $s \in S$, the performance ratio defined by

$$r_{p,s} = \frac{t_{p,s}}{\min\{t_{p,s} : s \in S\}},$$

where $t_{p,s}$ is the time spent by the solver s to solve problem p. The overall performance of solver s is represented by the function

$$P(t) = \frac{1}{n_p} size\{p \in P : r_{p,s} \leq t\},$$

where n_p is the number of problems considered. In words, $P(t)$ is the fraction of the number of problems that are solved by s within a factor t of the time spent by the fastest solver (for each problem). Plotting $P(t)$, we get a performance profile for a particular solver.

For the 53 equality constrained problems selected, the performance profiles of Lancelot-B, Ipopt, and DCI are shown in Figure 5.1. One can deduce from this figure that the DCI algorithm took less time than Lancelot-B and Ipopt to obtain the solution of 47% of the problems, while Ipopt was the best solver for 45% of the problems, and Lancelot-B took less time in only 9.4% of the cases. Ipopt outperforms DCI for t between 1.5 and 4.5, but, in general, we may say that DCI presented the best performance among the solvers.

DCI and Lancelot-B obtained an optimal solution (i.e., a stationary point for (1.1)) for all of the problems. The Ipopt code, in turn, converged to a point of local infeasibility when solving the LUKVLE16 problem and ran out of memory after spending 2920 seconds searching the solution of the LUKVLE11 problem. For all of the remaining problems, Ipopt also obtained an optimal solution.

To close this section, let us focus our analysis on the behavior of the restoration scheme adopted in DCI, summarized in Figure 5.2. Each point in the figure represents one CUTEr problem. The horizontal coordinate of a point is the percentage of the number of iterations in which only one restoration was done. The vertical coordinate is the percentage of iterations in which more than one restoration was needed. Diagonal lines were included to group the problems by the percentage of iterations with one or more restorations.

The results obtained for the 53 CUTEr problems showed that, on average, the DCI algorithm performed one restoration in 24.3% of the iterations, while it was necessary to perform more than one restoration in only 9.2% of the iterations. Summing the figures, we observe that no restoration was made in about two-thirds of the iterations, on average. Besides, if we consider only the iterations in which more than one restoration was done, the number of restorations was equal to 2 in 62.1% of the cases.

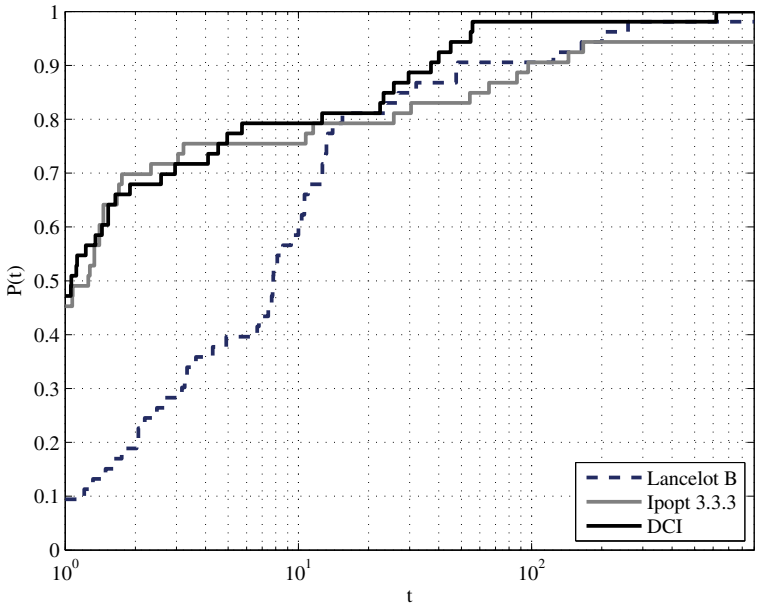


FIG. 5.1. *Performance profiles for* 53 *CUTEr problems.*

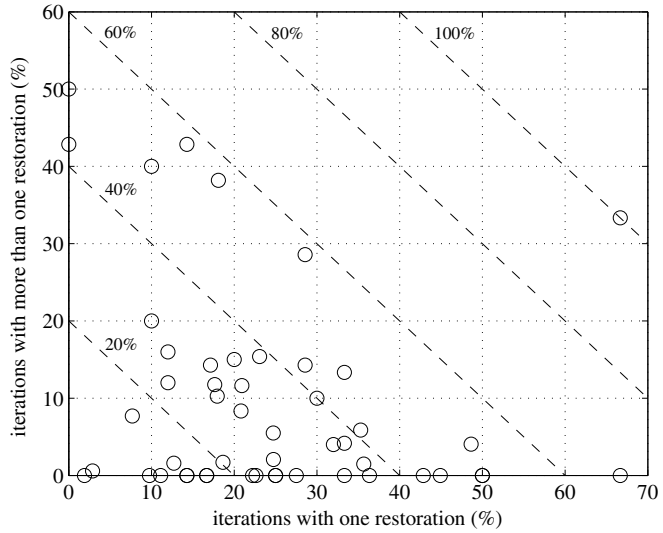


FIG. 5.2. *Percentage of iterations with one or more restorations for the* 53 *CUTEr problems.*

Our experiments with the CUTEr problems also revealed that the choice of an initial value for ρ_{max} is still an open problem. For several problems, a particular value of ρ^0_{max} has led to a much better performance of the algorithm if compared to (5.5). One possible way to circumvent this problem is to use a few iterations of the algorithm only to calibrate this parameter, prior to using the rules for updating it.

6. Conclusions. In this paper, we have presented a new algorithm for solving nonlinear programming problems with equality constraints. The method uses the idea of a trust cylinder to keep the infeasibility under control. The radius of this cylinder is reduced as the algorithm approaches the optimal point. The algorithm is globally convergent in the sense that its accumulation set has stationary points for (1.1). In addition, it is also superlinearly convergent under some mild assumptions.

Our current implementation of the algorithm works well when applied to medium-sized problems, so we believe it worthwhile to investigate its performance for larger problems. Some of the improvements that are to be made to the code after solving large-scale problems include

- the use of an augmented system approach to solve the linear systems;
- the reformulation of the algorithm so that inexact solutions for the linear subroutines are admitted;
- the use of BFGS approximations to the Hessian of the Lagrangian when computing the horizontal step;
- the definition of clever rules for choosing the initial value of ρ_{max}.

Furthermore, we also have plans to extend the algorithm to solve inequality constrained problems.

REFERENCES

[1] J. Abadie and J. Carpentier, *Some Numerical Experiments with the GRG Method for Nonlinear Programming*, Paper HR7422, Électricité de France, Paris, 1967.

[2] J. Abadie and J. Carpentier, *Generalization of the Wolfe reduced gradient method to the case of nonlinear constraints*, in Optimization, R. Fletcher, ed., Academic Press, London, 1969, pp. 37–47.

[3] P. R. Amestoy, T. A. Davis, and I. S. Duff, *An approximate minimum degree ordering algorithm*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 886–905.

[4] L. T. Biegler, J. Nocedal, and C. Schmid, *A reduced Hessian method for large-scale constrained optimization*, SIAM J. Optim., 5 (1995), pp. 314–347.

[5] S. Bonettini, V. Ruggiero, and F. Tinti, *On the solution of indefinite systems arising in nonlinear programming problems*, Numer. Linear Algebra Appl., 14 (2007), pp. 807–831.

[6] J. R. Bunch and L. Kaufman, *Some stable methods for calculating inertia and solving symmetric indefinite linear systems*, Math. Comp., 31 (1977), pp. 163–179.

[7] R. H. Byrd, *An example of irregular convergence in some constrained optimization methods that use the projected Hessian*, Math. Program., 32 (1985), pp. 232–237.

[8] R. H. Byrd, *On the convergence of constrained optimization methods with accurate Hessian information on a subspace*, SIAM J. Numer. Anal., 27 (1990), pp. 141–153.

[9] R. H. Byrd, J. C. Gilbert, and J. Nocedal, *A trust region method based on interior point techniques for nonlinear program.*, Math. Program., 89 (2000), pp. 149–186.

[10] R. H. Byrd, M. E. Hribar, and J. Nocedal, *An interior point algorithm for large-scale nonlinear programming*, SIAM J. Optim., 9 (1999), pp. 877–900.

[11] T. Coleman and A. R. Conn, *Nonlinear programming via an exact penalty function: Asymptotic analysis*, Math. Program., 24 (1982), pp. 123–136.

[12] T. F. Coleman and A. R. Conn, *On the local convergence of a quasi-Newton method for the nonlinear programming problem*, SIAM J. Numer. Anal., 21 (1984), pp. 755–769.

[13] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Trust-Region Methods*, SIAM, Philadelphia, 2000.

[14] J. E. Dennis, H. J. Martinez, and R. A. Tapia, *Convergence theory for the structured BFGS secant method with application to nonlinear least squares*, J. Optim. Theory Appl., 61 (1989), pp. 161–178.

[15] J. E. Dennis and L. N. Vicente, *On the convergence theory of trust-region-based algorithms for equality-constrained optimization*, SIAM J. Optim., 7 (1997), pp. 927–950.

[16] E. D. Dolan and J. J. Moré, *Benchmarking optimization software with performance profiles*, Math. Program., 91 (2002), pp. 201–213.

[17] M. El-Alem, *A global convergence theory for Dennis, El-Alem, and Maciel's class of trust-region algorithms for constrained optimization without assuming regularity*, SIAM J. Optim., 9 (1999), pp. 965–990.

[18] R. Fletcher, *Practical Methods of Optimization*, 2nd ed., John Wiley & Sons, Chichester, UK, 1987.

[19] N. I. M. Gould, M. E. Hribar, and J. Nocedal, *On the solution of equality constrained quadratic programming problems arising in optimization*, SIAM J. Sci. Comput., 23 (2001), pp. 1376–1395.

[20] N. I. M. Gould, D. Orban, and P. L. Toint, *GALAHAD, a library of thread-safe Fortran* 90 *packages for large-scale nonlinear optimization*, ACM Trans. Math. Software, 29 (2003), pp. 353–372.

[21] N. I. M. Gould, D. Orban, and P. L. Toint, *CUTEr (and SifDec), a constrained and unconstrained testing environment, revisited*, ACM Trans. Math. Software, 29 (2003), pp. 373–394.

[22] F. A. M. Gomes, M. C. Maciel, and J. M. Martínez, *Nonlinear programming algorithms using trust regions and augmented Lagrangians with nonmonotone penalty parameters*, Math. Program., 84 (1999), pp. 161–200.

[23] M. Lalee, J. Nocedal, and T. Plantenga, *On the implementation of an algorithm for large-scale equality constrained optimization*, SIAM J. Optim., 8 (1998), pp. 682–706.

[24] J. M. Martínez, *A trust-region SLCP model algorithm for non-linear programming*, in Foundations of Computational Mathematics, F. Cucker and M. Shub, eds., Springer-Verlag, Berlin, 1997, pp. 246–255.

[25] J. M. Martínez, *Two-phase model algorithm with global convergence for nonlinear programming*, J. Optim. Theory Appl., 96 (1998), pp. 397–436.

[26] J. M. Martínez, *Inexact restoration method with Lagrangian tangent decrease and new merit function for nonlinear programming*, J. Optim. Theory Appl., 111 (2001), pp. 39–58.

[27] J. M. Martínez and E. A. Pilotta, *Inexact restoration algorithm for constrained optimization*, J. Optim. Theory Appl., 104 (2000), pp. 135–163.

[28] J. M. Martínez and E. A. Pilotta, *Inexact restoration methods for nonlinear programming: Advances and perspectives*, in Optimization and Control with Applications, L. Qi, K. Teo, and X. Yang, eds., Springer-Verlag, New York, 2005, pp. 271–292.

[29] A. Miele, H. Y. Huang, and J. C. Heideman, *Sequential gradient-restoration algorithm for the minimization of constrained functions—ordinary and gradient versions*, J. Optim. Theory Appl., 4 (1969), pp. 213–243.

[30] A. Miele, A. V. Levy, and E. E. Cragg, *Modifications and extensions of the conjugate gradient-restoration algorithm for mathematical programming problems*, J. Optim. Theory Appl., 7 (1971), pp. 450–472.

[31] J. J. Moré and D. J. Thuente, *Line search algorithms with guaranteed sufficient decrease*, ACM Trans. Math. Software, 20 (1994), pp. 286–307.

[32] H. Mukai and E. Polak, *On the use of approximations in algorithms for optimization problems with equality and inequality constraints*, SIAM J. Numer. Anal., 15 (1978), pp. 674–693.

[33] M. J. D. Powell, *A hybrid method for nonlinear equations*, in Numerical Methods for Nonlinear Algebraic Equations, P. Rabinowitz, ed., Gordon and Breach, London, 1970, pp. 87–114.

[34] M. J. D. Powell, *The convergence of variable metric methods for nonlinearly constrained optimization calculations*, in Nonlinear Programming, 3, O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, eds., Academic Press, New York, 1978, pp. 27–63.

[35] M. Rom and M. Avriel, *Properties of the sequential gradient-restoration algorithm (SGRA). 1. Introduction and comparison with related methods*, J. Optim. Theory Appl., 62 (1989), pp. 77–98.

[36] M. Rom and M. Avriel, *Properties of the sequential gradient-restoration algorithm (SGRA). 2. Convergence analysis*, J. Optim. Theory Appl., 62 (1989), pp. 99–125.

[37] J. B. Rosen, *The gradient projection method for nonlinear programming. Part* I. *Linear constraints*, SIAM J. Appl. Math., 8 (1960), pp. 181–217.

[38] J. B. Rosen, *The gradient projection method for nonlinear programming. Part* II. *Nonlinear constraints*, SIAM J. Appl. Math., 9 (1961), pp. 514–532.

[39] O. Schenk and K. Gärtner, *On fast factorization pivoting methods for symmetric indefinite systems*, Electron. Trans. Numer. Anal., 23 (2006), pp. 158–179.

[40] T. Steihaug, *The conjugate gradient method and trust regions in large scale optimization*, SIAM J. Numer. Anal., 20 (1983), pp. 626–637.

[41] R. A. Tapia, *Quasi-Newton methods for equality constrained optimization: Equivalence of existing methods and a new implementation*, in Nonlinear Programming, 3, O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, eds., Academic Press, New York, 1978, pp. 125–164.

[42] Ph. L. Toint, *Towards an efficient sparsity exploiting Newton method for minimization*, in Sparse Matrices and Their Uses, I. S. Duff, ed., Academic Press, London, 1981, pp. 57–88.
[43] A. Wächter and L. T. Biegler, *On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming*, Math. Program., 106 (2006), pp. 25–57.
[44] P. Wolfe, *Methods of nonlinear programming*, in Recent Advances in Mathematical Programming, R. L. Graves and P. Wolfe, eds., McGraw–Hill, New York, 1963, pp. 67–86.
[45] Y. Yuan, *An only* 2*-step Q-superlinear convergence example for some algorithms that use reduced Hessian approximations*, Math. Programming, 32 (1985), pp. 224–231.

SIAM J. OPTIM.
Vol. 19, No. 3, pp. 1326–1343

OPTIMIZATION PROBLEMS WITH SECOND ORDER STOCHASTIC DOMINANCE CONSTRAINTS: DUALITY, COMPACT FORMULATIONS, AND CUT GENERATION METHODS*

GÁBOR RUDOLF† AND ANDRZEJ RUSZCZYŃSKI‡

Abstract. For stochastic optimization problems with second order stochastic dominance constraints we develop a new form of the duality theory featuring measures on the product of the probability space and the real line. We present two formulations involving small numbers of variables and exponentially many constraints: primal and dual. The dual formulation reveals connections between dominance constraints, generalized transportation problems, and the theory of measures with given marginals. Both formulations lead to two classes of cutting plane methods. Finite convergence of both methods is proved in the case of finitely many events. Numerical results for a portfolio problem are provided.

Key words. stochastic programming, stochastic dominance, duality, cutting plane methods

AMS subject classifications. Primary, 90C15, 90C48, 65K05; Secondary, 90C05, 60E15

DOI. 10.1137/070702473

1. Introduction.

Our objective is to develop new approaches to stochastic optimization problems with a constraint in the form of the second order stochastic dominance relation. Such problems, introduced and analyzed in [4, 5], are new models of risk-averse optimization, in which risk aversion is expressed by the stochastic dominance constraint. Due to its specific structure, the constraint poses new theoretical and computational challenges.

The relation of *stochastic dominance* (introduced in statistics in [18, 19] and in economics in [14, 27]) is defined as follows: Let X and Y be random variables on a probability space $(\Omega, \mathcal{F}, P)$ with distribution functions F_X and F_Y, respectively. We say that X *dominates* Y *in the first order* if $F_X(\eta) \leq F_Y(\eta)$ for all $\eta \in \mathrm{R}$, and we denote this relation by $X \succeq_{(1)} Y$. An equivalent condition is that for every nondecreasing function $u(\cdot)$ one has

$$\mathrm{E}[u(X)] \geq \mathrm{E}[u(Y)], \tag{1.1}$$

provided the expected values above are finite.

For two integrable random variables X and Y, we say that X *dominates* Y *in the second order* if $\int_{-\infty}^{\eta} F_X(t)\,dt \leq \int_{-\infty}^{\eta} F_Y(t)\,dt$ for all $\eta \in \mathrm{R}$, and we denote this relation by $X \succeq_{(2)} Y$. An equivalent condition is that for every concave nondecreasing function $u(\cdot)$ condition (1.1) holds true, provided that the expected values on both sides are finite.

We refer the readers to the monographs [21, 32] for a modern view on the stochastic dominance relations and other comparison methods for random outcomes.

*Received by the editors September 10, 2007; accepted for publication (in revised form) July 28, 2008; published electronically November 19, 2008. This research was supported by NSF grants CCR-0306558 and DMS-0603728.

http://www.siam.org/journals/siopt/19-3/70247.html

†RUTCOR, Rutgers University, 640 Bartholomew Rd., Piscataway, NJ 08854 (grudolf@new-rutcor.rutgers.edu).

‡Department of Management Science and Information Systems, Rutgers University, 94 Rockafeller Rd., Piscataway, NJ 08854 (rusz@business.rutgers.edu).

More generally, for an interval $I \subset \mathrm{R}$ let $X \succeq_{(2,I)} Y$ denote the relation

$$\int_{-\infty}^{\eta} F_X(t)\,dt \le \int_{-\infty}^{\eta} F_Y(t)\,dt \quad \forall\, \eta \in I.$$

It is a relaxation of the second order dominance relation. If the interval I is compact, then this relaxation allows us to overcome technical difficulties in dealing with the second order dominance relation, as discussed in [4, 5]. If the interval I is reduced to one point, then the relation $X \succeq_{(2,I)} Y$ becomes the *integrated chance constraint* of [16].

An alternative representation of the second order dominance relation can be derived by using the *shortfall* of a random variable X from a target $\eta \in \mathrm{R}$, defined as $\max(0, \eta - X)$ (and written compactly as $[\eta - X]_+$). By changing the order of integration one can easily verify that the expected value of the shortfall is given by the formula $\mathrm{E}([\eta - X]_+) = \int_{-\infty}^{\eta} F_X(t)\,dt$. Therefore, we can rewrite the relation $X \succeq_{(2,I)} Y$ in the following form:

$$\mathrm{E}([\eta - X]_+) \le \mathrm{E}([\eta - Y]_+) \quad \forall\, \eta \in I. \tag{1.2}$$

Consider a stochastic model in which our decisions $z \in Z$ affect a random outcome $X = G(z)$. We assume that $z \in Z \subset \mathcal{Z}$, where $\mathcal{Z}$ is a Banach space and Z is a convex closed set. The mapping $G : \mathcal{Z} \to \mathcal{L}_1(\Omega, \mathcal{F}, P)$ is assumed to be continuous and concave in the sense that for P-almost all $\omega \in \Omega$ the function $z \mapsto [G(z)](\omega)$ is concave. Finally, let $f : \mathcal{Z} \to \mathrm{R}$ be a concave objective functional (for example, $f(z) = \mathrm{E}G(z)$). We are interested in the following problem:

$$\begin{array}{ll} \underset{z}{\text{maximize}} & f(z) \\ \text{subject to} & G(z) \succeq_{(2,I)} Y, \\ & z \in Z. \end{array} \tag{1.3}$$

Here $Y \in \mathcal{L}_1(\Omega, \mathcal{F}, P)$ is a benchmark random outcome and I is an interval in R.

As the second order dominance relation carries over to expectations of concave nondecreasing utility functions, no risk-averse decision maker will prefer random outcome Y over random outcome $G(z)$ (if $I = \mathrm{R}$). Therefore, if the benchmark outcome Y represents an "acceptable" risk exposure, the risk exposure of $G(z)$ is even "more acceptable." Furthermore, suppose that the objective functional is monotone (consistent) with respect to the second order stochastic dominance relation, as defined in [24, 25, 26]: $G(z') \succeq_{(2)} G(z) \Rightarrow f(z') \ge f(z)$. For example, we may use $f(z) = \mathrm{E}[G(z)]$ or $f(z)$ being a negative of a coherent measure of risk. If the solution of problem (1.3) is unique, then no other feasible outcome $G(z')$ can strictly dominate the solution $G(z)$ (see [24, 25, 26]). The essence of the approach via stochastic dominance constraints is that the distribution of the outcome $G(z)$ is indirectly shaped by the distribution of the benchmark Y, which may also be an artificially constructed random variable. Reference [23] illustrates this modeling flexibility on an example of a portfolio problem.

The papers [4, 5] provide the optimality and duality theory for problem (1.3) in which Lagrange multipliers associated with the dominance constraints are identified with concave nondecreasing utility functions. In [7] an equivalent inverse form of the second order stochastic dominance constraint was analyzed, and it was shown that it is equivalent to a continuum of conditional (average) value at risk constraints [29].

Moreover, Lagrange multipliers associated with the inverse form of stochastic dominance constraints were identified in [7] with concave rank dependent utility functions of the dual utility theory [34]. In this way, model (1.3) is related to several classical models of risk-averse decision making.

However, an efficient solution of problem (1.3), even in the finite-dimensional linear case, remains a challenge.

In what follows we focus on the stochastic dominance constraint $G(z) \succeq_{(2,I)} Y$ as the novel element in model (1.3), leaving aside considerations about possible objective functionals. We also remark that setting the problem in a Banach space $\mathcal{Z}$ does not lead to any significant technical difficulties, as compared to the finite-dimensional case $\mathcal{Z} = \mathrm{R}^n$. Moreover, we hope to apply our formulation to multistage stochastic optimization problems, with $\mathcal{Z}$ representing the space of policies, which is usually modeled as a subspace of the space of integrable functions (see [30]).

Using (1.2) we obtain a more explicit formulation of (1.3):

$$\begin{aligned} &\underset{z}{\text{maximize}} && f(z) \\ &\text{subject to} && \mathrm{E}([\eta - G(z)]_+) \le \mathrm{E}([\eta - Y]_+) \quad \forall\, \eta \in I, \\ &&& z \in Z. \end{aligned} \tag{1.4}$$

When the functions $f(\cdot)$ and $G(\cdot)$ are affine and the set Z is a convex closed polyhedron, in section 2 we develop a linear programming formulation of problem (1.4). But even in the finite-dimensional case, this problem is difficult to solve because its size grows quadratically with the number of the elementary events considered.

Another approach to (1.4) is the dual method of [5]. It is a specialized nonsmooth optimization algorithm applied to the dual problem, in the space of concave nondecreasing functions playing the role of Lagrange multipliers associated with the dominance constraint. While efficient for some problems, especially portfolio problems of [8], the dual method is rather complicated.

Our objective is to develop new efficient linear programming formulations which exploit the specific structure of the stochastic dominance constraint in cut generation schemes. This results in a significant increase of the size of computationally tractable problems, as well as in a speedup in the solution of smaller instances. Furthermore, for problems with first order stochastic dominance constraints $G(z) \succeq_{(1)} Y$, which are typically much more difficult, due to the potential nonconvexity of the feasible region, model (1.3) serves as a powerful convex relaxation (see [6, 22, 23]). Thus, the speedup also benefits some advanced iterative methods of [23] for problems with first order constraints.

In sections 2 and 3 we present a primal cutting plane method based on formulation (1.4). In section 4 we develop a new version of the duality theory for an extended reformulation of problem (1.4). In section 5 we show how a reduction of the number of variables in the dual problem can be achieved by employing Strassen's theorem about the existence of measures on product spaces with given marginals. This leads to a dual cutting plane method in section 7. Finally, in section 8 we present numerical results, along with performance comparisons of the various methods, for portfolio optimization problems based on real data.

2. A linear representation of the second order stochastic dominance constraint. In order to solve (1.3) it is necessary to represent the relation $\succeq_{(2,I)}$ in a tractable form. The usual approach to achieve this is to introduce *shortfall functions*. In the finite-dimensional case they correspond to slack variables, but in

the infinite-dimensional case we need to introduce an appropriate space of the shortfall functions.

Denote by ℓ the Lebesgue measure on I, and let $\mathcal{B}$ be the σ-algebra of Borel subsets of I. We denote the Banach space of continuous functions on I by $\mathcal{C}(I)$. Let $\mathcal{S}$ be the vector space of all real-valued measurable functions s on $(I\times\Omega, \mathcal{B}\times\mathcal{F}, P\times\ell)$ satisfying the following conditions:

(i) for every $\eta\in I$ the function $s(\eta,\cdot)$ is an element of $\mathcal{L}_1(\Omega,\mathcal{F},P)$;
(ii) for P-almost all $\omega\in\Omega$ the function $s(\cdot,\omega)$ is an element of $\mathcal{C}(I)$;
(iii) the function $\omega\to\max_{\eta\in I}|s(\eta,\omega)|$ is an element of $\mathcal{L}_1(\Omega,\mathcal{F},P)$.

Owing to the Lebesgue theorem, the function $w(\eta)=\int_\Omega s(\eta,\omega)\,dP$ is an element of $\mathcal{C}(I)$. It can be verified directly from the definition that $\mathcal{S}$ is a Banach space with the norm

$$\|s\|=\int_\Omega \max_{\eta\in I}|s(\eta,\omega)|\,dP.$$

Immediately from (1.2) we obtain the following observation.

LEMMA 2.1. *Assume that $X,Y\in\mathcal{L}_1(\Omega,\mathcal{F},P)$. Then $X\succeq_{(2,I)}Y$ if and only if there exists a nonnegative function $s\in\mathcal{S}$ such that*

$$s(\eta,\omega)\ge\eta-X(\omega)\quad\forall\,\eta\in I\quad\forall\,\omega\in\Omega,$$
$$\int_\Omega s(\eta,\omega)\,dP\le\int_\Omega[\eta-Y(\omega)]_+\,dP\quad\forall\,\eta\in I.$$

Let us introduce the notation $v(\eta)=\mathrm{E}([\eta-Y]_+)=\int_\Omega[\eta-Y(\omega)]_+\,dP$ for the shortfalls of the benchmark variable. Applying Lemma 2.1, we can formulate another optimization problem which is equivalent to (1.3):

$$\begin{aligned}
&\underset{z,s}{\text{maximize}} && f(z)\\
&\text{subject to} && \int_\Omega s(\eta,\omega)\,dP\le v(\eta)\quad\forall\,\eta\in I,\\
&&& [G(z)](\omega)+s(\eta,\omega)\ge\eta\quad\forall\eta\in I\quad\forall\,\omega\in\Omega,\\
&&& s\ge 0,\\
&&& z\in Z,\ s\in\mathcal{S}.
\end{aligned}\tag{2.1}$$

If the functional $f(\cdot)$ and the mapping $G(\cdot)$ are affine and the set Z is polyhedral, then problem (2.1) becomes a linear programming problem in Banach spaces. When the distribution of the benchmark outcome is discrete, one can restrict the range of η in (2.1) to the realizations of the benchmark Y.

A potential drawback of the above approach is the introduction of the auxiliary variables $s(\cdot,\cdot)$ indexed by the set $I\times\Omega$. As we shall see in section 8, even in the finite-dimensional case, when $\mathcal{Z}=\mathrm{R}^n$ and the probability space Ω is finite, formulation (2.1) may be impractical to solve.

We now present an alternative representation which does not require additional variables $s(\cdot,\cdot)$. It is an extension of the representation developed in [17] for integrated chance constraints.

THEOREM 2.2. *Assume that $X,Y\in\mathcal{L}_1(\Omega,\mathcal{F},P)$. Then $X\succeq_{(2,I)}Y$ if and only if for all $\eta\in I$ and all events $A\in\mathcal{F}$*

$$\mathrm{E}\left((\eta-X)\,1_A\right)\le v(\eta).$$

Proof. For every event $A \in \mathcal{F}$

$$\mathrm{E}\left((\eta - X)\,1_A\right) \le \mathrm{E}([\eta - X]_+).$$

This inequality becomes an equation if $A = \{X < \eta\}$, and thus

$$\max_{A \in \mathcal{F}} \mathrm{E}\left((\eta - X)\,1_A\right) = \mathrm{E}([\eta - X]_+).$$

The theorem immediately follows from (1.2). □

Using this result we obtain another equivalent formulation of the optimization problem (1.2):

$$\begin{array}{ll} \underset{z}{\text{maximize}} & f(z) \\ \text{subject to} & \displaystyle\int_A \big(\eta - G(z)\big)\,dP \le v(\eta), \quad \forall\, \eta \in I, \quad \forall\, A \in \mathcal{F}, \\ & z \in Z. \end{array} \tag{2.2}$$

Although the auxiliary variables are no longer present, we have introduced an infinite family of constraints indexed by the set $I \times \mathcal{F}$. However, we shall show that the new family of constraints can be efficiently dealt with by a cut generation method.

3. A primal cutting plane method. In this section we assume that $I = \mathrm{R}$ and the probability space Ω is finite, with elementary events $\omega_1, \ldots, \omega_N$ and corresponding probabilities $p_1, \ldots, p_N$. The realizations of the benchmark outcome Y are denoted by $y_1, \ldots, y_D$, and the corresponding benchmark shortfalls are $v_j = \mathrm{E}([y_j - Y]_+)$. We also write $G_i(z)$ for $[G(z)](\omega_i)$.

It is known from [4, 5] that in the case of a discrete benchmark the second order dominance condition $G(z) \succeq_{(2)} Y$ is equivalent to finitely many inequalities:

$$\mathrm{E}([y_j - G(z)]_+) \le v_j, \quad j = 1, \ldots, D. \tag{3.1}$$

We can thus rewrite problem (2.2) as follows:

$$\begin{array}{ll} \underset{z}{\text{maximize}} & f(z) \\ \text{subject to} & \displaystyle\sum_{i \in A} p_i\big(y_j - G_i(z)\big) \le v_j \quad \forall\, j = 1, \ldots, D, \quad \forall\, A \subset \{1, \ldots, N\}, \\ & z \in Z. \end{array} \tag{3.2}$$

The last formulation allows for the construction of a cutting plane method. At iteration k we have a collection of subsets (events) $A_1, \ldots, A_{k-1}$ of $\{1, \ldots, N\}$. We solve a relaxation of (3.2):

$$\begin{array}{ll} \underset{z}{\text{maximize}} & f(z) \\ \text{subject to} & \displaystyle\sum_{i \in A_m} p_i\big(y_j - G_i(z)\big) \le v_j, \quad j = 1, \ldots, D, \quad m = 1, \ldots, k-1, \\ & z \in Z. \end{array} \tag{3.3}$$

If the solution z^k of this problem (which is assumed to exist) satisfies all constraints (3.1), then we stop. Otherwise, we find j^* for which (3.1) is violated, and we define

$$A_k = \{1 \le i \le N : y_{j^*} > G_i(z^k)\}.$$

The iteration index k is increased by one, and we solve (3.3) again.

Since (3.1) is violated,

$$\sum_{i\in A_k} p_i\big(y_{j^*} - G_i(z)\big) > v_{j^*},$$

and thus A_k is different than A_m, $m = 1,\dots,k-1$, used in problem (3.3). As the possible number of sets that can be added is finite, the method must stop at an optimal solution of (3.2). Examples in section 8 suggest that in practice a small number of sets A_k need to be generated in order to find the optimal solution.

4. Lagrangian duality. In this section we derive duality relations for the extended formulation (2.1). Our derivation uses ideas and techniques developed in [5]. The main difference is that we develop duality relations for the formulation (2.1) involving explicit shortfall variables, in contrast to the duality theory of [5], where we focused on the dominance constraint in the nonsmooth formulation (1.4).

The difficulty with formulation (2.1) is that no Slater condition can be formulated for the inequality constraint on the shortfall variables

$$[G(z)](\omega) + s(\eta,\omega) \ge \eta \quad \forall\, \eta \in I \quad \forall\, \omega \in \Omega,$$

because the nonnegative cone in the space $\mathcal{S}$ has no interior. Because of that, we cannot simply apply general duality schemes from [28] or [16]. We need to exploit the special structure of problem (2.1).

At first, we introduce several relevant topological vector spaces. We denote by $\mathbf{rca}(I)$ the space of finite signed measures on I and by $\mathcal{L}_\infty(\Omega,\mathcal{F},P)$ the space of essentially bounded measurable real functions on $(\Omega,\mathcal{F},P)$. Let $\mathcal{M}$ denote the vector space of signed measures on $(I\times\Omega,\mathcal{B}\times\mathcal{F})$ such that for every measure $\lambda\in\mathcal{M}$ the *marginal measures* λ_I and λ_Ω, defined by the equations

$$\lambda_I(B) = \lambda(B\times\Omega), \quad B\in\mathcal{B},$$
$$\lambda_\Omega(A) = \lambda(I\times A), \quad A\in\mathcal{F},$$

satisfy the following conditions:

$$\lambda_I \in \mathbf{rca}(I), \qquad \frac{d\lambda_\Omega}{dP} \in \mathcal{L}_\infty(\Omega,\mathcal{F},P). \tag{4.1}$$

Here we implicitly assume that λ_Ω is absolutely continuous with respect to P.

THEOREM 4.1. *The space $\mathcal{M}$ is the topological dual space to the space $\mathcal{S}$ that is, ℓ is a continuous linear functional on $\mathcal{S}$ if and only if there exists $\lambda\in\mathcal{M}$ such that for all $s\in\mathcal{S}$*

$$\ell(s) = \iint_{I\times\Omega} s(\eta,\omega)\,d\lambda. \tag{4.2}$$

Proof. Fix any $\lambda\in\mathcal{M}$, $\lambda\ge 0$, and consider the linear functional (4.2). Its value can be bounded as follows:

$$|\ell(s)| \le \iint_{I\times\Omega} \max_{\eta\in I}|s(\eta,\omega)|\,d\lambda = \int_\Omega \max_{\eta\in I}|s(\eta,\omega)|\,d\lambda_\Omega$$
$$= \int_\Omega \max_{\eta\in I}|s(\eta,\omega)|\,\frac{d\lambda_\Omega}{dP}(\omega)\,dP \le \Big\|\frac{d\lambda_\Omega}{dP}\Big\|_{\mathcal{L}_\infty}\|s\|.$$

For a general signed measure $\lambda \in \mathcal{M}$ we use its Jordan decomposition into a difference of two nonnegative measures $\lambda = \lambda^+ - \lambda^-$, and we define Ω^+ and Ω^- to be the support sets of λ^+ and λ^-, respectively. Using the last displayed inequality we obtain the estimate

$$\begin{aligned}
|\ell(s)| &\le \left| \iint\limits_{I\times\Omega^+} s(\eta,\omega)\, d\lambda_+ \right| + \left| \iint\limits_{I\times\Omega^-} s(\eta,\omega)\, d\lambda_- \right| \\
&\le \left\| \frac{d\lambda^+_\Omega}{dP} \right\|_{\mathcal{L}_\infty} \int_{\Omega^+} \max_{\eta\in I} |s(\eta,\omega)|\; dP + \left\| \frac{d\lambda^-_\Omega}{dP} \right\|_{\mathcal{L}_\infty} \int_{\Omega^-} \max_{\eta\in I} |s(\eta,\omega)|\; dP \\
&\le \max\left(\left\| \frac{d\lambda^+_\Omega}{dP} \right\|_{\mathcal{L}_\infty}, \left\| \frac{d\lambda^-_\Omega}{dP} \right\|_{\mathcal{L}_\infty} \right) \int_\Omega \max_{\eta\in I} |s(\eta,\omega)|\; dP = \left\| \frac{d\lambda_\Omega}{dP} \right\|_{\mathcal{L}_\infty} \|s\|,
\end{aligned}$$

and we conclude that the linear functional (4.2) is continuous. Thus $\mathcal{S}^* \supset \mathcal{M}$.

To prove the converse inclusion, consider the linear subspace of S:

$$\mathcal{S}_0 = \big\{ s \in \mathcal{S} : s = \varphi\xi,\ \varphi \in \mathcal{C}(I),\ \xi \in \mathcal{L}_1(\Omega,\mathcal{F},P) \big\}.$$

Let $\ell \in \mathcal{S}_0^*$. Fix $A \in \mathcal{F}$, and consider the functional $\varphi \mapsto \ell(\varphi 1_A)$. It is continuous on $\mathcal{C}(I)$. By the Riesz representation theorem, there exists a measure $\mu^\ell_A \in \mathbf{rca}(I)$ such that

$$\ell(\varphi 1_A) = \int_I \varphi(\eta)\, d\mu^\ell_A \quad \forall\, \varphi \in \mathcal{C}(I).$$

Define the measure λ^ℓ on $(I\times\Omega, \mathcal{B}\times\mathcal{F})$ by the formula

$$\lambda^\ell(B\times A) = \mu^\ell_A(B), \quad \forall\, B \in \mathcal{B}, \quad \forall\, A \in \mathcal{F}.$$

Then

$$\ell(\varphi 1_A) = \iint\limits_{I\times\Omega} \varphi(\eta) 1_A(\omega)\, d\lambda^\ell.$$

It follows that for every $s = \varphi\xi$ such that $\varphi \in \mathcal{C}(I)$ and ξ is a step function, i.e., $\xi = \sum_{k=1}^K \alpha_k 1_{A_k}$ with some $\alpha_k \in \mathrm{R}$ and $A_k \in \mathcal{F}$, $k = 1,\dots,K$, the functional ℓ has the form (4.2) with $\lambda = \lambda^\ell$. As the step functions are dense in $\mathcal{L}_1(\Omega,\mathcal{F},P)$, for every $\xi \in \mathcal{L}_1(\Omega,\mathcal{F},P)$ we can find a sequence of step functions $\xi^j \to \xi$, $j \to \infty$. Since ℓ is continuous, we obtain

$$\ell(\varphi\xi) = \lim_{j\to\infty} \ell(\varphi\xi^j) = \lim_{j\to\infty} \iint\limits_{I\times\Omega} \varphi(\eta)\xi_j(\omega)\, d\lambda^\ell = \iint\limits_{I\times\Omega} \varphi(\eta)\xi(\omega)\, d\lambda^\ell,$$

and thus the functional ℓ has the form (4.2) on $\mathcal{S}_0$. Moreover, the marginal measure λ^ℓ_I satisfies the first part of condition (4.1):

$$\lambda^\ell_I(B) = \lambda^\ell(B\times\Omega) = \mu^\ell_\Omega \in \mathbf{rca}(I).$$

Consider now functions $s(\eta,\omega) = \xi(\omega)$ with $\xi \in \mathcal{L}_1(\Omega,\mathcal{F},P)$. As ℓ is continuous, the functional $\xi \mapsto \ell(1\xi)$ must be continuous on $\mathcal{L}_1(\Omega,\mathcal{F},P)$. Since

$$\ell(1\xi) = \iint\limits_{I\times\Omega} \xi(\omega)\, d\lambda^\ell = \int_\Omega \xi(\omega)\, d\lambda^\ell_\Omega,$$

it is necessary that also the second part of (4.1) is satisfied by λ_Ω^ℓ. Thus $\mathcal{S}^* \subset \mathcal{S}_0^* \subset \mathcal{M}$. □

We can now formulate the *Lagrangian* $L : \mathcal{Z} \times \mathcal{S} \times \mathcal{M} \times \mathbf{rca}(I) \to \mathrm{R}$ of the optimization problem (2.1) as follows:

$$L(z,s,\lambda,\mu) = f(z) - \iint\limits_{I\times\Omega} \Big(\eta - [G(z)](\omega) - s(\eta,\omega)\Big)\, d\lambda + \int_I \left(v(\eta) - \int_\Omega s(\eta,\omega)\, dP\right) d\mu.$$

The corresponding *Lagrangian dual function* $L_D : \mathcal{M} \times \mathbf{rca}(I) \to \mathrm{R}$ is given by

$$\begin{aligned} L_D(\lambda,\mu) &= \sup_{s\ge 0,\ z\in Z} L(z,s,\lambda,\mu) \\ &= \sup_{s\ge 0,\ z\in Z} \left\{ f(z) + \iint\limits_{I\times\Omega} s(\eta,\omega)\, d(\lambda - \mu\times P) - \int_I \eta\, d\lambda_I + \int_\Omega [G(z)](\omega)\, d\lambda_\Omega + \int_I v(\omega)\, d\mu \right\}. \end{aligned}$$

By examining the second term of this expression we obtain

(4.3)
$$L_D(\lambda,\mu) = \begin{cases} -\displaystyle\int_I \eta\, d\lambda_I + \int_I v(\eta)\, d\mu + \sup_{z\in Z} \left\{ f(z) + \int_\Omega [G(z)](\omega)\, d\lambda_\Omega \right\} & \text{if } \lambda \le \mu \times P, \\ +\infty & \text{otherwise.} \end{cases}$$

This leads to the following dual problem:

$$\begin{aligned} &\underset{\lambda,\mu}{\text{minimize}} \ -\int_I \eta\, d\lambda_I + \int_I v(\omega)\, d\mu + \sup_{z\in Z} \left\{ f(z) + \int_\Omega [G(z)](\omega)\, d\lambda_\Omega \right\} \\ &\text{subject to } \lambda \le \mu \times P, \\ &\qquad\qquad \lambda \in \mathcal{M}^+, \quad \mu \in \mathbf{rca}^+(I). \end{aligned} \tag{4.4}$$

Here we use $\mathcal{M}^+$ and $\mathbf{rca}^+(I)$ to denote the sets of all nonnegative measures in $\mathcal{M}$ and $\mathbf{rca}(I)$, respectively.

THEOREM 4.2 (weak duality). *Let c_* and c_L denote the optimum values of the original problem* (2.1) *and the Lagrangian dual* (4.4)*, respectively. Then $c_* \le c_L$.*

Proof. Let (z,s) be feasible for (2.1). Then for every $(\lambda,\mu) \in \mathcal{M}^+ \times \mathbf{rca}^+(I)$ we have the inequalities

$$L_D(\lambda,\mu) \ge L(z,s,\lambda,\mu) \ge f(z).$$

Taking the infimum of the left-hand side with respect to (λ,μ) and the supremum of the right-hand side with respect to feasible (z,s), we obtain the assertion. □

In order to prove the strong duality relation we need a constraint qualification condition, introduced in [4, 5].

DEFINITION 4.3. *Problem* (2.1) *satisfies the* uniform dominance condition *if there exists $\tilde z \in Z$ such that*

$$\max_{\eta\in I} \Big\{ \mathrm{E}\big[(\eta - G(\tilde z))_+\big] - v(\eta) \Big\} < 0.$$

THEOREM 4.4 (strong duality). *Assume that problem* (2.1) *satisfies the uniform dominance condition and that it has an optimal solution. Then the dual problem* (4.4) *has an optimal solution and $c_* = c_L$.*

Proof. Due to Theorem 4.2, it is sufficient to find $(\hat{\lambda}, \hat{\mu}) \in \mathcal{M}^+ \times \mathbf{rca}^+(I)$ such that $L_D(\hat{\lambda}, \hat{\mu}) = c_*$.

Let $(\hat{z}, \hat{s})$ be an optimal solution of (2.1). Consider the equivalent problem formulation (1.4). Following [4] we can rewrite it in the abstract form:

$$\begin{aligned} &\underset{z}{\text{maximize}}\ f(z) \\ &\text{subject to } \Gamma(z) \in K, \\ &z \in Z, \end{aligned}$$

where $\Gamma : \mathcal{Z} \to \mathcal{C}(I)$ is a continuous operator defined as

$$[\Gamma(z)](\eta) = v(\eta) - \mathrm{E}([\eta - G(z)]_+), \quad \eta \in I.$$

The set K is the cone of nonnegative functions in $\mathcal{C}(I)$. Observe that the function $z \to \eta - G(z)$ is convex, for almost all $\omega \in \Omega$, and the function $x \to (x)_+$ is convex and nondecreasing. Therefore, the composition $\mathrm{E}[(\eta - G(z))_+]$ is a convex function of z. It follows that the operator Γ is concave with respect to the cone K; that is, for any z_1, z_2 in Z and all $\lambda \in [0, 1]$,

$$\Gamma(\lambda z_1 + (1 - \lambda) z_2) - [\lambda \Gamma(z_1) + (1 - \lambda)\Gamma(z_2)] \in K.$$

As the topological dual space to $\mathcal{C}(I)$ is $\mathbf{rca}(I)$, we can introduce the Lagrangian $\Lambda : \mathcal{Z} \times \mathbf{rca}(I) \to \mathrm{R}$,

$$\Lambda(z, \mu) = f(z) + \int_I [\Gamma(z)](\eta)\, d\mu. \tag{4.5}$$

Let us observe that the uniform dominance condition implies that the following generalized Slater condition is satisfied: There exists a point $\tilde{z} \in Z$ such that $\Gamma(\tilde{z}) \in \operatorname{int} K$. Therefore, we can use the necessary conditions of optimality in Banach spaces (see, e.g., [3, Theorem 3.4]). We conclude that there exists a measure $\hat{\mu} \in \mathbf{rca}^+(I)$ such that

$$\Lambda(\hat{z}, \hat{\mu}) = \max_{z \in Z} \Lambda(z, \hat{\mu}) \tag{4.6}$$

and

$$\int_I \big(v(\eta) - \mathrm{E}([\eta - G(\hat{z})]_+)\big)\, d\hat{\mu} = 0. \tag{4.7}$$

This means that $c_* = f(\hat{z}) = \Lambda(\hat{z}, \hat{\mu})$.

Define the set $U = \big\{(\beta, X, z) \in \mathrm{R} \times \mathcal{L}_1(\Omega, \mathcal{F}, P) \times Z : \beta \le f(z),\ X \le G(z)\big\}$. It follows from (4.6) that $\hat{\beta} = f(\hat{z})$, $\hat{X} = G(\hat{z})$, and $\hat{z}$ are the solution of the convex optimization problem

$$\underset{(\beta, X, z) \in U}{\text{maximize}}\ \beta - \iint_{I \times \Omega} [\eta - X]_+\, dP\, d\hat{\mu}. \tag{4.8}$$

Indeed, the best value of β is $f(z)$, and, due to the monotonicity of the function $x \to -[\eta - x]_+$, the best value of X is $G(z)$. By carrying out the partial maximization with respect to (β, X) we reduce (4.8) to the right-hand side of (4.6).

Consider the function $\varphi : \mathcal{S} \to \mathrm{R}$ defined by

$$\varphi(s) = \iint\limits_{I\times\Omega} \big[s(\eta,\omega)\big]_+ \, dP \, d\hat{\mu}$$

at the point $\hat{s}(\eta,\omega) = \eta - [G(\hat{z})](\omega)$, $\eta \in I$, $\omega \in \Omega$. By virtue of the necessary and sufficient condition of optimality for problem (4.8), there exists a subgradient $\gamma \in \partial\varphi(\hat{s})$ such that $(\hat{\beta}, \hat{X}, \hat{z})$ is also a solution of the problem

$$\underset{(\beta,X,z)\in U}{\text{maximize}} \ \beta + \iint\limits_{I\times\Omega} \gamma(\eta,\omega) X \, dP \, d\hat{\mu}. \tag{4.9}$$

By Strassen's disintegration theorem [33, Theorem 1],

$$\gamma(\eta,\omega) \in \partial\big(\eta - \hat{X}(\omega)\big)_+ = \begin{cases} \{1\} & \text{if } \eta > \hat{X}(\omega), \\ \{0\} & \text{if } \eta < \hat{X}(\omega), \\ [0,1] & \text{if } \eta = \hat{X}(\omega). \end{cases}$$

From the definition of the set U and from the fact that $\gamma(\eta,\omega) \ge 0$, for every value of z the best values of β and X in (4.9) are $f(z)$ and $G(z)$, respectively. It follows that $\hat{z}$ is an optimal solution of the problem

$$\underset{z\in Z}{\text{maximize}} \left\{ f(z) + \iint\limits_{I\times\Omega} \gamma(\eta,\omega) G(z) \, dP \, d\hat{\mu} \right\}.$$

Define the measure $\hat{\lambda}$ as absolutely continuous with respect to $\hat{\mu} \times P$ with the Radon–Nikodym derivative

$$\frac{d\hat{\lambda}}{d(\hat{\mu}\times P)} = \gamma.$$

Since $0 \le \gamma \le 1$, we have $0 \le \hat{\lambda} \le \hat{\mu} \times P$. From (4.3) we obtain

$$\begin{aligned} L_D(\hat{\lambda},\hat{\mu}) &= -\int_I \eta \, d\hat{\lambda}_I(\eta) + \int_I v(\eta) \, d\hat{\mu} + \sup_{z\in Z} \left\{ f(z) + \int_\Omega [G(z)](\omega) \, d\hat{\lambda}_\Omega \right\} \\ &= -\iint\limits_{I\times\Omega} \eta\gamma(\eta,\omega) \, dP \, d\hat{\mu} + \int_I v(\eta) \, d\hat{\mu} + f(\hat{z}) + \iint\limits_{I\times\Omega} [G(\hat{z})](\omega)\gamma(\eta,\omega) \, dP \, d\hat{\mu}. \end{aligned}$$

It follows from the definition of γ that the first and the last term in this expression can be written as

$$\begin{aligned} &-\iint\limits_{I\times\Omega} \eta\gamma(\eta,\omega) \, dP \, d\hat{\mu} + \iint\limits_{I\times\Omega} [G(\hat{z})](\omega)\gamma(\eta,\omega) \, dP \, d\hat{\mu} \\ &\qquad = -\iint\limits_{I\times\Omega} \big(\eta - [G(\hat{z})](\omega)\big)_+ \, dP \, d\hat{\mu} = -\int_I \mathrm{E}\big(\eta - [G(\hat{z})](\omega)\big)_+ \, d\hat{\mu}. \end{aligned}$$

Substituting into the last formula for $L_D(\hat{\lambda}, \hat{\mu})$ we conclude that

$$L_D(\hat{\lambda}, \hat{\mu}) = f(\hat{z}) + \int_I \Big[v(\eta) - \mathrm{E}\big(\eta - [G(\hat{z})](\omega)\big)_+ \Big] \, d\hat{\mu} = f(\hat{z}) = c_*.$$

In the last equation we have used the complementarity condition (4.7). □

Finally, let us observe that the condition $\lambda \leq \mu \times P$ appearing in the dual problem implies that $\lambda_I \leq \mu$. This is of importance for the solution method we describe later in section 7.

5. Reducing the space of Lagrange multipliers. Notice that apart from the condition $\lambda \leq \mu \times p$ the measure $\lambda \in \mathcal{M}$ on the product space $I \times \Omega$ appears in the dual optimization problem (4.4) via its marginal measures λ_I and λ_Ω. We can exploit this fact to achieve a reduction of the space of variables similar to that seen in the case of the primal problem. The main tool for this reduction is Strassen's theorem on the existence of measures with given marginals [33, Theorem 6]. We present here its version in the setting suitable for direct application to our problem.

THEOREM 5.1. *Let $\kappa \in \mathcal{M}^+$, $\beta \in \mathbf{rca}(I)$, and let α be a measure on $(\Omega, \mathcal{F})$. There exists a measure $\lambda \in \mathcal{M}^+$ having marginal measures $\lambda_I = \beta$ and $\lambda_\Omega = \alpha$ and such that $\lambda \leq \kappa$ if and only if*

$$\beta(B) + \alpha(A) \leq \psi + \kappa(B \times A) \quad \forall\, B \in \mathcal{B} \quad \forall\, A \in \mathcal{F},$$

where $\psi = \beta(I) = \alpha(\Omega)$.

Observe that setting $B = I$ we obtain $\alpha(A) \leq \kappa(B \times A)$ for all $A \in \mathcal{F}$. Employing the definition of $\mathcal{M}^+$ we conclude that it is necessary that $d\alpha/dP \in \mathcal{L}_\infty(\Omega, \mathcal{F}, P)$.

Applying Theorem 5.1 to the dual problem (4.4) with $\kappa = \mu \times P$, we obtain the following equivalent formulation of the dual problem:

$$\begin{aligned} \underset{\alpha,\beta,\mu,\psi}{\text{minimize}} \quad & -\int_I \eta \, d\beta + \int_I v \, d\mu + \sup_{z \in Z} \left(f(z) + \int_\Omega [G(z)](\omega) \, d\alpha \right) \\ \text{subject to} \quad & \beta(I) = \psi, \qquad \alpha(\Omega) = \psi, \\ & \beta(B) + \alpha(A) \leq \psi + \mu(B) P(A) \quad \forall\, B \in \mathcal{B} \quad \forall\, A \in \mathcal{F}, \\ & \alpha \geq 0, \quad \frac{d\alpha}{dP} \in \mathcal{L}_\infty(\Omega, \mathcal{F}, P), \quad \beta, \mu \in \mathbf{rca}^+(I). \end{aligned} \tag{5.1}$$

Note that the measure λ on the product space $I \times \Omega$ is eliminated from this formulation, at the cost of introducing new constraints indexed by the family $\mathcal{B} \times \mathcal{F}$. The merits of this trade-off become apparent for problems with discrete distributions, where we propose a column generation method.

6. An implied transportation problem. We now focus again on the finite probability space $\Omega = \{\omega_1, \ldots, \omega_N\}$ with corresponding probabilities $p_1, \ldots, p_N$. The realizations of the benchmark outcome Y are denoted by $y_1, \ldots, y_D$, and the corresponding benchmark shortfalls are $v_j = \mathrm{E}([y_j - Y]_+)$.

We recall for convenience the dual problem (4.4) in this case. The measure λ becomes an array λ_{ij}, $i = 1, \ldots, N$, $j = 1, \ldots, D$. The marginal measures are its row and column sums, respectively. We obtain the following formulation:

$$\begin{aligned} \underset{\lambda,\mu}{\text{minimize}} \quad & -\sum_{j=1}^{D} \sum_{i=1}^{N} \lambda_{ij} y_j + \sum_{j=1}^{D} \mu_j v_j + \sup_{z \in Z} \left\{ f(z) + \sum_{i=1}^{N} \sum_{j=1}^{D} \lambda_{ij} G_i(z) \right\} \\ \text{subject to} \quad & \lambda_{ij} \leq p_i \mu_j, \quad i = 1, \ldots, N, \quad j = 1, \ldots, D, \\ & \lambda \geq 0, \quad \mu \geq 0. \end{aligned} \tag{6.1}$$

Consider the marginal sums

$$\alpha_i = \sum_{j=1}^{D} \lambda_{ij}, \quad i = 1, \dots, N,$$
$$\beta_j = \sum_{i=1}^{N} \lambda_{ij}, \quad j = 1, \dots, D.$$

Vectors $\alpha \geq 0$ and $\beta \geq 0$ are marginal sums of a feasible dual variable λ if and only if the following conditions are satisfied:

(i) for some $\psi \geq 0$ we have $\sum_{i=1}^{N} \alpha_i = \sum_{j=1}^{D} \beta_j = \psi$;

(ii) there exists a transportation flow of value ψ in the network having N source nodes with supplies α, D destination nodes with demands β, and arc capacities equal to $p_i \mu_j$ for every arc (i, j), $i = 1, \dots, N$, $j = 1, \dots, D$.

In the discrete case Strassen's theorem takes on the form of the *maximum flow–minimum cut theorem* [11, 13]: For all $A \subset \{1, \dots, N\}$ and all $B \subset \{1, \dots, D\}$

$$\sum_{i \in A} \alpha_i + \sum_{j \in B} \beta_j - \sum_{i \in A} p_i \sum_{j \in B} \mu_j \leq \psi. \tag{6.2}$$

The dual formulation based on this fact and corresponding to (5.1) is

$$\begin{aligned} \underset{\alpha, \beta, \mu, \psi}{\text{minimize}} \quad & -\sum_{j=1}^{D} y_j \beta_j + \sum_{j=1}^{D} v_j \mu_j + \sup_{z \in Z} \left\{ f(z) + \sum_{i=1}^{N} \alpha_i G_i(z) \right\} \\ \text{subject to} \quad & \sum_{i=1}^{N} \alpha_i = \psi, \qquad \sum_{j=1}^{D} \beta_j = \psi, \\ & \sum_{i \in A} \alpha_i + \sum_{j \in B} \beta_j - \sum_{i \in A} p_i \sum_{j \in B} \mu_j \leq \psi \\ & \qquad \forall\, A \subset \{1, \dots, N\} \quad \forall\, B \subset \{1, \dots, D\} \\ & \alpha \geq 0, \quad \beta \geq 0, \quad \mu \geq 0, \quad \psi \geq 0. \end{aligned} \tag{6.3}$$

In this formulation ND dual variables of (6.1) are replaced by $N + D$ marginal sums, at the cost of introducing 2^{N+D} new constraints, indexed by all possible sets A and B.

If the functions $f(\cdot)$ and $G_i(\cdot)$ are affine and the set Z is defined by linear equations and inequalities, then problem (6.3) becomes a linear programming problem. In a standard way, the term $\sup_{z \in Z} \{f(z) + \sum_{i=1}^{N} \alpha_i G_i(z)\}$ can be replaced by an affine function of α and linear inequalities involving α. All of these manipulations are the same as in linear programming duality theory. We illustrate them for a portfolio example in section 8.

The main difficulty associated with problem (6.3) is the large number of constraints. We show in section 7 a way to overcome this difficulty by generating only a subset of relevant constraints.

7. A dual column generation method. Formulation (6.3) suggests a cutting plane method of the following form: At iteration k we have pairs of sets $A_m \subset \{1, \dots, N\}$ and $B_m \subset \{1, \dots, D\}$, $m = 1, \dots, k-1$. We solve a relaxation of problem

(6.3):

$$\begin{aligned}
&\underset{\alpha,\beta,\mu,\psi}{\text{minimize}} \quad -\sum_{j=1}^{D} y_j\beta_j + \sum_{j=1}^{D} v_j\mu_j + \sup_{z\in Z}\left\{ f(z) + \sum_{i=1}^{N} \alpha_i G_i(z)\right\} \\
&\text{subject to} \quad \sum_{i=1}^{N} \alpha_i = \psi, \qquad \sum_{j=1}^{D} \beta_j = \psi, \\
&\qquad \sum_{i\in A_m} \alpha_i + \sum_{j\in B_m} \beta_j - \sum_{i\in A_m} p_i \sum_{j\in B_m} \mu_j \le \psi, \quad m = 1,\dots,k-1, \\
&\qquad \alpha \ge 0, \quad \beta \ge 0, \quad \mu \ge 0, \quad \psi \ge 0.
\end{aligned} \tag{7.1}$$

The next step is to verify inequalities (6.2) for all possible sets A and B at the optimal solution $(\alpha^k, \beta^k, \mu^k, \psi^k)$. To this end, we find a pair A_k, B_k which solves the problem

$$\underset{\substack{A\subset\{1,\dots,N\}\\ B\subset\{1,\dots,D\}}}{\text{maximize}} \quad -\psi^k + \sum_{i\in A} \alpha_i^k + \sum_{j\in B} \beta_j^k - \sum_{i\in A} p_i \sum_{j\in B} \mu_j^k. \tag{7.2}$$

Defining the complement event $A^c = \{1,\dots,N\} \setminus A$ we observe that the first three terms in (7.2) describe the required inflow to the set of nodes $A^c \cup B$. The last term in (7.2) is the total capacity of the arcs leading to this set, that is, the arcs starting in A and ending in B. It follows that problem (7.2) is a problem of finding a minimal cut in a bipartite graph. It can be solved in a very efficient way by special network algorithms, as described in [1]. One method, which is closely related to our transformation, is the following. We formulate the maximum flow problem:

$$\begin{aligned}
&\underset{\lambda}{\text{maximize}} \quad \sum_{i=1}^{N}\sum_{j=1}^{D} \lambda_{ij} \\
&\text{subject to} \quad \sum_{j=1}^{D} \lambda_{ij} \le \alpha_i^k, \quad i = 1,\dots,N, \qquad (7.3)\\
&\qquad \sum_{i=1}^{N} \lambda_{ij} \le \beta_j^k, \quad j = 1,\dots,D, \qquad (7.4)\\
&\qquad 0 \le \lambda_{ij} \le p_i\mu_j^k, \quad i = 1,\dots,N, \quad j = 1,\dots,D.
\end{aligned}$$

If the flow equals ψ^k, then the optimal solution of (7.1) is also optimal for (6.3). Otherwise, we denote the Lagrange multipliers associated with (7.3) by ζ_i, $i = 1,\dots,N$, and the Lagrange multipliers associated with (7.4) by ξ_j, $j = 1,\dots,D$ (they all are equal to either 0 or 1). We set

$$A_k = \{i : \zeta_i = 0\}, \qquad B_k = \{j : \xi_j = 0\},$$

and we add the pair (A_k, B_k) to the pairs of sets included in (7.1), increase k by 1, and continue.

Observe that if the maximum in (7.2) is positive (and thus the maximum flow in the last displayed problem is smaller than ψ^k), the new cut is different than the cuts already included in problem (7.1). As the number of possible cuts is finite, the method must eventually stop at an optimal solution. In that case the flow in the network gives us the optimal values of the multipliers λ in the dual problem (6.1).

8. Numerical illustration. Let $R_1, \dots, R_n$ be random return rates of assets $1, \dots, n$. We denote the fractions of the initial capital invested in these assets by $z_1, \dots, z_n$. Clearly, the portfolio return rate equals

$$G(z) = R_1 z_1 + \cdots + R_n z_n.$$

The set of possible asset allocations is the simplex

$$Z = \{z \in \mathrm{R}^n : z_1 + \cdots + z_n = 1,\ z_k \geq 0,\ k = 1, \dots, n\},$$

but the approach outlined here easily extends to more general polyhedral sets Z. Finally, let a benchmark random return rate Y be given; for example, Y may represent the return rate of an index or the return rate of the current portfolio. The dominance-constrained portfolio optimization problem takes on the form

$$\begin{aligned} \underset{z}{\text{maximize}}\ & \mathrm{E}\big[R_1 z_1 + \cdots + R_n z_n\big] \\ \text{subject to}\ & R_1 z_1 + \cdots + R_n z_n \succeq_{(2)} Y, \\ & z \in Z. \end{aligned}$$

This model was introduced as an example in [4] and analyzed in [8].

As discussed in the introduction, no risk-averse decision maker will prefer a portfolio with return rate Y over a portfolio with return rate $R_1 z_1 + \cdots + R_n z_n$. Therefore, the risk exposure of the portfolio return rate is "more acceptable" than that of Y. In our model we use the expected value of the portfolio return rate as the objective functional, and thus the entire burden of controlling risk is carried by the stochastic dominance constraint. As the distribution of the returns at the solution is indirectly shaped by the distribution of the benchmark Y, it is essential that Y be "acceptable," for example, the return rate of a broad market index. However, it is easy to additionally incorporate risk-averse objective functionals to our model, such as coherent measures of risk (see [2, 9, 10, 20, 31] and the references therein).

In the discrete distribution case, we denote the return of asset k in event i by r_{ik}, $i = 1, \dots, N$, $k = 1, \dots, n$, the probabilities of the elementary events by p_i, $i = 1, \dots, N$, the realizations of the benchmark returns by y_j, and the benchmark shortfalls by

$$v_j = \sum_{i=1}^{N} p_i (y_j - y_i)_+.$$

We obtain the following problem:

$$\begin{aligned} \underset{z}{\text{maximize}}\quad & \sum_{i=1}^{N} \sum_{k=1}^{n} p_i r_{ik} z_k \\ \text{subject to}\quad & \sum_{i=1}^{N} p_i \left(y_j - \sum_{k=1}^{n} r_{ik} z_k \right)_+ \leq v_j, \quad j = 1, \dots, N, \\ & z \in Z. \end{aligned}$$

The piecewise linear constraint is dealt with by the primal cutting plane method.

TABLE 8.1
Dimensions of the three formulations.

Scenarios	Linear Programming		Primal Formulation		Dual Formulation	
	Variables	Constraints	Variables	Constraints	Variables	Constraints
50	3000	2551	500	1	152	502
100	10500	10101	500	1	302	502
150	23000	22651	500	1	452	502
200	40500	40201	500	1	602	502
500	250500	250001	500	1	1502	502
750	563000	562501	500	1	2252	502
1000	1000500	1000001	500	1	3002	502

The dual problem (6.3) takes on the following form:

$$
\begin{aligned}
\underset{\alpha,\beta,\mu,\psi,\zeta}{\text{minimize}} \quad & -\sum_{j=1}^{N} y_j\beta_j + \sum_{j=1}^{N} v_j\mu_j + \zeta \\
\text{subject to} \quad & \sum_{i=1}^{N} \alpha_i = \psi, \qquad \sum_{j=1}^{N} \beta_j = \psi, \\
& \sum_{i=1}^{N} (p_i + \alpha_i) r_{ik} \le \zeta, \quad k = 1, \ldots, n, \\
& \sum_{i\in A} \alpha_i + \sum_{j\in B} \beta_j - \sum_{i\in A} p_i \sum_{j\in B} \mu_j \le \psi \quad \forall\, A \subset \{1, \ldots, N\} \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall\, B \subset \{1, \ldots, N\} \\
& \alpha \ge 0, \quad \beta \ge 0, \quad \mu \ge 0, \quad \psi \ge 0.
\end{aligned}
$$

The variable ζ in the objective function of this problem represents the term

$$
\begin{aligned}
\sup_{z\in Z} \left\{ f(z) + \sum_{i=1}^{N} \alpha_i G_i(z) \right\} &= \sup_{z\in Z} \left\{ \sum_{i=1}^{N} \sum_{k=1}^{n} p_i r_{ik} z_k + \sum_{i=1}^{N} \sum_{k=1}^{n} \alpha_i r_{ik} z_k \right\} \\
&= \max_{1\le k\le n} \sum_{i=1}^{N} (p_i + \alpha_i) r_{ik}.
\end{aligned}
$$

The constraints involving the sets A and B are dealt with by the dual cutting plane method.

We considered several problem instances of different sizes, obtained from historical data on realizations of joint daily returns of $n = 500$ assets in N days, for seven different values of N ranging from 50 to 1000. We used the returns in each day as equally probable realizations of the n-dimensional random vector R. The benchmark outcome Y was the return rate of the Standard & Poors 500 index. All calculations were carried out on a 2.00 GHz Pentium 4 PC with 1.00 GB of RAM by using the AMPL modeling language [12] and with version 9.1 of the CPLEX solver [15].

Table 8.1 compares the sizes of the three formulations: The straightforward linear programming model (2.1), the primal cutting plane formulation (2.2), and the dual cutting plane formulation (5.1). In the last two cases we report the initial numbers of constraints only, without the cuts indexed by the sets $A \in \mathcal{F}$ and $B \in \mathcal{B}$. The numbers of cuts, which were actually generated in the course of the solution, are reported in

TABLE 8.2
Performance of the three approaches.

Scenarios	Linear Programming		Primal Method			Dual Method		
	CPU	Iterations	CPU	Cuts	Iterations	CPU	Cuts	Iterations
50	3.44	570	0.55	9	9	13.81	68	6883
100	20.23	3161	2.03	33	75	407.26	259	156304
150	372.52	7272	3.49	53	267	9144.25	552	1155166
200	373.63	16666	3.90	61	180	-	-	-
500	-	-	6.59	88	924	-	-	-
750	-	-	9.74	123	477	-	-	-
1000	-	-	10.23	117	530	-	-	-

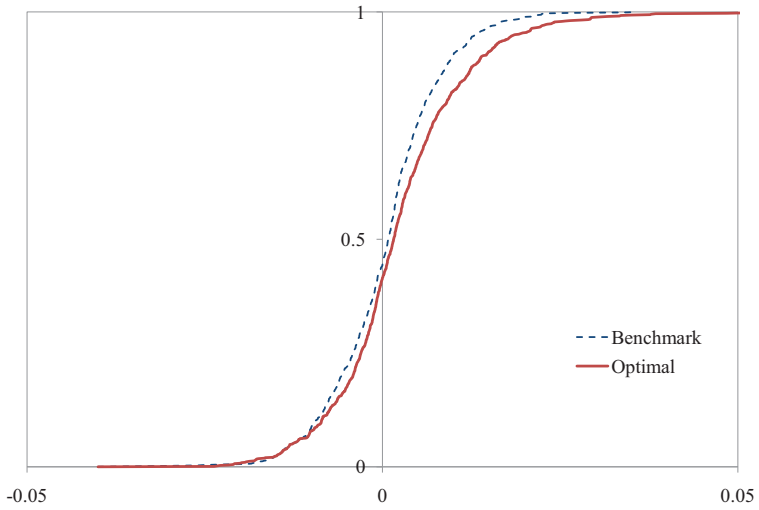


FIG. 8.1. *Cumulative distribution functions of the return rates of the benchmark and optimal portfolios in the* 1000 *scenario example.*

Table 8.2. This table provides also the CPU times of the simplex solver in the three cases and the total numbers of simplex iterations performed.

It can be seen from these results that the primal cut generation method is quite efficient, and it dramatically outperforms the direct linear programming approach. This is consistent with the results of [17] for integrated chance constraints. In fact, the direct linear programming model was too large for our computer for 500 scenarios and more. The dual method is much slower for this problem class, mainly due to minimal differences between many cuts and severe numerical difficulties associated with that. For problems with $N = 200$ scenarios and more, we interrupted the calculation because of excessive time. Apparently, the number of Strassen cuts is too large. However, we still believe that the dual formulation is interesting in its own right and that one day it may find its application.

Finally, Figure 8.1 compares the cumulative distribution functions of the return rates of the benchmark portfolio (the S&P 500 index) and of the solution to the dominance-constrained problem for the case of 1000 scenarios. The optimal portfolio contains only 11 assets, but we can see that they are sufficient to shape the distribution function in a favorable way. Close inspection reveals that the optimal distribution function is not entirely below the benchmark (this would mean first order stochastic

dominance); in the range between -0.02 and -0.015 it is slightly above. However, the expected shortfall (1.2) is always smaller at the solution than at the benchmark. This is in line with the results of [23], where similar examples are presented.

Acknowledgments. The authors are grateful to two anonymous referees for their insightful comments which helped to improve the presentation of this paper.

REFERENCES

[1] R.K. Ahuja, T.L. Magnanti, and J.B. Orlin, *Network Flows. Theory, Algorithms, and Applications*, Prentice–Hall, Englewood Cliffs, NJ, 1993.
[2] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, *Coherent measures of risk*, Math. Finance, 9 (1999), pp. 203–228.
[3] J.F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer-Verlag, New York, 2000.
[4] D. Dentcheva and A. Ruszczyński, *Optimization with stochastic dominance constraints*, SIAM J. Optim., 14 (2003), pp. 548–566.
[5] D. Dentcheva and A. Ruszczyński, *Optimality and duality theory for stochastic optimization problems with nonlinear dominance constraints*, Math. Program., 99 (2004), pp. 329–350.
[6] D. Dentcheva and A. Ruszczyński, *Semi-infinite probabilistic optimization: First order stochastic dominance constraints*, Optim., 53 (2004), pp. 583–601.
[7] D. Dentcheva and A. Ruszczyński, *Inverse stochastic dominance constraints and rank dependent expected utility theory*, Math. Program., 108 (2006), pp. 297–311.
[8] D. Dentcheva and A. Ruszczyński, *Portfolio optimization with stochastic dominance constraints*, J. Banking Finance, 30 (2006), pp. 433–451.
[9] H. Föllmer and A. Schied, *Convex measures of risk and trading constraints*, Finance Stoch., 6 (2002), pp. 429–447.
[10] H. Föllmer and A. Schied, *Stochastic Finance. An Introduction in Discrete Time*, Walter de Gruyter, Berlin, 2004.
[11] L.R. Ford and D.R. Fulkerson, *Maximal flow through a network*, Canad. J. Math., 8 (1956), pp. 399–404.
[12] R. Fourer, D.M. Gay, and B.W. Kernighan, *AMPL: A Modeling Language for Mathematical Programming*, Duxbury Press, Belmont, CA, 2002.
[13] D. Gale, *A theorem on flows in networks*, Pacific J. Math., 7 (1957), pp. 1073–1082.
[14] J. Hadar and W. Russell, *Rules for ordering uncertain prospects*, Amer. Econ. Rev., 59 (1969), pp. 25–34.
[15] *ILOG CPLEX* 9.0 *Users Manual*, ILOG, Incline Village, NV, 2005.
[16] W.K. Klein Haneveld, *Duality in Stochastic Linear and Dynamic Programming*, Lecture Notes in Econom. Math. Systems 274, Springer-Verlag, New York, 1986.
[17] W.K. Klein Haneveld and M.H. Van Der Vlerk, *Integrated chance constraints: Reduced forms and an algorithm*, Comput. Manag. Sci., 3 (2006), pp. 245–269.
[18] E. Lehmann, *Ordered families of distributions*, Ann. Math. Statist., 26 (1955), pp. 399–419.
[19] H.B. Mann and D.R. Whitney, *On a test of whether one of two random variables is stochastically larger than the other*, Ann. Math. Statist., 18 (1947), pp. 50–60.
[20] N. Miller and A. Ruszczyński, *Risk-adjusted probability measures in portfolio optimization with coherent measures of risk*, Eur. J. Oper. Res., 191 (2008), pp. 193–206.
[21] A. Müller and D. Stoyan, *Comparison Methods for Stochastic Models and Risks*, John Wiley & Sons, Chichester, 2002.
[22] N. Noyan, G. Rudolf, and A. Ruszczyński, *Relaxations of linear programming problems with first order stochastic dominance constraints*, Oper. Res. Lett., 34 (2006), pp. 653–659.
[23] N. Noyan and A. Ruszczyński, *Valid inequalities and restrictions for stochastic programming problems with first order stochastic dominance constraints*, Math. Program., 114 (2008), pp. 249–275.
[24] W. Ogryczak and A. Ruszczyński, *From stochastic dominance to mean–risk models: Semideviations as risk measures*, Eur. J. Oper. Res., 116 (1999), pp. 33–50.
[25] W. Ogryczak and A. Ruszczyński, *On consistency of stochastic dominance and mean–semideviation models*, Math. Program., 89 (2001), pp. 217–232.
[26] W. Ogryczak and A. Ruszczyński, *Dual stochastic dominance and related mean–risk models*, SIAM J. Optim., 13 (2002), pp. 60–78.
[27] J.P. Quirk and R. Saposnik, *Admissibility and measurable utility functions*, Rev. Econ. Stud., 29 (1962), pp. 140–146.

[28] R.T. Rockafellar, *Conjugate Duality and Optimization*, SIAM, Philadelphia, 1974.
[29] R.T. Rockafellar and S. Uryasev, *Conditional value-at-risk for general loss distributions*, J. Banking Finance, 26 (2002), pp. 1443–1471.
[30] A. Ruszczynski and A. Shapiro, eds., *Stochastic Programming*, Handbooks Oper. Res. Management Sci. 10, Elsevier, Amsterdam, 2003, North Holland.
[31] A. Ruszczyński and A. Shapiro, *Optimization of convex risk functions*, Math. Oper. Res., 31 (2006), pp. 433–452.
[32] M. Shaked and J.G. Shanthikumar, *Stochastic Orders and Their Applications*, Academic Press, Boston, 1994.
[33] V. Strassen, *The existence of probability measures with given marginals*, Ann. Math. Statist., 38 (1965), pp. 423–439.
[34] M.E. Yaari, *The dual theory of choice under risk*, Econometrica, 55 (1987), pp. 95–115.

SIAM J. OPTIM.
Vol. 19, No. 3, pp. 1344–1367

A MINIMAX THEOREM WITH APPLICATIONS TO MACHINE LEARNING, SIGNAL PROCESSING, AND FINANCE*

SEUNG-JEAN KIM† AND STEPHEN BOYD†

Abstract. This paper concerns a fractional function of the form $x^T a/\sqrt{x^T Bx}$, where B is positive definite. We consider the game of choosing x from a convex set, to maximize the function, and choosing (a, B) from a convex set, to minimize it. We prove the existence of a saddle point and describe an efficient method, based on convex optimization, for computing it. We describe applications in machine learning (robust Fisher linear discriminant analysis), signal processing (robust beamforming and robust matched filtering), and finance (robust portfolio selection). In these applications, x corresponds to some design variables to be chosen, and the pair (a, B) corresponds to the statistical model, which is uncertain.

Key words. convex optimization, minimax theorem, robust optimization

AMS subject classifications. 65K0, 90C25, 90C47, 91B28

DOI. 10.1137/060677586

1. Introduction.

This paper concerns a fractional function of the form

$$f(x, a, B) = \frac{x^T a}{\sqrt{x^T Bx}}, \tag{1}$$

where $x, a \in \mathbb{R}^n$ and $B = B^T \in \mathbb{R}^{n\times n}$. We assume that $x \in \mathcal{X} \subseteq \mathbb{R}^n\backslash\{0\}$ and $(a, B) \in \mathcal{U} \subseteq \mathbb{R}^n \times \mathbb{S}^n_{++}$. Here $\mathbb{S}^n_{++}$ denotes the set of $n \times n$ symmetric positive definite matrices.

We list some of the basic properties of the function f. It is (positive) homogeneous (of degree zero) in x: for all $t > 0$,

$$f(tx, a, B) = f(x, a, B).$$

If

$$x^T a \geq 0 \text{ for all } x \in \mathcal{X} \text{ and for all } a, \text{ with } (a, B) \in \mathcal{U}, \tag{2}$$

then for fixed $(a, B) \in \mathcal{U}$, f is quasi-concave in x, and for fixed $x \in \mathcal{X}$, f is quasi-convex in (a, B). This can be seen as follows: for $\gamma \geq 0$, the set

$$\{x \mid f(a, B, x) \geq \gamma\} = \left\{x \;\middle|\; \gamma\sqrt{x^T Bx} \leq x^T a\right\}$$

is convex (since it is a second-order cone in $\mathbb{R}^n$), and the set

$$\{(a, B) \mid f(a, B, x) \leq \gamma\} = \left\{(a, B) \;\middle|\; \gamma\sqrt{x^T Bx} \geq x^T a\right\}$$

is convex (since $\sqrt{x^T Bx}$ is concave in B).

*Received by the editors December 12, 2006; accepted for publication (in revised form) May 6, 2008; published electronically November 21, 2008. This work was funded in part by the Precourt Institute on Energy Efficiency, Army award W911NF-07-1-0029, NSF award 0529426, NASA award NNX07AEIIA, AFOSR award FA9550-06-1-0514, and AFOSR award FA9550-06-1-0312.
http://www.siam.org/journals/siopt/19-3/67758.html

†Information Systems Laboratory, Electrical Engineering Department, Stanford University, Stanford, CA 94305-9510 (sjkim@stanford.edu, boyd@stanford.edu).

A zero-sum game and related problems. In this paper we consider the zero-sum game of choosing x from a convex set $\mathcal{X}$, to maximize the function, and choosing (a, B) from a convex compact set $\mathcal{U}$, to minimize it. The game is associated with the following two problems:

- max-min problem

$$\begin{array}{ll} \text{maximize} & \inf\limits_{(a,B)\in\mathcal{U}} f(x,a,B) \\ \text{subject to} & x \in \mathcal{X}, \end{array} \tag{3}$$

  with variables $x \in \mathbb{R}^n$,
- min-max problem

$$\begin{array}{ll} \text{minimize} & \sup\limits_{x\in\mathcal{X}} f(x,a,B) \\ \text{subject to} & (a,B) \in \mathcal{U}, \end{array} \tag{4}$$

  with variables $a \in \mathbb{R}^n$ and $B = B^T \in \mathbb{R}^{n\times n}$.

Problems of the form (3) arise in several disciplines including machine learning (robust Fisher linear discriminant analysis), signal processing (robust beamforming and robust matched filtering), and finance (robust portfolio selection). In these applications, x corresponds to some design variables to be chosen, and the pair (a, B) corresponds to the first and second moments of a random vector, say, $\mathbf{Z}$, which are uncertain. We want to choose x so that the combined random variable $x^T\mathbf{Z}$ is well separated from zero. The ratio of the mean of the random variable to the standard deviation $f(x, a, B)$ measures the extent to which the random variable can be well separated from zero. The max-min problem is to find the design variables that are optimal in a worst-case sense, where worst-case means over all possible statistics. The min-max problem is to find the least-favorable statistical model, with the design variables chosen optimally for the statistics.

Minimax properties. The minimax inequality or *weak minimax property*

$$\sup_{x\in\mathcal{X}} \inf_{(a,B)\in\mathcal{U}} f(x,a,B) \le \inf_{(a,B)\in\mathcal{U}} \sup_{x\in\mathcal{X}} f(x,a,B) \tag{5}$$

always holds for any $\mathcal{X} \subseteq \mathbb{R}$ and any $\mathcal{U} \subseteq \mathbb{S}^n_{++}$. The minimax equality or *strong minimax property*

$$\sup_{x\in\mathcal{X}} \inf_{(a,B)\in\mathcal{U}} f(x,a,B) = \inf_{(a,B)\in\mathcal{U}} \sup_{x\in\mathcal{X}} f(x,a,B) \tag{6}$$

holds if $\mathcal{X}$ is convex, $\mathcal{U}$ is convex and compact, and (2) holds, which follows from Sion's quasi-convex–quasi-concave minimax theorem [25].

In this paper we will show that the strong minimax property holds with a weaker assumption than (2):

$$\text{there exists } \bar{x} \in \mathcal{X} \text{ such that } \bar{x}^T a > 0 \text{ for all } a \text{ with } (a, B) \in \mathcal{U}. \tag{7}$$

To state the minimax result, we first describe an equivalent formulation of the min-max problem (4).

PROPOSITION 1. *Suppose that $\mathcal{X}$ is a cone in $\mathbb{R}^n$ that does not contain the origin, with $\mathcal{X} \cup \{0\}$ convex and closed, and $\mathcal{U}$ is a compact subset of $\mathbb{R}^n \times \mathbb{S}^n_{++}$. Suppose further that* (7) *holds. Then, the min-max problem* (4) *is equivalent to*

$$\begin{array}{ll} \textit{minimize} & (a+\lambda)^T B^{-1}(a+\lambda) \\ \textit{subject to} & (a,B) \in \mathcal{U}, \quad \lambda \in \mathcal{X}^*, \end{array} \tag{8}$$

where $a \in \mathbb{R}^n$, $B = B^T \in \mathbb{R}^{n\times n}$, and $\lambda \in \mathbb{R}^n$ are the variables and $\mathcal{X}^*$ is the dual

cone of $\mathcal{X}$ *given by*

$$\mathcal{X}^* = \left\{\lambda \in \mathbb{R}^n \mid \lambda^T x \geq 0 \ \forall x \in \mathcal{X}\right\},$$

in the following sense: if $(a^\star, B^\star, \lambda^\star)$ *solves* (8)*, then* $(a^\star, B^\star)$ *solves* (4)*, and conversely if* $(a^\star, B^\star)$ *solves* (4)*, then there exists* $\lambda^\star \in \mathcal{X}^*$ *such that* $(a^\star, B^\star, \lambda^\star)$ *solves* (8)*. Moreover,*

$$\inf_{(a,B)\in\mathcal{U}} \sup_{x\in\mathcal{X}} f(x,a,B) = \left(\inf_{(a,B)\in\mathcal{U},\ \lambda\in\mathcal{X}^*} (a+\lambda)^T B^{-1}(a+\lambda)\right)^{1/2}.$$

Finally, (8) *always has a solution, and for any solution* $(a^\star, B^\star, \lambda^\star)$*,*

$$a^\star + \lambda^\star \neq 0.$$

The proof is deferred to the appendix.

The dual cone $\mathcal{X}^*$ is always convex. The objective of (8) is convex since a function of the form $f(x,X) = x^T X^{-1} x$, called a matrix fractional function, is convex over $\mathbb{R}^n \times \mathbb{S}^n_{++}$; see, e.g., [7, section 3.1.7]. Therefore, (8) is a convex problem. We conclude that the min-max problem (4) can be reformulated as the convex problem (8).

We can solve the max-min problem (3), using a minimax result for the fractional function $f(x,a,B)$.

THEOREM 1. *Suppose that* $\mathcal{X}$ *is a cone in* $\mathbb{R}^n$ *that does not contain the origin, with* $\mathcal{X} \cup \{0\}$ *convex and closed, and* $\mathcal{U}$ *is a convex compact subset of* $\mathbb{R}^n \times \mathbb{S}^n_{++}$*. Suppose further that* (7) *holds. Let* $(a^\star, B^\star, \lambda^\star)$ *be a solution to the convex problem* (8) *(whose existence is guaranteed in Proposition* 1*). Then,*

$$x^\star = B^{\star -1}(a^\star + \lambda^\star) \in \mathcal{X},$$

and the triple $(x^\star, a^\star, B^\star)$ *satisfies the* saddle-point property

$$f(x,a^\star,B^\star) \leq f(x^\star,a^\star,B^\star) \leq f(x^\star,a,B) \quad \forall x \in \mathcal{X} \quad \forall (a,B) \in \mathcal{U}. \tag{9}$$

The proof is deferred to the appendix.

We show that the assumption (7) is needed for the strong minimax property to hold. Consider $\mathcal{X} = \mathbb{R}^n \backslash \{0\}$ and $\mathcal{U} = \mathcal{B}_1 \times \{I\}$, where $\mathcal{B}_1$ is the Euclidean ball of radius one. Then, all of the assumptions hold except for (7). We have

$$\sup_{x\neq 0} \inf_{(a,B)\in\mathcal{U}} \frac{x^T a}{\sqrt{x^T B x}} = \sup_{x\neq 0} \inf_{a\in\mathcal{B}_1} \frac{x^T a}{\sqrt{x^T x}} = \sup_{x\neq 0} \frac{-\|x\|}{\|x\|} = -1$$

and

$$\inf_{(a,B)\in\mathcal{U}} \sup_{x\neq 0} \frac{x^T a}{\sqrt{x^T B x}} = \inf_{a\in\mathcal{B}_1} \sup_{x\neq 0} \frac{x^T a}{\|x\|} = \inf_{a\in\mathcal{B}_1} \frac{\|x\|\|a\|}{\|x\|} = 0.$$

From a standard result [2, section 2.6] in minimax theory, the saddle-point property (9) means that

$$\begin{aligned} f(x^\star,a^\star,B^\star) &= \sup_{x\in\mathcal{X}} f(x,a^\star,B^\star) \\ &= \inf_{(a,B)\in\mathcal{U}} f(x^\star,a,B) \\ &= \sup_{x\in\mathcal{X}} \inf_{(a,B)\in\mathcal{U}} f(x,a,B) \\ &= \inf_{(a,B)\in\mathcal{U}} \sup_{x\in\mathcal{X}} f(x,a,B). \end{aligned}$$

As a consequence, $x^\star$ solves (3).

More computational results. The max-min problem (3) has a unique solution up to (positive) scaling.

PROPOSITION 2. *Under the assumptions of Theorem* 1, *the max-min problem* (3) *has a unique solution up to (positive) scaling, meaning that for any two solutions* $x^\star$ *and* $y^\star$, *there is a positive number* $\alpha > 0$ *such that* $x^\star = \alpha y^\star$.

The proof is deferred to the appendix.

The convex problem (8) can be reformulated as a standard convex optimization problem. Using the Schur complement technique [7, Appendix 5.5], we can see that

$$(a+\lambda)^T B^{-1}(a+\lambda) \le t$$

if and only if the linear matrix inequality (LMI)

$$\begin{bmatrix} t & (a+\lambda)^T \\ a+\lambda & B \end{bmatrix} \succeq 0$$

holds. (Here $A \succeq 0$ means that A is positive semidefinite.) The convex problem (8) is therefore equivalent to

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & (a,B) \in \mathcal{U}, \quad \lambda \in \mathcal{X}^*, \quad \begin{bmatrix} t & (a+\lambda)^T \\ a+\lambda & B \end{bmatrix} \succeq 0, \end{array}$$

where the variables are $t \in \mathbb{R}$, $a \in \mathbb{R}^n$, $B = B^T \in \mathbb{R}^{n\times n}$, and $\lambda \in \mathbb{R}^n$. When the uncertainty sets $\mathcal{U}$ can be represented by LMIs, this problem is a semidefinite program (SDP). (Several high-quality open-source solvers for SDPs are available, e.g., SeDuMi [26], SDPT3 [27], and DSDP5 [1].) The reader is referred to [6, 29] for more on semidefinite programming and LMIs.

Outline of the paper. In the next section, we give a probabilistic interpretation of the saddle-point property established above. In sections 3–5, we give the applications of the minimax result in machine learning, signal processing, and portfolio selection. We give our conclusions in section 6. The appendix contains the proofs that are omitted from the main text.

2. A probabilistic interpretation.

2.1. Probabilistic linear separation. Suppose $z \sim \mathcal{N}(a,B)$ and $x \in \mathbb{R}^n$. Here, we use $\mathcal{N}(a,B)$ to denote the Gaussian distribution with mean a and covariance B. Then, $x^T z \sim \mathcal{N}(x^T a, x^T B x)$, so

$$\textbf{Prob}\left(x^T z \ge 0\right) = \Phi\left(\frac{x^T a}{\sqrt{x^T B x}}\right), \tag{10}$$

where Φ is the cumulative distribution function of the standard normal distribution.

Theorem 1 with $\mathcal{U} = \{(a,B)\}$ tells us that the right-hand side of (10) is maximized (over $x \in \mathcal{X}$) by $x = B^{-1}(a+\lambda^\star)$, where $\lambda^\star$ solves the convex problem (8) with $\mathcal{U} = \{(a,B)\}$. In other words, $x = B^{-1}(a+\lambda^\star)$ gives the hyperplane through the origin that maximizes the probability of z being on its positive side. The associated maximum probability is $\Phi([(a+\lambda^\star)^T B^{-1}(a+\lambda^\star)]^{1/2})$. Thus, $(a+\lambda^\star)^T B^{-1}(a+\lambda^\star)$ (which is the objective of (8)) can be used to measure the extent to which a hyperplane perpendicular to $x \in \mathcal{X}$ can separate a random signal $z \sim \mathcal{N}(a,B)$ from the origin.

We give another interpretation. Suppose that we know the mean $\mathbf{E}\, z = a$ and the covariance $\mathbf{E}(z-a)(z-a)^T = B$ of z, but its third and higher moments are unknown.

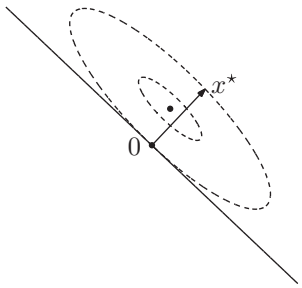


FIG. 1. *Illustration of $x^\star = B^{-1}a$. The center of the two confidence ellipsoids (whose boundaries are shown as dashed line curves) is a, and their shapes are determined by B.*

Here $\mathbf{E}$ denotes the expectation operation. Then, $\mathbf{E}\, x^T z = x^T a$ and $\mathbf{E}(x^T z - x^T a)^2 = x^T B x$, so by the Chebyshev bound, we have

$$\mathbf{Prob}\left(x^T z \geq 0\right) \geq \Psi\left(\frac{x^T a}{\sqrt{x^T B x}}\right), \tag{11}$$

where

$$\Psi(u) = \frac{\max\{u, 0\}^2}{1 + \max\{u, 0\}^2}.$$

This bound is sharp; in other words, there is a distribution for z with mean a and covariance B for which equality holds in (11) [3, 30]. Since Ψ is increasing, this probability is also maximized by $x = B^{-1}(a + \lambda^\star)$. Thus, $x = B^{-1}(a + \lambda^\star)$ gives the hyperplane through the origin and perpendicular to $x \in \mathcal{X}$ that maximizes the Chebyshev lower bound for $\mathbf{Prob}(x^T z \geq 0)$. The maximum value of the Chebyshev lower bound is $p^\star/(1 + p^\star)$, where $p^\star = \left[(a + \lambda^\star)^T B^{-1}(a + \lambda^\star)\right]^{1/2}$. This quantity assesses the maximum extent to which a hyperplane perpendicular to $x \in \mathcal{X}$ can separate from the origin a random signal z, whose first and second moments are known but otherwise arbitrary. This quantity is an increasing function of $p^\star$, so the hyperplane perpendicular to $x \in \mathcal{X}$ that maximally separates from the origin a Gaussian random signal $z \sim \mathcal{N}(a, B)$ also maximally separates, in the sense of the Chebyshev bound, a signal with known mean and covariance.

When $\mathcal{X} = \mathbb{R}^n \backslash \{0\}$, we have $\mathcal{X}^* = 0$, so $x = B^{-1}a$ maximizes the right-hand side of (10). We can give its graphical interpretation. We find the confidence ellipsoid of the Gaussian distribution $\mathcal{N}(a, B)$, whose boundary touches the origin. This ellipsoid is tangential to the hyperplane through the origin and perpendicular to $x = B^{-1}a$. Figure 1 illustrates this interpretation in $\mathbb{R}^2$.

2.2. Robust linear separation. We now assume that the mean and covariance are uncertain but known to belong to a convex compact subset $\mathcal{U}$ of $\mathbb{R}^n \times \mathbb{S}^n_{++}$. We make one assumption: for each $(a, \Sigma) \in \mathcal{U}$, we have $a \neq 0$. In other words, we rule out the possibility that the mean is zero.

Theorem 1 tells us that there exists a triple $(x^\star, a^\star, B^\star)$, with $x^\star \in \mathcal{X}$ and $(a^\star, B^\star) \in \mathcal{U}$, such that

$$\Phi\left(\frac{x^T a^\star}{\sqrt{x^T B^\star x}}\right) \leq \Phi\left(\frac{x^{\star T} a^\star}{\sqrt{x^{\star T} B^\star x^\star}}\right) \leq \Phi\left(\frac{x^{\star T} a}{\sqrt{x^{\star T} B x^\star}}\right) \quad \forall x \in \mathcal{X} \quad \forall (a, B) \in \mathcal{U}. \tag{12}$$

Here we use the fact that Φ is strictly increasing.

From the saddle-point property (12), we can see that $x^\star$ solves

$$\begin{array}{ll} \text{maximize} & \inf\limits_{(a,B)\in\mathcal{U}, z\sim\mathcal{N}(a,B)} \mathbf{Prob}\left(x^T z \geq 0\right) \\ \text{subject to} & x \in \mathcal{X}, \end{array} \tag{13}$$

and the pair $(a^\star, B^\star)$ solves

$$\begin{array}{ll} \text{minimize} & \sup\limits_{x\in\mathcal{X}, z\sim\mathcal{N}(a,B)} \mathbf{Prob}\left(x^T z > 0\right) \\ \text{subject to} & (a,B) \in \mathcal{U}. \end{array} \tag{14}$$

Problem (13) is to find a hyperplane through the origin and perpendicular to $x \in \mathcal{X}$ that separates robustly a normal random variable z on $\mathbb{R}^n$ with uncertain first and second moments belonging to $\mathcal{U}$. Problem (14) is to find the least-favorable model in terms of the separation probability (when the random variable is normal). It follows from (10) that (13) is equivalent to the max-min problem (3), and (14) is equivalent to (4) and hence to the convex problem (8) by Proposition 1. These two problems can be solved using convex optimization.

We close by pointing out that the same results hold with the Chebyshev bound as the separation probability.

3. Robust Fisher discriminant analysis. As another application, we consider a robust classification problem.

3.1. Fisher linear discriminant analysis. In linear discriminant analysis, we want to separate two classes which can be identified with two random variables in $\mathbb{R}^n$. Fisher linear discriminant analysis (FLDA) is a widely used technique for pattern classification, proposed by Fisher in the 1930s. The reader is referred to standard textbooks on statistical learning, e.g., [13], for more on FLDA.

For a (linear) discriminant characterized by $w \in \mathbb{R}^n$, the degree of discrimination is measured by the Fisher discriminant ratio

$$F(w, \mu_+, \mu_-, \Sigma_+, \Sigma_-) = \frac{\left(w^T(\mu_+ - \mu_-)\right)^2}{w^T(\Sigma_+ + \Sigma_-)w},$$

where μ_+ and Σ_+ (μ_+ and Σ_-) denote the mean and covariance, respectively, of examples drawn from the positive (negative) class. A discriminant that maximizes the Fisher discriminant ratio is given by

$$\bar{w} = (\Sigma_+ + \Sigma_-)^{-1}(\mu_+ - \mu_-),$$

which gives the maximum Fisher discriminant ratio

$$\sup_{w\neq 0} F(w, \mu_+, \mu_-, \Sigma_+, \Sigma_-) = (\mu_+ - \mu_-)^T(\Sigma_+ + \Sigma_-)^{-1}(\mu_+ - \mu_-).$$

Once the optimal discriminant is found, we can form the (binary) classifier

$$\phi(x) = \operatorname{sgn}\left(\bar{w}^T x + v\right), \tag{15}$$

where

$$\operatorname{sgn}(z) = \begin{cases} +1, & z > 0, \\ -1, & z \leq 0, \end{cases}$$

and v is the bias or threshold. The classifier picks the outcome, given x, according to the linear boundary between the two binary outcomes (defined by $\bar{w}^T x + v = 0$).

We can give a probabilistic interpretation of FLDA. Suppose that $x \sim \mathcal{N}(\mu_+, \Sigma_+)$ and $y \sim \mathcal{N}(\mu_-, \Sigma_-)$. We want to find w that maximizes $\mathbf{Prob}(w^T x > w^T y)$. Here,

$$x - y \sim \mathcal{N}(\mu_+ - \mu_-, \Sigma_+ + \Sigma_-),$$

so

$$\mathbf{Prob}\left(w^T x > w^T y\right) = \mathbf{Prob}\left(w^T (x-y) > 0\right) = \Phi\left(\frac{w^T(\mu_+ - \mu_-)}{\sqrt{w^T(\Sigma_+ + \Sigma_-)w}}\right).$$

This probability is called the nominal discrimination probability. Evidently, FLDA amounts to maximizing the fractional function

$$f(w, \mu_+ - \mu_-, \Sigma_+ + \Sigma_-) = \frac{w^T(\mu_+ - \mu_-)}{\sqrt{w^T(\Sigma_+ + \Sigma_-)w}}.$$

3.2. Robust Fisher linear discriminant analysis. In FLDA, the problem data or parameters (i.e., the first and second moments of the two random variables) are not known but are estimated from sample data. FLDA can be sensitive to the variation or uncertainty in the problem data, meaning that the discriminant computed from an estimate of the parameters can give very poor discrimination for another set of problem data that is also a reasonable estimate of the parameters. Robust FLDA attempts to systematically alleviate this sensitivity problem by explicitly incorporating a model of data uncertainty in the classification problem and optimizing for the worst-case scenario under this model; see [17] for more on robust FLDA and its extension.

We assume that the problem data μ_+, μ_-, Σ_+, and Σ_- are uncertain but known to belong to a convex compact subset $\mathcal{V}$ of $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{S}^n_{++} \times \mathbb{S}^n_{++}$. We make the following assumption:

$$\text{for each } (\mu_+, \mu_-, \Sigma_+, \Sigma_-) \in \mathcal{V}, \text{ we have } \mu_+ \neq \mu_-. \tag{16}$$

This assumption simply means that for each possible value of the means and covariances, the two classes are distinguishable via FLDA.

The *worst-case analysis problem* of finding the worst-case means and covariances for a given discriminant w can be written as

$$\begin{array}{ll} \text{minimize} & f(w, \mu_+ - \mu_-, \Sigma_+ + \Sigma_-) \\ \text{subject to} & (\mu_+, \mu_-, \Sigma_+, \Sigma_-) \in \mathcal{V}, \end{array} \tag{17}$$

with variables μ_+, μ_-, Σ_+, and Σ_-. Optimal points for this problem, say, $(\mu_+^{\mathrm{wc}}, \mu_-^{\mathrm{wc}}, \Sigma_+^{\mathrm{wc}}, \Sigma_-^{\mathrm{wc}})$, are called the *worst-case means and covariances*, which depend on w. With the worst-case means and covariances, we can compute the *worst-case discrimination probability*

$$\mathbf{P}_{\mathrm{wc}}(w) = \Phi\left(\frac{w^T(\mu_+^{\mathrm{wc}} - \mu_-^{\mathrm{wc}})}{\sqrt{w^T(\Sigma_+^{\mathrm{wc}} + \Sigma_-^{\mathrm{wc}})w}}\right)$$

(over the set $\mathcal{U}$ of possible means and covariances).

The *robust FLDA problem* is to find a discriminant that maximizes the worst-case Fisher discriminant ratio:

$$\begin{array}{ll} \text{maximize} & \inf\limits_{(\mu_+,\mu_-,\Sigma_+,\Sigma_-)\in\mathcal{V}} f(w, \mu_+ - \mu_-, \Sigma_+ + \Sigma_-) \\ \text{subject to} & w \neq 0, \end{array} \tag{18}$$

with variable w. Here we choose a linear discriminant that maximizes the Fisher discrimination ratio, with the worst possible means and covariances that are consistent with our data uncertainty model. Any solution to (18) is called a *robust optimal Fisher discriminant*.

The robust FLDA problem (18) has the form (3) with

$$\mathcal{U} = \{(\mu_+ - \mu_-, \Sigma_+ + \Sigma_-) \in \mathbb{R}^n \times \mathbb{S}^n_{++} \mid (\mu_+, \mu_-, \Sigma_+, \Sigma_-) \in \mathcal{U}\}.$$

In this problem, each element of the set $\mathcal{U}$ is a pair of the mean and covariance of the difference of the two random variables. For this problem, we can see from (16) that assumption (7) holds. The robust FLDA problem can therefore be solved by using the minimax result described above.

3.3. Numerical example. We illustrate the result with a classification problem in $\mathbb{R}^2$. The nominal means and covariances of the two classes are

$$\bar{\mu}_+ = (1, 0), \qquad \bar{\mu}_- = (-1, 0), \qquad \bar{\Sigma}_+ = \bar{\Sigma}_- = I \in \mathbb{R}^{2\times 2}.$$

We assume that only μ_+ is uncertain and lies within the ellipse

$$\mathcal{E} = \left\{\mu_+ \in \mathbb{R}^2 \mid \mu_+ = \bar{\mu}_+ + Pu,\ \|u\| \leq 1\right\},$$

where the matrix P which determines the shape of the ellipse is

$$P = \begin{bmatrix} 0.78 & 0.64 \\ 0.64 & 0.78 \end{bmatrix} \in \mathbb{R}^{2\times 2}.$$

Figure 2 illustrates the setting described above. Here the shaded ellipse corresponds to $\mathcal{E}$, and the dashed line curves are the set of points μ_+ and μ_- that satisfy

$$\left\|{\Sigma_+}^{-1/2}(\mu_+ - \bar{\mu}_+)\right\| = \|\mu_+ - \bar{\mu}_+\| = 1, \qquad \left\|{\Sigma_-}^{-1/2}(\mu_- - \bar{\mu}_-)\right\| = \|\mu_- - \bar{\mu}_-\| = 1.$$

The nominal optimal discriminant which maximizes the Fisher discriminant ratio with the nominal means and covariances is given by $w^{\text{nom}} = (1, 0)$. The robust optimal discriminant w^{rob} is computed using the method described above. Figure 2 shows two linear decision boundaries

$$x^T w^{\text{nom}} = 0, \qquad x^T w^{\text{rob}} = 0$$

determined by the two discriminants. Since the mean of the positive class is uncertain and the uncertainty is significant in a certain direction, the robust discriminant is tilted toward the direction.

Table 1 summarizes the results. Here, $\mathbf{P}_{\text{nom}}$ is the nominal discrimination probability and $\mathbf{P}_{\text{wc}}$ is the worst-case discrimination probability. The nominal optimal discriminant achieves $\mathbf{P}_{\text{nom}} = 0.92$, which corresponds to 92% of correct discrimination without uncertainty. However, with uncertainty present, its nominal discrimination probability degrades rapidly; the worst-case discrimination probability for the nominal optimal discriminant is 78%. The robust optimal discriminant performs well in the presence of uncertainty. It has a worst-case discrimination probability around 83%, 5% higher than that of the nominal optimal discriminant.

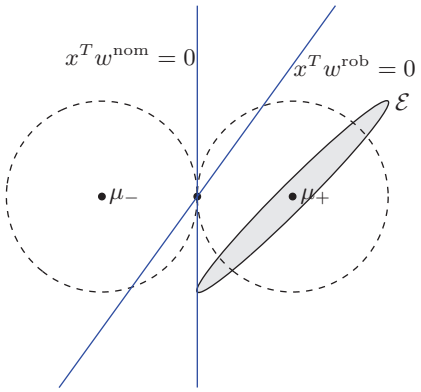


FIG. 2. *A simple example for robust FLDA.*

TABLE 1
Robust discriminant analysis results.

	$\mathbf{P}_{\text{nom}}$	$\mathbf{P}_{\text{wc}}$
Nominal optimal discriminant	0.92	0.78
Robust optimal discriminant	0.87	0.83

4. Robust matched filtering. We consider a signal model of the form

$$y(t) = s(t)a + v(t) \in \mathbb{R}^n,$$

where a is the steering vector, $s(t) \in \{0, 1\}$ is the binary source signal, $y(t) \in \mathbb{R}^n$ is the received signal, and $v(t) \sim \mathcal{N}(0, \Sigma)$ is the noise. We consider the problem of estimating $s(t)$, based on an observed sample of y. In other words, the sample is generated from one of the two possible distributions $\mathcal{N}(0, \Sigma)$ and $\mathcal{N}(a, \Sigma)$, and we are to guess which one.

After reviewing a basic result on optimal detection with the setting described above, we show how the minimax result given above allows us to design a robust detector that takes into account the uncertainty in the model parameters, namely, the steering vector and the noise covariance.

4.1. Matched filtering. A (deterministic) detector is a function ψ from $\mathbb{R}^n$ (the set of possible observed values) into $\{0, 1\}$ (the set of possible signal values or hypotheses). It can be expressed as

$$\psi(y) = \begin{cases} 0, & h(y) < t, \\ 1, & h(y) > t, \end{cases} \tag{19}$$

which thresholds a detection or test statistic, a function of the received signal, $h(y) \in \mathbb{R}$. Here t is the threshold that determines the boundary between the two hypotheses. A detector with a detection statistic of the form $h(y) = w^T y$ is called linear.

The performance of a detector ψ can be summarized by the pair $(P_{\text{fp}}, P_{\text{tp}})$, where

$$P_{\text{fp}} = \mathbf{Prob}(\psi(y) = 1 \mid s(t) = 0)$$

is the *false positive or alarm rate* (the probability that the signal is falsely detected when in fact there is no signal) and

$$P_{\text{tp}} = \mathbf{Prob}(\psi(y) = 1 \mid s(t) = 1)$$

is the *true positive rate* (the probability that the signal is detected correctly). The optimal detector design problem is a bicriterion problem, with objectives $P_{\rm fn}$ and $P_{\rm fp}$. The optimal trade-off curve between $P_{\rm fn}$ and $P_{\rm fp}$ is called the *receiver operating characteristic* (ROC).

The filtered output, with weight vector $w \in \mathbb{R}^n$, is given by

$$w^T y(t) = s(t) w^T a + w^T v(t).$$

The power of the steering vector $w^T a$ (which is deterministic) at the filtered output is given by $(w^T a)^2$, and the power of the undesired signal $w^T v$ at the filtered output is $w^T \Sigma w$. The signal to noise ratio (SNR) is

$$S(w, a, \Sigma) = \frac{\left(w^T a\right)^2}{w^T \Sigma w}.$$

The optimal ROC curve is obtained using a linear detection statistic $h(y) = w^{\star T} y$ with $w^\star$ maximizing

$$f(w, a, \Sigma) = \frac{w^T a}{\sqrt{w^T \Sigma w}},$$

which is the square root of the SNR (SSNR). (See, e.g., [28].) The weight vector that maximizes SSNR is given by $w = \Sigma^{-1} a$. When the covariance is a scaled identity matrix, the matched filter $w = a$ is optimal. Even when Σ is not a scaled identity matrix, the optimal weight vector is called the matched filter.

4.2. Robust matched filtering. Matched filtering is often sensitive to the uncertainty in the input parameters, namely, the steering vector and the noise covariance. Robust matched filtering attempts to alleviate the sensitivity problem by taking into account an uncertainty model in the detection problem. (The reader is referred to the tutorial [15] for more on robust signal detection.)

We assume that the desired signal and covariance matrix are uncertain but known to belong to a convex compact subset $\mathcal{U}$ of $\mathbb{R}^n \times \mathbb{S}^n_{++}$. We make a technical assumption:

$$a \neq 0 \qquad \forall (a, \Sigma) \in \mathcal{U}. \tag{20}$$

In other words, we rule out the possibility that the signal we want to detect is zero.

The *worst-case SSNR analysis problem* of finding a steering vector and a covariance that minimize SSNR for a given weight vector w can be written as

$$\begin{array}{ll} \text{minimize} & f(w, a, \Sigma) \\ \text{subject to} & (a, \Sigma) \in \mathcal{U}, \end{array} \tag{21}$$

with variables a and Σ. The optimal value of this problem is the *worst-case SSNR* (over the uncertainty set $\mathcal{U}$).

The *robust matched-filtering problem* is to find a weight vector that maximizes the worst-case SSNR, which can be cast as

$$\begin{array}{ll} \text{maximize} & \inf\limits_{(a,\Sigma)\in\mathcal{U}} f(x, a, B) \\ \text{subject to} & w \neq 0, \end{array} \tag{22}$$

with variables w. (The thresholding rule $h(y) = w^{\star T} y$ that uses a solution $w^\star$ of this problem as the weight vector yields the robust ROC curve that characterizes limits of performance in the worst-case sense.)

The robust signal detection setting described above is exactly the minimax setting described in the introduction, where a is the steering vector and B is the noise covariance. For this problem, we can see from the compactness of $\mathcal{U}$ and (20) that assumption (7) holds. We can solve the robust matched-filtering problem (22), using the minimax result for the fractional function (1).

We close by pointing out that we can handle convex constraints on the weight vector. For example, in robust beamforming, a special type of robust matched-filtering problem, we often want to choose the weight vector that maximizes the worst-case SSNR, subject to a unit array gain for the desired wave and rejection constraints on interferences [22]. This problem can also be solved using Theorem 1.

4.3. Numerical example. As an illustrative example, we consider the case when $a = (2, 3, 2, 2)$ is fixed (with no uncertainty) and the noise covariance Σ is uncertain and has the form

$$\begin{bmatrix} 1 & - & + & - \\ & 1 & ? & + \\ & & 1 & ? \\ & & & 1 \end{bmatrix}.$$

(Only the upper triangular part is shown because the matrix is symmetric.) Here, "+" means that $\Sigma_{ij} \in [0, 1]$, "−" means that $\Sigma_{ij} \in [-1, 0]$, and "?" means that $\Sigma_{ij} \in [-1, 1]$. Of course we assume $\Sigma \succ 0$. The nominal noise covariance is taken as

$$\bar{\Sigma} = \begin{bmatrix} 1 & -.5 & .5 & -.5 \\ & 1.0 & 0.0 & .5 \\ & & 1.0 & 0.0 \\ & & & 1.0 \end{bmatrix}.$$

Here, the upper-triangular part is shown since the matrix is symmetric. With the nominal covariance, we compute the nominal optimal weight vector or filter.

The least-favorable covariance, found by solving the convex problem (8) corresponding to the problem data above, is given by

$$\Sigma^{\mathrm{lf}} = \begin{bmatrix} 1.00 & 0.00 & .38 & -.12 \\ & 1.00 & .41 & .74 \\ & & 1.00 & .23 \\ & & & 1.00 \end{bmatrix}.$$

With the least-favorable covariance, we compute the robust optimal weight vector or filter.

Table 2 summarizes the results. The nominal optimal filter achieves an SSNR of 5.5 without uncertainty. In the presence of uncertainty, the SSNR achieved by the filter can degrade rapidly; the worst-case SSNR level for the nominal optimal filter is 3.0. The robust filter performs well in the presence of model mismatch; it has the worst-case SSNR of 3.6, which is 20% larger than that of the nominal optimal filter.

5. Worst-case Sharpe ratio maximization. The minimax result has an important application in robust portfolio selection.

5.1. Mean-variance asset allocation. Since the pioneering work of Markowitz [20], mean-variance (MV) analysis has been a topic of extensive research. In MV analysis, the (percentage) returns of risky assets $1, \ldots, n$ over a period are modeled

TABLE 2
Robust matched-filtering results.

	Nominal SSNR	Worst-case SSNR
Nominal optimal filter	5.5	3.0
Robust optimal filter	4.9	3.6

as a random vector $a = (a_1, \ldots, a_n)$ in $\mathbb{R}^n$. The input data or parameters needed for MV analysis are the mean μ and the covariance matrix Σ of a:

$$\mu = \mathbf{E}\, a, \qquad \Sigma = \mathbf{E}\, (a-\mu)(a-\mu)^T.$$

We assume that there is a risk-free asset with deterministic return $\mu_{\rm rf}$ and zero variance.

A portfolio $w \in \mathbb{R}^{n+1}$ is a finite linear combination of the assets. Let w_i denote the amount of asset i held throughout the period. A long position in asset i corresponds to $w_i > 0$, and a short position in asset i corresponds to $w_i < 0$. The return of a portfolio $w = (w_1, \ldots, w_n)$ is a (scalar) random variable $w^T a = \sum_{i=1}^n w_i a_i$, whose mean and volatility (standard deviation) are $\mu^T w$ and $\sqrt{w^T \Sigma w}$, respectively. We assume that an admissible portfolio $w = (w_1, \ldots, w_n)$ is constrained to lie in a convex compact subset $\mathcal{A}$ of $\mathbb{R}^n$. The portfolio budget constraint on w can be expressed, without loss of generality, as $\mathbf{1}^T w = 1$. Here $\mathbf{1}$ is the vector of all ones. The set of admissible portfolios subject to the portfolio budget constraint is given by

$$\mathcal{W} = \left\{ w \mid w \in \mathcal{A},\ \mathbf{1}^T w = 1 \right\}.$$

The performance of an admissible portfolio is often measured by its reward-to-variability or Sharpe ratio (SR):

$$S(w, \mu, \Sigma) = \frac{\mu^T w - \mu_{\rm rf}}{\sqrt{w^T \Sigma w}}.$$

The admissible portfolio that maximizes the ratio over $\mathcal{W}$ is called the *tangency portfolio* (TP). The SR achieved by this portfolio is called the market price of risk. The TP plays an important role in asset pricing theory and practice (see, e.g., [8, 19, 24]).

If the n risky assets with (single period) returns follow $a \sim \mathcal{N}(\mu, \Sigma)$, then

$$w^T a \sim \mathcal{N}\left(w^T \mu, w^T \Sigma w\right),$$

so the probability of outperforming the risk-free return $\mu_{\rm rf}$ is

$$\mathbf{Prob}\left(a^T w > \mu_{\rm rf}\right) = \Phi\left(\frac{\mu^T w - \mu_{\rm rf}}{\sqrt{w^T \Sigma w}}\right).$$

This probability is maximized by the TP.

SR maximization is related to the safety-first approach to portfolio selection [23], through the Chebyshev bound. Suppose $\mathbf{E}\, a = \mu$, $\mathbf{E}\, (a-\mu)^T (a-\mu) = \Sigma$ and otherwise arbitrary. Then, $\mathbf{E}\, a^T w = \mu^T x$ and $\mathbf{E}\, (a^T x - \mu^T x)^2 = w^T \Sigma w$, so it follows from the Chebyshev bound that

$$\mathbf{Prob}\left(a^T w \geq \mu_{\rm rf}\right) \geq \Psi\left(\frac{\mu^T w - \mu_{\rm rf}}{\sqrt{w^T \Sigma w}}\right).$$

In the safety-first approach [23], we want to find a portfolio that maximizes the bound. Since Ψ is increasing, this bound is also maximized by the TP.

5.2. Worst-case SR maximization. The input parameters are estimated with error. Conventional MV allocation is often sensitive to the uncertainty or the estimation error in the parameters, meaning that optimal portfolios computed with an estimate of the parameters can give very poor performance for another set of parameters that is similar and statistically hard to distinguish from the estimate; see, e.g., [4, 5, 14, 21], to name a few. Robust MV portfolio analysis attempts to systematically alleviate the sensitivity problem of conventional MV allocation by explicitly incorporating an uncertainty model on the input data or parameters in a portfolio selection problem and carrying out the analysis for the worst-case scenario under this model. Recent work on robust portfolio optimization includes [9, 10, 11, 12, 18].

In this section, we consider the robust counterpart of the SR maximization problem. The reader is referred to [16] for the importance of this problem in robust MV analysis. In this paper, we focus on the computational aspects of the robust counterpart.

We assume that the expected return μ and covariance Σ of the asset returns are uncertain but known to belong to a convex compact subset $\mathcal{U}$ of $\mathbb{R}^n \times \mathbb{S}^n_{++}$. We also assume there exists an admissible portfolio $\bar{w} \in \mathcal{W}$ of risky assets whose worst-case mean return is greater than the risk-free return:

$$\text{there exists a portfolio } \bar{w} \in \mathcal{W} \text{ such that } \mu^T w > \mu_{\rm rf} \text{ for all } (\mu, \Sigma) \in \mathcal{U}. \tag{23}$$

Worst-case SR maximization. The zero-sum game of choosing w from $\mathcal{W}$, to maximize the SR, and choosing (μ, Σ) from $\mathcal{U}$, to minimize the SR, is associated with the following two problems:

- worst-case SR maximization problem of finding an admissible portfolio w that maximizes the worst-case SR (over the given model $\mathcal{U}$ of uncertainty)

$$\begin{array}{ll} \text{maximize} & \inf\limits_{(\mu,\Sigma)\in\mathcal{U}} S(w, \mu, \Sigma) \\ \text{subject to} & w \in \mathcal{W}, \end{array} \tag{24}$$

- worst-case market price of risk analysis (MPRA) problem of finding the least-favorable statistics (over the uncertainty set $\mathcal{U}$), with portfolio weights chosen optimally for the asset return statistics,

$$\begin{array}{ll} \text{minimize} & \sup\limits_{w\in\mathcal{W}} S(w, \mu, \Sigma) \\ \text{subject to} & (\mu, \Sigma) \in \mathcal{U}. \end{array} \tag{25}$$

The SR is not a fractional function of the form (1), so we cannot apply Theorem 1 directly to the zero-sum game given above. We can get around this difficulty by using the fact that when the domain is restricted to $\mathcal{W}$, the SR has the form (1)

$$\frac{\mu^T w - \mu_{\rm rf}}{\sqrt{w^T \Sigma w}} = \frac{w^T(\mu - \mu_{\rm rf}\mathbf{1})}{\sqrt{w^T \Sigma w}} = f(w, \mu - \mu_{\rm rf}\mathbf{1}, \Sigma) \qquad \forall w \in \mathcal{W}$$

and

$$S(w, \mu, \Sigma) = f(tw, \mu - \mu_{\rm rf}\mathbf{1}, \Sigma), \quad w \in \mathcal{W}, \quad t > 0, \tag{26}$$

whenever $S(w, \mu, \Sigma) > 0$.

The set

$$\mathcal{X} = \mathrm{cl}\,\{tw \in \mathbb{R}^n \mid w \in \mathcal{W},\ t > 0\} \backslash \{0\},$$

where $\operatorname{cl} A$ means the closure of the set A and $A\backslash B$ means the complement of B in A, is a cone in $\mathbb{R}^n$, with $\mathcal{X} \cup \{0\}$ closed and convex. Assumption (23), along with the compactness of $\mathcal{U}$, means that

$$\inf_{(\mu,\Sigma)\in\mathcal{U}} \bar{w}^T(\mu - \mu_{\rm rf}\mathbf{1}) > 0.$$

We can therefore apply Theorem 1 to the zero-sum game of choosing w from $\mathcal{X}$, to maximize $f(x, \mu-\mu_{\rm rf}\mathbf{1}, \Sigma)$, and choosing (μ, Σ) from $\mathcal{U}$, to minimize $f(x, \mu-\mu_{\rm rf}\mathbf{1}, \Sigma)$.

The max-min and min-max problems associated with the game are

- max-min problem

$$\begin{array}{ll} \text{maximize} & \inf\limits_{(\mu,\Sigma)\in\mathcal{U}} f(x, \mu - \mu_{\rm rf}\mathbf{1}, \Sigma) \\ \text{subject to} & x \in \mathcal{X}, \end{array} \tag{27}$$

- min-max problem

$$\begin{array}{ll} \text{minimize} & \sup\limits_{x\in\mathcal{X}} f(x, \mu - \mu_{\rm rf}\mathbf{1}, \Sigma) \\ \text{subject to} & (\mu, \Sigma) \in \mathcal{U}. \end{array} \tag{28}$$

According to Theorem 1, the two problems have the same optimal value:

$$\sup_{x\in\mathcal{X}} \inf_{(\mu,\Sigma)\in\mathcal{U}} f(x, \mu - \mu_{\rm rf}\mathbf{1}, \Sigma) = \inf_{(\mu,\Sigma)\in\mathcal{U}} \sup_{x\in\mathcal{X}} f(x, \mu - \mu_{\rm rf}\mathbf{1}, \Sigma). \tag{29}$$

As a result, the SR satisfies the minimax equality

$$\sup_{w\in\mathcal{W}} \inf_{(\mu,\Sigma)\in\mathcal{U}} S(w, \mu, \Sigma) = \inf_{(\mu,\Sigma)\in\mathcal{U}} \sup_{w\in\mathcal{W}} S(w, \mu, \Sigma),$$

which follows from (26) and (29).

From Proposition 1, we can see that the min-max problem (28) is equivalent to the convex problem

$$\begin{array}{ll} \text{minimize} & (\mu - \mu_{\rm rf}\mathbf{1} + \lambda)^T\Sigma^{-1}(\mu - \mu_{\rm rf}\mathbf{1} + \lambda) \\ \text{subject to} & (\mu, \Sigma) \in \mathcal{U}, \quad \lambda \in \mathcal{W}^{\oplus} \end{array} \tag{30}$$

in which the optimization variables are $\mu \in \mathbb{R}^n$, $\Sigma = \Sigma^T \in \mathbb{R}^{n\times n}$, and $\lambda \in \mathbb{R}^n$. Here $\mathcal{W}^{\oplus}$ is the positive conjugate cone $\mathcal{W}$, which is equal to the dual cone $X^\star$ of $\mathcal{X}$:

$$\mathcal{X}^* = \mathcal{W}^{\oplus} = \left\{\lambda \in \mathbb{R}^n \mid \lambda^T w \geq 0 \ \forall w \in \mathcal{W}\right\}.$$

The convex problem (30) has a solution, say, $(\mu^\star, \Sigma^\star, \lambda^\star)$. Then,

$$x^\star = {\Sigma^\star}^{-1}(\mu^\star - \mu_{\rm rf}\mathbf{1} + \lambda^\star) \in \mathcal{X}$$

is a unique solution of the max-min problem (27) (up to positive scaling). Moreover, the saddle-point property

$$f(x, \mu^\star - \mu_{\rm rf}\mathbf{1}, \Sigma^\star) \leq f(x^\star, \mu^\star - \mu_{\rm rf}\mathbf{1}, \Sigma^\star) \leq f(x^\star, \mu - \mu_{\rm rf}\mathbf{1}, \Sigma), \quad x \in \mathcal{X}, \quad (\mu, \Sigma) \in \mathcal{U}, \tag{31}$$

holds. We can see from (26) that

$$S(w, \mu^\star, \Sigma^\star) \leq S(w^\star, \mu^\star, \Sigma^\star) \leq S(w^\star, \mu, \Sigma) \quad \forall w \in \mathcal{W} \quad \forall (\mu, \Sigma) \in \mathcal{U}. \tag{32}$$

Finally, since $\mathbf{1}^T x \geq 0$ for all $x \in \mathcal{X}$, we have

$$\mathbf{1}^T \Sigma^{\star -1}(\mu^\star - \mu_{\rm rf}\mathbf{1} + \lambda^\star) \geq 0.$$

If $x^\star$ satisfies $\mathbf{1}^T x^\star > 0$, the portfolio

$$w^\star = \left(1/\mathbf{1}^T x^\star\right) x^\star = \frac{1}{\mathbf{1}^T \Sigma^{\star -1}(\mu^\star - \mu_{\rm rf}\mathbf{1} + \lambda^\star)} \Sigma^{\star -1}(\mu^\star - \mu_{\rm rf}\mathbf{1} + \lambda^\star) \tag{33}$$

satisfies the budget constraint and is admissible (i.e., $w^\star \in \mathcal{W}$); i.e., it is a solution to the worst-case SR maximization (24). Moreover, it is the unique solution to the worst-case SR maximization (24). The case of $\mathbf{1}^T x^\star = 0$ may arise when the set $\mathcal{W}$ is unbounded. In this case, the worst-case SR maximization problem (24) has no solution, so the game involving the SR has no saddle point.

Minimax properties of the SR. The results established above are summarized in the following proposition.

PROPOSITION 3. *Suppose that the uncertainty set $\mathcal{U}$ is compact and convex. Suppose further that Assumption* (23) *holds. Let $(\mu^\star, \Sigma^\star, \lambda^\star)$ be a solution to the convex problem* (30). *Then, we have the following:*

(i) *If $\mathbf{1}^T \Sigma^{-1}(\mu - \mu_{\rm rf}\mathbf{1} + \lambda^\star) > 0$, then the triple $(w^\star, \mu^\star, \Sigma^\star)$ with $w^\star$ in* (33) *satisfies the saddle-point property* (32), *and $w^\star$ is the unique solution to the worst-case SR maximization problem* (24).

(ii) *If $\mathbf{1}^T \Sigma^{-1}(\mu - \mu_{\rm rf}\mathbf{1} + \lambda^\star) = 0$, then the optimal value of the worst-case SR maximization problem* (24) *is not achieved by any portfolio in $\mathcal{W}$.*

Moreover, the minimax equality

$$\sup_{w \in \mathcal{W}} \inf_{(\mu,\Sigma) \in \mathcal{U}} S(w, \mu, \Sigma) = \inf_{(\mu,\Sigma) \in \mathcal{U}} \sup_{w \in \mathcal{W}} S(w, \mu, \Sigma)$$

holds regardless of the existence of a solution.

The worst-case MPRA problem (25) is equivalent to the min-max problem (28), which is in turn equivalent to the convex problem (30). This proposition shows that the TP of the least-favorable model $(\mu^\star, \Sigma^\star)$ solves the worst-case SR maximization problem (24). The saddle-point property (32) means that the portfolio $w^\star$ in (33) is the TP of the least-favorable model $(\mu^\star, \Sigma^\star)$. The portfolio is called the *robust TP*.

5.3. Numerical example. We illustrate the result with a synthetic example, with $n = 7$ risky assets. The risk-free return is taken as $\mu_{\rm rf} = 5$.

Setup. The nominal returns $\bar{\mu}_i$ and variances $\bar{\sigma}_i^2$ of the risky assets are taken as

$$\bar{\mu} = [10.3 \;\; 10.5 \;\; 5.5 \;\; 10.5 \;\; 110 \;\; 14.4 \;\; 10.1]^T,$$
$$\bar{\sigma} = [11.3 \;\; 18.1 \;\; 6.8 \;\; 22.7 \;\; 24.0 \;\; 14.7 \;\; 20.9]^T.$$

The nominal correlation matrix $\bar{\Omega}$ is

$$\bar{\Omega} = \begin{bmatrix} 1.00 & .07 & -.12 & .43 & -.11 & .44 & .25 \\ & 1.00 & .73 & -.14 & .39 & .28 & .10 \\ & & 1.00 & .14 & .50 & .52 & -.13 \\ & & & 1.00 & .04 & .35 & .38 \\ & & & & 1.00 & .70 & .04 \\ & & & & & 1.00 & -.09 \\ & & & & & & 1.00 \end{bmatrix}.$$

The nominal covariance is

$$\bar{\Sigma} = \operatorname{diag}(\bar{\sigma})\bar{\Omega}\operatorname{diag}(\bar{\sigma}),$$

where we use $\operatorname{diag}(u_1, \ldots, u_m)$ to denote the diagonal matrix with diagonal entries $u_1, \ldots, u_m$.

The mean uncertainty model used in our study is

$$\begin{aligned} |\mu_i - \bar{\mu}_i| &\le 0.3|\bar{\mu}_i|, \quad i = 1, \ldots, 7, \\ \left|\mathbf{1}^T\mu - \mathbf{1}^T\bar{\mu}\right| &\le 0.15\left|\mathbf{1}^T\bar{\mu}\right|. \end{aligned}$$

These constraints mean that the possible variation in the expected return of each asset is at most 30%, and the possible variation in the expected return of the portfolio $(1/n)\mathbf{1}$ (in which a fraction $1/n$ of the budget is allocated to each asset of the n assets) is at most 15%. The covariance uncertainty model used in our study is

$$\begin{aligned} \left|\Sigma_{ij} - \bar{\Sigma}_{ij}\right| &\le 0.3\left|\bar{\Sigma}_{ij}\right|, \quad i, j = 1, \ldots, 7, \\ \left\|\Sigma - \bar{\Sigma}\right\|_F &\le 0.15\left\|\bar{\Sigma}\right\|_F. \end{aligned}$$

(Here, $\|A\|_F$ denotes the Frobenius norm of A, i.e., $\|A\|_F = (\sum_{i,j=1}^n A_{ij}^2)^{1/2}$.) These constraints mean that the possible variation in each component of the covariance matrix is at most 30% and the possible deviation of the covariance from the nominal covariance is at most 15% in terms of the Frobenius norm.

We consider the case when short selling is allowed in a limited way as follows:

$$w = w_{\text{long}} - w_{\text{short}}, \quad w_{\text{long}}, w_{\text{short}} \succeq 0, \quad \mathbf{1}^T w_{\text{short}} \le \eta \mathbf{1}^T w_{\text{long}}, \tag{34}$$

where η is a positive constant and w_{long} and w_{short} represent the total long and short positions at the beginning of the period, respectively. ($w \succeq 0$ means that w is componentwise nonnegative.) The last constraint limits the total short position to some fraction η of the total long position. In our numerical study, we take $\gamma = 0.3$.

The asset constraint set is given by the cone

$$\mathcal{W} = \left\{ w \in \mathbb{R}^n \mid w = w_{\text{long}} - w_{\text{short}},\ A\begin{bmatrix} w_{\text{long}} \\ w_{\text{short}} \end{bmatrix} \preceq 0 \right\},$$

where

$$A = \begin{bmatrix} -I & 0 \\ 0 & -I \\ -\gamma\mathbf{1}^T & \mathbf{1}^T \end{bmatrix} \in \mathbb{R}^{(2n+1)\times(2n)}.$$

A simple argument based on linear programming duality shows that the dual cone of $\mathcal{X} = \mathcal{W}$ is given by

$$\mathcal{X}^* = \left\{ \lambda \in \mathbb{R}^n \;\middle|\; \text{there exists } y \succeq 0 \text{ such that } A^T y + \begin{bmatrix} \lambda \\ -\lambda \end{bmatrix} = 0 \right\}.$$

Comparison results. We can find the robust TP by applying Theorem 1 to the corresponding problem (27) with the asset allocation constraints and uncertainty model described above. The nominal TP can be found using Theorem 1 with the singleton $\mathcal{U} = \{(\bar{\mu}, \bar{\Sigma})\}$.

TABLE 3
Nominal and worst-case SRs of the nominal and robust TPs.

	Nominal SR	Worst-case SR
Nominal TP	0.74	0.22
Robust TP	0.57	0.36

TABLE 4
Outperformance probability of the nominal and robust TPs.

	$\mathbf{P}_{\mathrm{nom}}$	$\mathbf{P}_{\mathrm{wc}}$
Nominal TP	0.77	0.59
Robust TP	0.71	0.64

Table 3 shows the nominal and worst-case SRs of the nominal optimal and robust optimal allocations. In comparison with the market portfolio, the robust market portfolio shows a relatively small decrease in the SR, in the presence of possible variations in the parameters. The SR of the robust market portfolio decreases about 39% from 0.57 to 0.36, while the SR of the nominal market portfolio decreases about 70% from 0.74 to 0.22.

Table 4 shows the probabilities of outperforming the risk-free asset for the nominal optimal and robust optimal weight allocations, when the asset returns follow a normal distribution. Here, $\mathbf{P}_{\mathrm{nom}}$ is the probability of beating the risk-free asset without uncertainty, called the outperformance probability, and $\mathbf{P}_{\mathrm{wc}}$ is the worst-case probability of outperforming the risk-free asset with uncertainty. The nominal optimal TP achieves $\mathbf{P}_{\mathrm{nom}} = 0.77$, which corresponds to 77% of outperforming the risk-free asset without uncertainty. However, in the presence of uncertainty in the parameters, its performance degrades rapidly; the worst-case outperformance probability for the nominal optimal discriminant is 59%. The robust optimal allocation performs well in the presence of uncertainty in the parameters, with the worst-case outperformance probability 5% higher than that of the nominal optimal allocation.

6. Conclusions. The fractional function $f(x, a, B) = a^T x/\sqrt{x^T B x}$ comes up in many contexts, some of which are discussed above. In this paper, we have established a minimax result for this function and a general computational method, based on convex optimization, for computing a saddle point.

The arguments used to establish the minimax result do not appear to be extensible to other fractional functions that have a similar form. For instance, the extension to a general fractional function of the form

$$g(x, A, B) = \frac{x^T A x}{x^T B x},$$

which is the Rayleigh quotient of the matrix pair $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times n}$ evaluated at $x \in \mathbb{R}^n$, is not possible; see, e.g., [31] for a counterexample. However, the arguments can be extended to the special case when A is a dyad, i.e., $A = aa^T$, with $a \in \mathbb{R}^n$, and $\mathcal{X} = \mathbb{R}^n\backslash\{0\}$. In this case, the minimax equality

$$\sup_{x\neq 0}\inf_{(a,B)\in\mathcal{U}} \frac{\left(x^T a\right)^2}{x^T B x} = \inf_{(a,B)\in\mathcal{U}}\sup_{x\neq 0} \frac{\left(x^T a\right)^2}{x^T B x}$$

holds with assumption (7); see [17] for the proof.

Appendix A. Proofs.

A.1. Proof of Proposition 1. We first show that the optimal value of (8) is positive. We start by noting that

$$\inf_{(a,B)\in\mathcal{U}} \frac{\bar{x}^T a}{\sqrt{\bar{x}^T B \bar{x}}} > 0, \tag{35}$$

with $\bar{x}$ in (7), and

$$\inf_{(a,B)\in\mathcal{U},\lambda\in\mathcal{X}^*} \frac{\bar{x}^T(a+\lambda)}{\sqrt{\bar{x}^T B\bar{x}}} = \inf_{(a,B)\in\mathcal{U}} \inf_{\lambda\in\mathcal{X}^*} \frac{\bar{x}^T(a+\lambda)}{\sqrt{\bar{x}^T B\bar{x}}} = \inf_{(a,B)\in\mathcal{U}} \frac{\bar{x}^T a}{\sqrt{\bar{x}^T B\bar{x}}}. \tag{36}$$

Here, we have used (35) and $\inf_{\lambda\in\mathcal{X}^*} \bar{x}^T\lambda = 0$. By the Cauchy–Schwarz inequality, $x^T(a+\lambda)/\sqrt{x^T Bx}$ is maximized over nonzero x by $x = B^{-1}(a+\lambda)$, so

$$\sup_{x\neq 0} x^T(a+\lambda)/\sqrt{x^T Bx} = \left[(a+\lambda)^T B^{-1}(a+\lambda)\right]^{1/2}.$$

It follows from the minimax inequality (5), (35), and (36) that

$$\begin{aligned}
\inf_{(a,B)\in\mathcal{U},\lambda\in\mathcal{X}^*} \left[(a+\lambda)^T B^{-1}(a+\lambda)\right]^{1/2} &= \inf_{(a,B)\in\mathcal{U},\lambda\in\mathcal{X}^*} \sup_{x\neq 0} \frac{x^T(a+\lambda)}{\sqrt{x^T Bx}} \\
&\geq \sup_{x\neq 0} \inf_{(a,B)\in\mathcal{U},\lambda\in\mathcal{X}^*} \frac{x^T(a+\lambda)}{\sqrt{x^T Bx}} \\
&\geq \inf_{(a,B)\in\mathcal{U},\lambda\in\mathcal{X}^*} \frac{\bar{x}^T(a+\lambda)}{\sqrt{\bar{x}^T B\bar{x}}} \\
&> 0.
\end{aligned}$$

(Here, we use the fact that the weak minimax property for $x^T(a+\lambda)/\sqrt{x^T Bx}$ holds for any $\mathcal{U} \subseteq \mathbb{R}^n \times \mathbb{S}^n_{++}$ and $\mathcal{X} \subseteq \mathbb{R}^n$.)

We next show that (8) has a solution. There is a sequence

$$\left\{ \left(a^{(i)} + \lambda^{(i)}, B^{(i)}\right) \;\middle|\; \left(a^{(i)}, B^{(i)}\right) \in \mathcal{U},\ \lambda^{(i)} \in \mathcal{X}^*,\ i = 1, 2, \dots \right\}$$

such that

$$\lim_{i\to\infty} \left(a^{(i)} + \lambda^{(i)}\right)^T {B^{(i)}}^{-1} \left(a^{(i)} + \lambda^{(i)}\right) = \inf_{(a,B)\in\mathcal{U},\lambda\in\mathcal{X}^*} (a+\lambda)^T B^{-1}(a+\lambda). \tag{37}$$

Since $\mathcal{U}$ is a compact subset of $\mathbb{R}^n \times \mathbb{S}^n_{++}$, we have

$$\sup\left\{\lambda_{\max}\left(B^{-1}\right) \;\middle|\; \text{for all } B \text{ with } (a,B)\in\mathcal{U}\right\} < \infty.$$

(Here $\lambda_{\max}(B)$ is the maximum eigenvalue of B.) Then, $S_1 = \{a^{(i)} + \lambda^{(i)} \in \mathbb{R}^n \mid i = 1, 2, \dots\}$ must be bounded. (Otherwise, there arises a contradiction to (37).) Since $\mathcal{U}$ is compact, the sequence $S_2 = \{(a^{(i)}, B^{(i)}) \in \mathcal{U} \mid i = 1, 2, \dots\}$ is bounded, which along with the boundedness of S_1 means that $S_3 = \{\lambda^{(i)} \in \mathbb{R}^n \mid i = 1, 2, \dots\}$ is also bounded. The bounded sequences S_2 and S_3 have convergent subsequences, which converge to, say, $(a^\star, B^\star)$ and $\lambda^\star$, respectively. Since $\mathcal{U}$ and $\mathcal{X}^*$ are closed, $(a^\star, B^\star) \in \mathcal{U}$ and $\lambda^\star \in \mathcal{X}^*$. The triple $(a^\star, B^\star, \lambda^\star)$ achieves the optimal value of (8). Since the optimal value is positive, $a^\star + \lambda^\star \neq 0$.

The equivalence between (4) and (8) follows from the following implication:

$$\sup_{x\in\mathcal{X}} x^T a > 0 \quad \Longrightarrow \quad \sup_{x\in\mathcal{X}} \frac{x^T a}{\sqrt{x^T B x}} = \inf_{\lambda\in\mathcal{X}^*} \left[(a+\lambda)^T B^{-1}(a+\lambda)\right]^{1/2}. \tag{38}$$

Then, (4) is equivalent to

$$\begin{array}{ll} \text{minimize} & \inf\limits_{\lambda\in\mathcal{X}^*} (a+\lambda)^T B^{-1}(a+\lambda) \\ \text{subject to} & (a,B)\in\mathcal{U}. \end{array}$$

It is now easy to see that (4) is equivalent to (8).

To establish the implication, we show that

$$\sup_{x\in\mathcal{X}} \frac{x^T a}{\sqrt{x^T B x}} = \sup_{x\neq 0} \inf_{\lambda\in\mathcal{X}^*} \frac{x^T(a+\lambda)}{\sqrt{x^T B x}}. \tag{39}$$

First, suppose that $x \in \mathcal{X}$. Then, $\lambda^T x \geq 0$ for any $\lambda \in \mathcal{X}^*$ and $0 \in \mathcal{X}^*$, so $\inf_{\lambda\in\mathcal{X}^*} \lambda^T x = 0$. Thus,

$$\inf_{\lambda\in\mathcal{X}^*} \frac{x^T(a+\lambda)}{\sqrt{x^T B x}} = \frac{x^T a}{\sqrt{x^T B x}}.$$

Next, suppose that $x \notin \mathcal{X}\cup\{0\}$. Note from $\mathcal{X}^{**} = \mathcal{X}\cup\{0\}$ that there exists a nonzero $\bar{\lambda} \in \mathcal{X}^*$, with $\bar{\lambda}^T x < 0$. Then,

$$\inf_{\lambda\in\mathcal{X}^*} \frac{x^T(a+\lambda)}{\sqrt{x^T B x}} \leq \inf_{t>0} \left(\frac{x^T a}{\sqrt{x^T B x}} + \frac{t x^T \bar{\lambda}}{\sqrt{x^T B x}} \right) = -\infty \quad \forall x \notin \mathcal{X}\cup\{0\}.$$

When $\sup_{x\in\mathcal{X}} x^T a > 0$, we have from (39) that

$$\inf_{\lambda\in\mathcal{X}^*} \sup_{x\neq 0} \frac{x^T(a+\lambda)}{\sqrt{x^T B x}} > 0.$$

By the Cauchy–Schwarz inequality,

$$\inf_{\lambda\in\mathcal{X}^*} \sup_{x\neq 0} \frac{x^T(a+\lambda)}{\sqrt{x^T B x}} = \inf_{\lambda\in\mathcal{X}^*} \left[(a+\lambda)^T B^{-1}(a+\lambda)\right]^{1/2} = \left[\inf_{\lambda\in\mathcal{X}^*} (a+\lambda)^T B^{-1}(a+\lambda)\right]^{1/2}.$$

Since $(a+\lambda)^T B^{-1}(a+\lambda)$ is strictly concave in λ, we can see that there is $\lambda^\star$ such that

$$\inf_{\lambda\in\mathcal{X}^*} (a+\lambda)^T B^{-1}(a+\lambda) = (a+\lambda^\star)^T B^{-1}(a+\lambda^\star). \tag{40}$$

Then,

$$\sup_{x\neq 0} \frac{x^T(a+\lambda^\star)}{\sqrt{x^T B x}} = \inf_{\lambda\in\mathcal{X}^*} \sup_{x\neq 0} \frac{x^T(a+\lambda)}{\sqrt{x^T B x}}.$$

As will be seen soon, $x^\star = B^{-1}(a+\lambda^\star)$ satisfies

$$\frac{x^{\star T}(a+\lambda^\star)}{\sqrt{x^{\star T} B x^\star}} = \inf_{\lambda\in\mathcal{X}^*} \frac{x^{\star T}(a+\lambda)}{\sqrt{x^{\star T} B x^\star}}. \tag{41}$$

Therefore,

$$\sup_{x\neq 0}\inf_{\lambda\in\mathcal{X}^*}\frac{x^T(a+\lambda)}{\sqrt{x^TBx}}=\inf_{\lambda\in\mathcal{X}^*}\sup_{x\neq 0}\frac{x^T(a+\lambda)}{\sqrt{x^TBx}}.$$

Taken together, the results established above show that

$$\begin{aligned}\sup_{x\in\mathcal{X}}\frac{x^Ta}{\sqrt{x^TBx}}&=\sup_{x\neq 0}\inf_{\lambda\in\mathcal{X}^*}\frac{x^T(a+\lambda)}{\sqrt{x^TBx}}=\inf_{\lambda\in\mathcal{X}^*}\sup_{x\neq 0}\frac{x^T(a+\lambda)}{\sqrt{x^TBx}}\\&=\inf_{\lambda\in\mathcal{X}^*}\left[(a+\lambda)^TB^{-1}(a+\lambda)\right]^{1/2}.\end{aligned}$$

We complete the proof by establishing (41). To this end, we derive explicitly the optimality condition for $\lambda^\star$ to satisfy (40):

$$2x^{\star T}(\lambda-\lambda^\star)\geq 0\quad\forall\lambda\in\mathcal{X}^*, \tag{42}$$

with $x^\star=B^{\star-1}(a+\lambda^\star)$. (See [7, section 4.2.3].) We now show that $x^\star$ satisfies (41). To this end, we note that $\bar\lambda$ is optimal for (41) if and only if

$$\left\langle\nabla_\lambda\frac{\left(x^{\star T}(a+\lambda)\right)^2}{x^{\star T}Bx^\star}\bigg|_{\bar\lambda},\left(\lambda-\bar\lambda\right)\right\rangle\geq 0\quad\forall\lambda\in\mathcal{X}^*.$$

Here $\nabla_\lambda h(\lambda)|_{\bar\lambda}$ denotes the gradient of h at the point $\bar\lambda$. We can write the optimality condition as

$$2\frac{x^{\star T}\left(a+\bar\lambda\right)}{x^{\star T}\bar Bx^\star}x^{\star T}\left(\lambda-\bar\lambda\right)\geq 0\quad\forall\lambda\in\mathcal{X}^*.$$

Substituting $\bar\lambda=\lambda^\star$ and noting that $(a+\lambda)^{\star T}x^\star/x^{\star T}B^\star x^\star=1$, the optimality condition reduces to (42). Thus, we have shown that $\lambda^\star$ is optimal for (41).

A.2. Proof of Theorem 1. We will establish the following claims:

- $x^\star=B^{\star-1}(a^\star+\lambda^\star)\in\mathcal{X}$.
- $(x^\star,a^\star,\lambda^\star,B^\star)$ satisfies the saddle-point property
(43)
$$\frac{x^T(a^\star+\lambda^\star)}{\sqrt{x^TB^\star x}}\leq\frac{x^{\star T}(a^\star+\lambda^\star)}{\sqrt{x^{\star T}B^\star x^\star}}\leq\frac{x^{\star T}(a+\lambda)}{\sqrt{x^{\star T}Bx^\star}}\ \ \forall x\neq 0\ \ \forall\lambda\in\mathcal{X}^*\ \ \forall(a,B)\in\mathcal{U}.$$
- $x^\star$ and $\lambda^\star$ are orthogonal to each other:
(44)
$$x^{\star T}\lambda^\star=0.$$

The claims of Theorem 1 follow directly from the claims above. By definition of the dual cone, we have $\lambda^{\star T}x\geq 0$ for all $x\in\mathcal{X}$ and $0\in\mathcal{X}^*$. It follows from (43) and (44) that

$$\begin{aligned}\frac{x^Ta^\star}{\sqrt{x^TB^\star x}}&\leq\frac{x^T(a^\star+\lambda^\star)}{\sqrt{x^TB^\star x}}\leq\frac{x^{\star T}(a^\star+\lambda^\star)}{\sqrt{x^{\star T}B^\star x^\star}}=\frac{x^{\star T}a^\star}{\sqrt{x^{\star T}B^\star x^\star}}\\&\leq\frac{x^{\star T}a}{\sqrt{x^{\star T}Bx^\star}}\ \forall x\in\mathcal{X}\ \forall(a,B)\in\mathcal{U}.\end{aligned}$$

The saddle-point property (43) is equivalent to showing that

$$\sup_{x\neq 0} \frac{x^T(a^\star+\lambda^\star)}{\sqrt{x^TB^\star x}} = \frac{x^{\star T}(a^\star+\lambda^\star)}{\sqrt{x^{\star T}B^\star x^\star}} \tag{45}$$

and

$$\inf_{(a,B)\in\mathcal{U},\lambda\in\mathcal{X}^*} \frac{x^{\star T}(a+\lambda)}{\sqrt{x^{\star T}Bx^\star}} = \frac{x^{\star T}(a^\star+\lambda^\star)}{\sqrt{x^{\star T}B^\star x^\star}}. \tag{46}$$

Here, (45) follows from the Cauchy–Schwarz inequality.

We establish (46) by showing an equivalent claim

$$\inf_{(c,B)\in\mathcal{V}} \frac{x^{\star T}c}{\sqrt{x^{\star T}Bx^\star}} = \frac{x^{\star T}c^\star}{\sqrt{x^{\star T}B^\star x^\star}},$$

where

$$c^\star = a^\star+\lambda^\star, \qquad \mathcal{V} = \{(a+\lambda,B)\in\mathbb{R}^n\times\mathbb{S}^n_{++} \mid (a,B)\in\mathcal{U},\ \lambda\in\mathcal{X}^*\}.$$

The set $\mathcal{V}$ is closed and convex.

We know that $(c^\star,B^\star)$ is optimal for the convex problem

$$\begin{array}{ll}\text{minimize} & g(c,B)=c^TB^{-1}c\\ \text{subject to} & (c,B)\in\mathcal{V},\end{array} \tag{47}$$

with variables $c\in\mathbb{R}^n$ and $B=B^T\in\mathbb{R}^{n\times n}$. From the optimality condition of this problem that $(c^\star,B^\star)$ satisfies, we will prove that $(c^\star,B^\star)$ is also optimal for the problem

$$\begin{array}{ll}\text{minimize} & x^{\star T}c/\sqrt{x^{\star T}Bx^\star}\\ \text{subject to} & (c,B)\in\mathcal{V},\end{array} \tag{48}$$

with variables $c\in\mathbb{R}^n$ and $B=B^T\in\mathbb{R}^{n\times n}$. The proof is based on an extension of the arguments used to establish (41).

We derive explicitly the optimality condition for the convex problem (47). The pair $(c^\star,B^\star)$ must satisfy the optimality condition

$$\left\langle \nabla_c g(c,B)|_{(c^\star,B^\star)}, (c-c^\star)\right\rangle + \left\langle \nabla_B g(c,B)|_{(c^\star,B^\star)}, (B-B^\star)\right\rangle \geq 0 \quad \forall (c,B)\in\mathcal{V}$$

(see [7, section 4.2.3]). Here $(\nabla_c f(c,B)|_{(\bar c,\bar B)}, \nabla_B g(c,B)|_{(\bar c,\bar B)})$ denotes the gradient of f at the point (c,B). Using $\nabla_c(c^TB^{-1}c)=2B^{-1}c$, $\nabla_B(c^TB^{-1}c)=-B^{-1}cc^TB^{-1}$, and $\langle X,Y\rangle=\mathbf{Tr}(XY)$ for $X,Y\in\mathbb{S}^n$, where $\mathbf{Tr}$ denotes trace, we can express the optimality condition as

$$2c^{\star T}B^{\star -1}(c-c^\star) - \mathbf{Tr}\, B^{\star -1}c^\star c^{\star T}B^{\star -1}(B-B^\star)\geq 0 \quad \forall (c,B)\in\mathcal{V}$$

or equivalently

$$2x^{\star T}(c-c^\star) - x^{\star T}(B-B^\star)x^\star \geq 0 \quad \forall (c,B)\in\mathcal{V}, \tag{49}$$

with $x^\star = B^{\star -1}c^\star$.

To establish the optimality of $(c^\star, B^\star)$ for (48), we show that a solution of (48) is also a solution to the optimization problem

$$\begin{array}{ll} \text{minimize} & \left(x^{\star T} c\right)^2 / \left(x^{\star T} B x^\star\right) \\ \text{subject to} & (c, B) \in \mathcal{V}, \end{array} \tag{50}$$

with variables $c \in \mathbb{R}^n$ and $B = B^T \in \mathbb{R}^{n \times n}$ and vice versa. To show that (50) is a convex optimization problem, we must show that the objective is a convex function of c and B. To do so, we express the objective as the composition

$$\frac{\left(x^{\star T} c\right)^2}{x^{\star T} B x^\star} = g(H(c, B)),$$

where $g(u, t) = u^2/t$ and H is the function

$$H(c, B) = \left(x^{\star T} c, x^{\star T} B x^\star\right).$$

The function H is linear (as a mapping from c and B into $\mathbb{R}^2$), and the function g is convex (provided $t > 0$, which holds here). Thus, the composition f is a convex function of a and B. (See [7, section 3].)

This equivalence between (48) and (50) follows from

$$x^{\star T} c / \left(x^{\star T} B x^\star\right)^{1/2} > 0 \quad \forall (c, B) \in \mathcal{V},$$

which is a direct consequence of the optimality condition (49):

$$\begin{aligned} 2x^{\star T} c &\geq 2x^{\star T} c^\star + x^{\star T}(B - B^\star)x^\star \\ &= x^{\star T} c^\star + x^{\star T}(c^\star - B^\star x^\star) + x^{\star T} B x^\star \\ &= x^{\star T} B^{\star -1} x^\star + x^{\star T} B x^\star \\ &> 0 \quad \forall (c, B) \in \mathcal{V}. \end{aligned}$$

We now show that $(c^\star, B^\star)$ is optimal for (50) and hence for (48). The optimality condition for (50) is that a pair $(\bar{c}, \bar{B})$ is optimal for (50) if and only if

$$\left\langle \nabla_c \frac{\left(x^{\star T} c\right)^2}{x^{\star T} B x^\star}\bigg|_{(\bar{c}, \bar{B})}, (c - \bar{c}) \right\rangle + \left\langle \nabla_B \frac{\left(x^{\star T} c\right)^2}{x^{\star T} B x^\star}\bigg|_{(\bar{c}, \bar{B})}, \left(B - \bar{B}\right) \right\rangle \geq 0 \quad \forall (c, B) \in \mathcal{V}$$

(see [7, section 4.2.3]). Using

$$\nabla_c \frac{\left(x^{\star T} c\right)^2}{x^{\star T} B x^\star} = 2\frac{c^T x^\star}{x^{\star T} B x^\star} x^\star, \qquad \nabla_B \frac{\left(x^{\star T} c\right)^2}{x^{\star T} B x^\star} = -\frac{\left(c^T x^\star\right)^2}{\left(x^{\star T} B x^\star\right)^2} x^\star x^{\star T},$$

we can write the optimality condition as

$$\begin{aligned} &2\frac{x^{\star T}\bar{c}}{x^{\star T}\bar{B}x^\star} x^{\star T}(c - \bar{c}) - \mathbf{Tr}\, \frac{\left(x^{\star T}\bar{c}\right)^2}{\left(x^{\star T}\bar{B}x^\star\right)^2} x^\star x^{\star T}\left(B - \bar{B}\right) \\ &= 2\frac{x^{\star T}\bar{c}}{x^{\star T}\bar{B}x^\star} x^{\star T}(c - \bar{c}) - \frac{\left(x^{\star T}\bar{c}\right)^2}{\left(x^{\star T}\bar{B}x^\star\right)^2} x^{\star T}\left(B - \bar{B}\right) x^\star \\ &\geq 0 \quad \forall (c, B) \in \mathcal{V}. \end{aligned}$$

Substituting $\bar{c} = c^\star$, $\bar{B} = B^\star$, and noting that $c^{\star T}x^\star / x^{\star T}B^\star x^\star = 1$, the optimality condition reduces to

$$2x^{\star T}(c - c^\star) - x^{\star T}(B - B^\star)x^\star \geq 0 \quad \forall (c, B) \in \mathcal{V},$$

which is precisely (49). Thus, we have shown that $(c^\star, B^\star)$ is optimal for (50), which in turn means that it is also optimal for (48).

We next show by way of contradiction that $x^\star \in \mathcal{X}$. Suppose that $x^\star \notin \mathcal{X}$. Then, it follows from $\mathcal{X}^{**} = \mathcal{X} \cup \{0\}$ that there is $\bar{\lambda} \in \mathcal{X}^*$ such that $\bar{\lambda}^T x^\star < 0$. For any fixed $(\bar{a}, \bar{B})$ in $\mathcal{U}$, we can see from (43) (already established) that

$$\inf_{(a,B)\in\mathcal{U},\ \lambda\in\mathcal{X}^*} \frac{x^{\star T}(a+\lambda)}{\sqrt{x^{\star T}Bx^\star}} \leq \inf_{\lambda\in\mathcal{X}^*} \frac{x^{\star T}(\bar{a}+\lambda)}{\sqrt{x^{\star T}\bar{B}x^\star}} \leq \frac{x^{\star T}\bar{a}}{\sqrt{x^{\star T}\bar{B}x^\star}} + \inf_{t\geq 0} \frac{tx^{\star T}\bar{\lambda}}{\sqrt{x^{\star T}\bar{B}x^\star}} = -\infty.$$

However, this is contradictory to the fact that

$$\frac{x^{\star T}(a^\star + \lambda^\star)}{\sqrt{x^{\star T}B^\star x^\star}} = \inf_{(a,B)\in\mathcal{U},\ \lambda^*\in\mathcal{X}^*} \frac{x^{\star T}(a+\lambda)}{\sqrt{x^{\star T}Bx^\star}}$$

must be finite.

We complete the proof by showing $\lambda^{\star T}x^\star = 0$. Since $0 \in \mathcal{X}^*$, the saddle-point property (43) implies that

$$\frac{x^{\star T}(a^\star + \lambda^\star)}{\sqrt{x^{\star T}B^\star x^\star}} \leq \frac{x^{\star T}a^\star}{\sqrt{x^{\star T}B^\star x^\star}},$$

which means $x^{\star T}\lambda^\star \leq 0$. Since $\lambda \in \mathcal{X}^*$ and $x^\star \in \mathcal{X}$, we also have $x^{\star T}\lambda^\star \geq 0$.

A.3. Proof of Proposition 2. Let γ be the optimal value of (3):

$$\gamma = \sup_{x\in\mathcal{X}} \inf_{(a,B)\in\mathcal{U}} \frac{x^T a}{\sqrt{x^T Bx}}. \tag{51}$$

We can see that for any $x \in \mathcal{X}$, the set $X = \{(\sqrt{x^T Bx}, x^T a) \mid (a, B) \in \mathcal{U}\}$ cannot lie entirely above the line $r = \gamma\sigma$ in the (σ, r) space.

Using the Cauchy–Schwarz inequality, we can show that for any nonzero x and y

$$\frac{1}{2}\left(\sqrt{x^T Bx} + \sqrt{y^T By}\right) \geq \left(\left(\frac{x+y}{2}\right)^T B \left(\frac{x+y}{2}\right)\right)^{1/2}. \tag{52}$$

Here equality holds if and only if x and y are linearly dependent.

Suppose that there are two solutions $x^\star$ and $y^\star$ which are not linearly dependent. Then, the two sets

$$X = \left\{\left(\sqrt{x^{\star T}Bx^\star}, x^{\star T}a\right) \;\middle|\; (a,B) \in \mathcal{U}\right\}, \quad Y = \left\{\left(\sqrt{y^{\star T}By^\star}, y^{\star T}a\right) \;\middle|\; (a,B) \in \mathcal{U}\right\}$$

lie on and above, but cannot lie entirely above, the line $r = \gamma\sigma$ in the (σ, r) space. If $x^\star$ and $y^\star$ are not linearly dependent, then it follows from (52) and the compactness of $\mathcal{U}$ that the set $Z = \{(\sqrt{z^{\star T}Bz^\star}, z^{\star T}a) \mid (a, B) \in \mathcal{U}\}$, with $z^\star = (x^\star + y^\star)/2$, lies entirely above the line $r = \gamma\sigma$. Therefore, we have

$$\inf_{(a,B)\in\mathcal{U}} \frac{z^{\star T}a}{\sqrt{z^{\star T}Bz^\star}} > \gamma,$$

which is contradictory to the definition of γ given in (51).

Acknowledgments. The authors thank Alessandro Magnani, Almir Mutapcic, and Young-Han Kim for helpful comments and suggestions.

REFERENCES

[1] S. Benson and Y. Ye, *DSDP*5: *Software for Semidefinite Programming*, http://www-unix.mcs.anl.gov/DSDP/(2005).
[2] D. Bertsekas, A. Nedić, and A. Ozdaglar, *Convex Analysis and Optimization*, Athena Scientific, Cambridge, MA, 2003.
[3] D. Bertsimas and I. Popescu, *Optimal inequalities in probability theory: A convex optimization approach*, SIAM J. Optim., 15 (2005), pp. 780–804.
[4] M. Best and P. Grauer, *On the sensitivity of mean-variance-efficient portfolios to changes in asset means: Some analytical and computational results*, Rev. Financ. Stud., 4 (1991), pp. 315–342.
[5] F. Black and R. Litterman, *Global portfolio optimization*, Financ. Analysts J., 48 (1992), pp. 28–43.
[6] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM, Philadelphia, 1994.
[7] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, 2004.
[8] J. Cochrane, *Asset Pricing*, 2nd ed., Princeton University Press, Princeton, NJ, 2001.
[9] O. Costa and J. Paiva, *Robust portfolio selection using linear matrix inequalities*, J. Econom. Dynam. Control, 26 (2002), pp. 889–909.
[10] L. E. Ghaoui, M. Oks, and F. Oustry, *Worst-case value-at-risk and robust portfolio optimization: A conic programming approach*, Oper. Res., 51 (2003), pp. 543–556.
[11] D. Goldfarb and G. Iyengar, *Robust portfolio selection problems*, Math. Oper. Res., 28 (2003), pp. 1–38.
[12] B. Halldórsson and R. Tütüncü, *An interior-point method for a class of saddle point problems*, J. Optim. Theory Appl., 116 (2003), pp. 559–590.
[13] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer Ser. Statist., Springer-Verlag, New York, 2001.
[14] P. Jorion, *Bayes-Stein estimation for portfolio analysis*, J. Financ. Quant. Anal., 21 (1979), pp. 279–292.
[15] S. Kassam and V. Poor, *Robust techniques for signal processing: A survey*, Proc. IEEE, 73 (1985), pp. 433–481.
[16] S.-J. Kim and S. Boyd, *Two-fund separation under model mis-specification*, manuscript. Available from http://www.stanford.edu/~boyd/rob_two_fund_sep.html.
[17] S.-J. Kim, A. Magnani, and S. Boyd, *Robust Fisher discriminant analysis*, in Adv. Neural Inform. Process. Syst., MIT Press, Cambridge, MA, 2006.
[18] M. Lobo and S. Boyd, *The worst-case risk of a portfolio*, manuscript. Available from http://faculty.fuqua.duke.edu/2000.
[19] D. Luenberger, *Investment Science*, Oxford University Press, New York, 1998.
[20] H. Markowitz, *Portfolio selection*, J. Finance, 7 (1952), pp. 77–91.
[21] R. Michaud, *The Markowitz optimization enigma: Is 'optimized' optimal?*, Financ. Analysts J., 45 (1989), pp. 31–42.
[22] A. Mutapcic, S.-J. Kim, and S. Boyd, *Array signal processing with robust rejection constraints via second-order cone programming*, in Proceedings of the 40th Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, 2006.
[23] A. Roy, *Safety first and the holding of assets*, Econometrica, 20 (1952), pp. 413–449.
[24] W. Sharpe, *The Sharpe ratio*, J. Portfolio Manag., 21 (1994), pp. 49–58.
[25] M. Sion, *On general minimax theorems*, Pacific J. Math., 8 (1958), pp. 171–176.
[26] J. Sturm, *Using SEDUMI* 1.02*, a Matlab Toolbox for Optimization Over Symmetric Cones*, http://www.fewcal.kub.nl/sturm/software/sedumi.html/(2001).
[27] K. Toh, R. H. Tütüncü, and M. Todd, *SDPT*3 *version* 3.02*. A Matlab software for semidefinite-quadratic-linear programming*, http://www.math.nus.edu.sg/~mattohkc/sdpt3.html/ (2002).
[28] H. Van Trees, *Detection, Estimation, and Modulation Theory, Part* I, Wiley-Interscience, New York, 2001.
[29] L. Vandenberghe and S. Boyd, *Semidefinite programming*, SIAM Rev., 38 (1996), pp. 49–95.
[30] L. Vandenberghe, S. Boyd, and K. Comanor, *Generalized Chebyshev bounds via semidefinite programming*, SIAM Rev., 49 (2007), pp. 52–64.
[31] S. Verdú and H. Poor, *On minimax robustness: A general approach and applications*, IEEE Trans. Inform. Theory, 30 (1984), pp. 328–340.

SIAM J. OPTIM.
Vol. 19, No. 3, pp. 1368–1391

TWO ALGORITHMS FOR THE MINIMUM ENCLOSING BALL PROBLEM*

E. ALPER YILDIRIM†

Abstract. Given $\mathcal{A} := \{a^1, \ldots, a^m\} \subset \mathbb{R}^n$ and $\epsilon > 0$, we propose and analyze two algorithms for the problem of computing a $(1+\epsilon)$-approximation to the radius of the minimum enclosing ball of $\mathcal{A}$. The first algorithm is closely related to the Frank–Wolfe algorithm with a proper initialization applied to the dual formulation of the minimum enclosing ball problem. We establish that this algorithm converges in $O(1/\epsilon)$ iterations with an overall complexity bound of $O(mn/\epsilon)$ arithmetic operations. In addition, the algorithm returns a "core set" of size $O(1/\epsilon)$, which is independent of both m and n. The latter algorithm is obtained by incorporating "away" steps into the former one at each iteration and achieves the same asymptotic complexity bound as the first one. While the asymptotic bound on the size of the core set returned by the second algorithm also remains the same as the first one, the latter algorithm has the potential to compute even smaller core sets in practice, since, in contrast to the former one, it allows "dropping" points from the working core set at each iteration. Our analysis reveals that the leading terms in the asymptotic complexity analysis are reasonably small. In contrast to the first algorithm, we also establish that the second algorithm asymptotically exhibits linear convergence, which provides further insight into our computational results, indicating that the latter algorithm indeed terminates faster with smaller core sets in comparison with the first one. We also discuss how our algorithms can be extended to compute an approximation to the minimum enclosing ball of more general input sets without sacrificing the iteration complexity and the bound on the core set size. In particular, we establish the existence of a core set of size $O(1/\epsilon)$ for a much wider class of input sets. We adopt the real number model of computation in our analysis.

Key words. minimum enclosing balls, core sets, approximation algorithms

AMS subject classifications. 90C25, 90C46, 65K05

DOI. 10.1137/070690419

1. Introduction.

Given a finite set of points $\mathcal{A} := \{a^1, \ldots, a^m\} \subset \mathbb{R}^n$, we are concerned with the problem of computing an approximation to the minimum enclosing ball of $\mathcal{A}$, which we shall denote by MEB($\mathcal{A}$).

For $c \in \mathbb{R}^n$ and a nonnegative $\rho \in \mathbb{R}$, let $\mathcal{B}_{c,\rho} \subset \mathbb{R}^n$ denote the ball centered at c with radius ρ, i.e.,

$$\mathcal{B}_{c,\rho} := \{x \in \mathbb{R}^n : \|x - c\| \leq \rho\},$$

where $\|\cdot\|$ denotes the Euclidean norm.

Given $\epsilon > 0$, a ball $\mathcal{B}_{c,\rho}$ is said to be a $(1+\epsilon)$-approximation to MEB($\mathcal{A}$) if

$$\mathcal{A} \subset \mathcal{B}_{c,\rho}, \quad \rho \leq (1+\epsilon)\rho_{\mathcal{A}}, \tag{1}$$

where $\mathcal{B}_{c_{\mathcal{A}},\rho_{\mathcal{A}}} := \text{MEB}(\mathcal{A})$.

A subset $\mathcal{X} \subseteq \mathcal{A}$ is said to be an ϵ-core set (or a core set) of $\mathcal{A}$ if

$$\rho_{\mathcal{X}} \leq \rho_{\mathcal{A}} \leq (1+\epsilon)\rho_{\mathcal{X}}, \tag{2}$$

where $\mathcal{B}_{c_{\mathcal{X}},\rho_{\mathcal{X}}} := \text{MEB}(\mathcal{X})$. Small core sets play an important role in designing efficient algorithms for large-scale problems, since they provide a compact representation of

*Received by the editors May 3, 2007; accepted for publication (in revised form) May 22, 2008; published electronically November 21, 2008.
http://www.siam.org/journals/siopt/19-3/69041.html
†Department of Industrial Engineering, Bilkent University, 06800 Bilkent, Ankara, Turkey (yildirim@bilkent.edu.tr). This research was partially supported by TÜBİTAK (Turkish Scientific and Technological Research Council) Grant 107M411.

the input set $\mathcal{A}$. If a small ϵ-core set $\mathcal{X}$ is available, then solving the problem on $\mathcal{X}$ already yields a good approximation to MEB($\mathcal{A}$). Since the center $c_{\mathcal{A}}$ of MEB($\mathcal{A}$) lies in the convex hull of $\mathcal{A}$ (cf. section 2), it follows from Carathéodory's theorem that there always exists a 0-core set of size at most $n + 1$.

Minimum enclosing balls have numerous important applications in clustering, nearest neighbor search, data classification, support vector machines, machine learning, facility location, collision detection, computer graphics, and military operations. We refer the reader to [24] and the references therein. In particular, many of these applications give rise to large-scale instances of the MEB problem, and a reasonably small accuracy suffices for such applications.

The minimum enclosing ball problem has a fairly rich literature dating back to at least the 19th century [37]. One of the earliest known solution methods is given by Sylvester [38], which is attributed to Peirce, and later rediscovered by Chrystal [9]. The reader is referred to [6] for a detailed account of the earlier history of this problem. More recent references include [25, 14, 28, 11, 8, 35, 7, 23, 22, 27, 31, 40, 16, 17, 4, 2, 24, 42, 13, 12, 29, 30, 21, 44, 32].

The earliest known algorithm due to Chrystal and Peirce [38, 9] computes the exact minimum enclosing ball of m points in the plane in $O(m^2)$ operations in the worst case. For a fixed dimension n, the minimum enclosing ball of m points can be computed in $O(m)$ operations [27, 40]. However, the dependence on the dimension n is exponential. Bădoiu, Har-Peled, and Indyk [4] established the existence of an ϵ-core set of size $O(1/\epsilon^2)$. Note that the size of the core set is independent of m and n. Based on this result, their algorithm can compute a $(1+\epsilon)$-approximation to MEB($\mathcal{A}$) in $O\left(mn/\epsilon^2 + (1/\epsilon^{10})\log(1/\epsilon)\right)$ operations. Bădoiu and Clarkson [2] and Kumar, Mitchell, and Yıldırım [24] independently discovered the existence of an ϵ-core set of size $O(1/\epsilon)$. As noted in [2], this improved core set result can be combined with the algorithm of [4] to obtain an improved running time of $O\left(mn/\epsilon + 1/\epsilon^5\right)$. The algorithm of [24] achieves a slightly improved complexity bound of $O\left(mn/\epsilon + (1/\epsilon^{4.5})\log(1/\epsilon)\right)$ using second-order cone programming combined with column generation. In addition, Bădoiu and Clarkson [2] proposed another simple algorithm that computes a $(1 + \epsilon)$-approximation in $O(mn/\epsilon^2)$ operations. In another paper, the same authors established a tight upper bound of $\lceil 1/\epsilon \rceil$ on the size of an ϵ-core set [3]. However, their construction is based on the assumption that $n \geq \lfloor 1/\epsilon \rfloor$. The algorithm of Panigrahy [33] computes a $(1 + \epsilon)$-approximation in $O(mn/\epsilon)$ operations. Note that this algorithm has the best known dependence on ϵ, and each of these algorithms is polynomial for fixed ϵ. If ϵ is viewed as part of the input data, the minimum enclosing ball can be formulated as an instance of convex programming problem and can be solved using the ellipsoid method in $O\left(n^3 m \log(1/\epsilon)\right)$ operations [19]. Alternatively, interior-point methods yield an overall complexity bound of $O\left(n^2 m^{3/2} \log(1/\epsilon)\right)$ operations if the problem is formulated as an instance of second-order cone programming [24].

In this paper, we focus on large-scale instances of the minimum enclosing ball problem for which a reasonably small value of ϵ is satisfactory. Throughout this paper, we adopt the real number model of computation [5], i.e., we assume that arithmetic operations with real numbers and comparisons can be done at unit cost. We propose and analyze two algorithms that compute a $(1 + \epsilon)$-approximation to MEB($\mathcal{A}$) for a given $\epsilon > 0$. Our first algorithm is closely related to the Frank–Wolfe algorithm [15] applied to the dual formulation of the problem. At each iteration, the algorithm can only add points to the working core set. The second algorithm is obtained by incorporating "away" steps into each iteration of the first one (see, e.g., [41, 20]). As such, the

latter algorithm has the potential to compute a smaller core set than the former one, since it allows "dropping" points from the working core set at each iteration. A similar algorithm has recently been proposed for the minimum-volume enclosing ellipsoid problem [39]. Both of our algorithms compute a $(1+\epsilon)$-approximation to $\mathrm{MEB}(\mathcal{A})$ in $O(mn/\epsilon)$ operations, which matches the currently best known dependence on ϵ. In addition, each algorithm explicitly computes an ϵ-core set of size $O(1/\epsilon)$. Our analysis reveals that the leading terms in the complexity analysis are reasonably small, which further contributes to the efficiency of our algorithms. Furthermore, we establish that the second algorithm asymptotically exhibits linear convergence. Our computational results indicate that the sizes of the core sets returned by our algorithms are generally much smaller than the corresponding worst-case estimates. Furthermore, as expected, the latter algorithm almost always outperforms the former one both in terms of the running time and the core set size.

We also discuss how our algorithms can be extended to compute an approximate minimum enclosing ball of more general input sets. In particular, we establish that the asymptotic core set size of $O(1/\epsilon)$ extends to a much larger class of input sets.

We first compare our algorithms with the one proposed by Panigrahy [33], which computes a $(1+\epsilon)$-approximation to the minimum enclosing ball of a finite set of points in $O(mn/\epsilon)$ arithmetic operations. Panigrahy's algorithm starts with a ball whose radius is known to be smaller than that of the minimum enclosing ball and maintains an upper bound ζ on the difference between these two radii. At each iteration, the algorithm moves the current ball toward the furthest point from the center until the ball touches that particular point without changing the radius of the ball. After repeating such iterations $O(1/\zeta)$ times, the algorithm either provides a certificate that an approximate solution has been computed or decides that either the radius can be increased or the error bound ζ can be decreased. The whole procedure is then repeated using the new parameters for the radius and the error bound. Similarly to Panigrahy's algorithm, each of our algorithms also constructs a sequence of balls, and our first algorithm moves the center toward the furthest point from the center of the current ball at each iteration. However, the center moves by only a fraction of this distance. Furthermore, the second algorithm also allows us to move the current center away from the closest point in the working core set. Unlike Panigrahy's algorithm, our algorithms construct balls of strictly increasing radii in each iteration, and the radius and the error bound are updated at each iteration. While Panigrahy's algorithm checks the termination criterion after each set of $O(1/\zeta)$ iterations, our algorithms employ a simpler termination criterion in each iteration. This strategy has the potential advantage of earlier termination than that predicted by the theoretical worst-case estimate. Finally, while Panigrahy exclusively works with an input set of finite points with more general enclosing shapes, our algorithms can easily be modified to compute an approximation of the minimum enclosing ball of a much wider class of input sets without sacrificing the core set bound of $O(1/\epsilon)$.

After the first version of this manuscript had been submitted, Clarkson [10] announced several results concerning the convergence properties of the Frank–Wolfe algorithm, which is the main ingredient in both of our algorithms. He studied the problem of maximizing a general concave function over the unit simplex, of which the dual formulation of the minimum enclosing problem is a special case. By giving a general definition of an *additive* ϵ-core set, he established core set results for several variants of the Frank–Wolfe algorithm in a more general setting. Due to the special structure of the objective function in the dual formulation of the minimum enclosing

ball problem, his additive core set definition almost matches with our multiplicative core set definition given by (2). He presented an improved complexity bound of $O(mn/\epsilon)$ for a slightly modified version of the algorithm of [2], which matches the complexity bounds of our algorithms.

On the other hand, he establishes an ϵ-core set size of $O(1/\epsilon^2)$ for the general problem using his Algorithm 1.1, which, apart from the choice of the initial solutions, coincides with our first algorithm that computes a core set of size $O(1/\epsilon)$. He proposes a more sophisticated algorithm (cf. Algorithm 4.2 in [10]), which requires the computation of the optimal solution of a sequence of subproblems restricted to the smaller faces of the unit simplex, to establish the improved core set result of $O(1/\epsilon)$. In addition, he also studies a variant of the Frank–Wolfe algorithm that uses "away" steps (cf. Algorithm 5.1), for which he establishes a core set size of $O(1/\epsilon)$. However, this algorithm differs from our second algorithm, since it again requires the computation of the optimal solution of a sequence of subproblems. In contrast, we establish a core set size of $O(1/\epsilon)$ using much simpler (and lazier) algorithms. We derive explicit small constants inside the asymptotic bounds. Finally, we extend each of our algorithms to much more general input sets and establish the existence of core sets of size $O(1/\epsilon)$. Therefore, while Clarkson's results apply to a more general class of problems, we achieve the same or stronger results using simpler algorithms for the special case of the minimum enclosing ball problem, but we allow more general input sets.

This paper is organized as follows. In the remainder of this section, we define our notation. In section 2, we discuss optimization formulations for the minimum enclosing ball problem. Section 3 presents our first algorithm. The second algorithm is the topic of section 4. We discuss the extensions of our algorithms in section 5. The computational results are presented in section 6. Finally, we conclude the paper with some future research directions in section 7.

1.1. Notation. Vectors are denoted by lowercase Roman letters. For a vector p, p_i denotes its ith component. Inequalities on vectors apply to each component. We reserve e^j for the jth unit vector. Uppercase Roman letters are reserved for matrices. We use $\log(\cdot)$ to denote the natural logarithm. Functions and operators are denoted by uppercase Greek letters. Scalars except for m and n are represented by lowercase Greek letters unless they represent components of a vector or elements of a sequence of scalars, vectors, or matrices. We reserve i, j, and k for such indexing purposes. Uppercase script letters are used for all other objects such as sets, balls, and ellipsoids.

2. Optimization formulations. In this section, we review the optimization formulations of the minimum enclosing ball problem. We remark that most of the material of this section already appears in the earlier literature (see, e.g., [11, 26]). Some results and proofs are included for the sake of completeness.

Let $\mathcal{A} := \{a^1, \ldots, a^m\} \subset \mathbb{R}^n$. The minimum enclosing ball of $\mathcal{A}$ can be computed by solving the following optimization problem:

$$(\mathcal{P}_1) \quad \begin{array}{ll} \min_{c,\rho} & \rho \\ \text{subject to} & \\ & \left\|a^i - c\right\| \le \rho, \quad i = 1, \ldots, m, \end{array}$$

where $c \in \mathbb{R}^n$ and $\rho \in \mathbb{R}$ are the decision variables. By squaring the constraints and defining $\gamma := \rho^2$, $(\mathcal{P}_1)$ can be converted into the following optimization problem with

smooth, convex quadratic constraints:

$$\begin{array}{lll} (\mathcal{P}_2) & \min_{c,\gamma} & \gamma \\ & \text{subject to} & \\ & & (a^i)^T a^i - 2(a^i)^T c + c^T c \leq \gamma, \quad i = 1, \ldots, m. \end{array}$$

The Lagrangian dual of $(\mathcal{P}_2)$ is given by

$$\begin{array}{lll} (\mathcal{D}) & \max_u & \Phi(u) := \sum_{i=1}^m u_i (a^i)^T a^i - \left(\sum_{i=1}^m u_i a^i \right)^T \left(\sum_{i=1}^m u_i a^i \right) \\ & \text{subject to} & \\ & & \sum_{i=1}^m u_i = 1, \\ & & u_i \geq 0, \quad i = 1, \ldots, m, \end{array}$$

where $u \in \mathbb{R}^m$ is the decision variable.

Since $(\mathcal{P}_2)$ is a concave maximization problem with linear constraints, it follows from the Karush–Kuhn–Tucker optimality conditions that $(c_{\mathcal{A}}, \gamma_{\mathcal{A}}) \in \mathbb{R}^n \times \mathbb{R}$ is an optimal solution of $(\mathcal{P}_2)$ if and only if there exists $u^* \in \mathbb{R}^m$ such that

$$\sum_{i=1}^m u_i^* = 1, \tag{3a}$$

$$c_{\mathcal{A}} = \sum_{i=1}^m u_i^* a^i, \tag{3b}$$

$$(a^i)^T a^i - 2(a^i)^T c_{\mathcal{A}} + (c_{\mathcal{A}})^T c_{\mathcal{A}} \leq \gamma_{\mathcal{A}}, \quad i = 1, \ldots, m, \tag{3c}$$

$$u_i^* \left((a^i)^T a^i - 2(a^i)^T c_{\mathcal{A}} + (c_{\mathcal{A}})^T c_{\mathcal{A}} - \gamma_{\mathcal{A}} \right) = 0, \quad i = 1, \ldots, m, \tag{3d}$$

$$u^* \geq 0. \tag{3e}$$

A simple manipulation of the optimality conditions reveals that

$$\gamma_{\mathcal{A}} = \Phi(u^*), \tag{4}$$

which implies that $u^* \in \mathbb{R}^m$ is an optimal solution of $(\mathcal{D})$ and that strong duality holds between $(\mathcal{P}_2)$ and $(\mathcal{D})$. Note that the center $c_{\mathcal{A}}$ of the minimum enclosing ball of $\mathcal{A}$ is given by a convex combination of the elements of $\mathcal{A}$ by (3b). In addition, it follows from (3d) that only the components of u^* corresponding to the points on the boundary of MEB($\mathcal{A}$) can have a positive value.

LEMMA 2.1. *Let $\mathcal{A} = \{a^1, \ldots, a^m\}$. The minimum enclosing ball of $\mathcal{A}$ exists and is unique. Let $u^* \in \mathbb{R}^m$ denote the optimal solution of $(\mathcal{D})$. Then, MEB($\mathcal{A}$) $= \mathcal{B}_{c_{\mathcal{A}}, \rho_{\mathcal{A}}}$, where*

$$c_{\mathcal{A}} = \sum_{i=1}^m u_i^* a^i, \quad \rho_{\mathcal{A}} = \sqrt{\Phi(u^*)}. \tag{5}$$

Proof. Note that $\mathcal{A} \subset \mathcal{B}_{0,\rho^u}$, where $\rho^u := \max_{i=1,\ldots,m} \left\| a^i \right\|$. By adding the redundant constraint $\gamma \leq (\rho^u)^2$ to $(\mathcal{P}_2)$, the feasible region becomes a closed and bounded set and the objective function is continuous, which establishes the existence of MEB($\mathcal{A}$). If there were two different minimum enclosing balls, one can then construct a ball of smaller radius that encloses the intersection of the two balls and

hence also $\mathcal{A}$, which is a contradiction. The relationships (5) directly follow from the discussions preceding the lemma. □

By Lemma 2.1, MEB($\mathcal{A}$) can be computed by solving the dual problem ($\mathcal{D}$), which will be the basis of both of our algorithms in this paper. We close this section by the following technical result, which will play an important role in finding a good initial feasible solution in our algorithms. The reader is referred to [18, 4, 12] for the proof of this result.

LEMMA 2.2. *Let $\mathcal{A} = \{a^1, \ldots, a^m\}$, and let MEB($\mathcal{A}$) $= \mathcal{B}_{c_\mathcal{A}, \rho_\mathcal{A}}$. Then, any closed half-space that contains $c_\mathcal{A}$ also contains at least one point $a^j \in \mathcal{A}$ such that $\|a^j - c_\mathcal{A}\| = \rho_\mathcal{A}$.*

3. The first algorithm. Given $\mathcal{A} := \{a^1, \ldots, a^m\} \subset \mathbb{R}^n$ and $\epsilon > 0$, we present our first algorithm that computes a $(1+\epsilon)$-approximation to MEB($\mathcal{A}$) in this section.

Algorithm 3.1 The first algorithm that computes a $(1 + \epsilon)$-approximation to MEB($\mathcal{A}$).

Require: Input set of points $\mathcal{A} = \{a^1, \ldots, a^m\} \subset \mathbb{R}^n, \epsilon > 0$.

1: $\alpha \leftarrow \arg\max_{i=1,\ldots,m} \|a^i - a^1\|^2$, $\beta \leftarrow \arg\max_{i=1,\ldots,m} \|a^i - a^\alpha\|^2$;
2: $u_i^0 \leftarrow 0, \quad i = 1, \ldots, m$;
3: $u_\alpha^0 \leftarrow 1/2, \quad u_\beta^0 \leftarrow 1/2$;
4: $\mathcal{X}_0 \leftarrow \{a^\alpha, a^\beta\}$;
5: $c^0 \leftarrow \sum_{i=1}^m u_i^0 a^i$;
6: $\gamma^0 \leftarrow \Phi(u^0)$;
7: $\kappa \leftarrow \arg\max_{i=1,\ldots,m} \|a^i - c^0\|^2$;
8: $\delta_0 \leftarrow \left(\|a^\kappa - c^0\|^2 / \gamma^0\right) - 1$;
9: $k \leftarrow 0$;
10: While $\delta_k > (1+\epsilon)^2 - 1$, do
11: **loop**
12: $\lambda^k \leftarrow \delta_k / [2(1+\delta_k)]$;
13: $k \leftarrow k + 1$;
14: $u^k \leftarrow (1 - \lambda^{k-1}) u^{k-1} + \lambda^{k-1} e^\kappa$;
15: $c^k \leftarrow (1 - \lambda^{k-1}) c^{k-1} + \lambda^{k-1} a^\kappa$;
16: $\mathcal{X}_k \leftarrow \mathcal{X}_{k-1} \cup \{a^\kappa\}$;
17: $\gamma^k \leftarrow \Phi(u^k)$;
18: $\kappa \leftarrow \arg\max_{i=1,\ldots,m} \|a^i - c^k\|^2$;
19: $\delta_k \leftarrow \left(\|a^\kappa - c^k\|^2 / \gamma^k\right) - 1$;
20: **end loop**
21: **Output** $c^k, \mathcal{X}_k, u^k, \sqrt{(1+\delta_k)\gamma^k}$.

We now describe Algorithm 3.1 in more detail. In step 1, the algorithm computes the furthest point $a^\alpha \in \mathcal{A}$ from $a^1 \in \mathcal{A}$ and then computes the furthest point $a^\beta \in \mathcal{A}$ from a^α. Steps 2 and 3 initialize the vector $u^0 \in \mathbb{R}^m$. Note that u^0 is a feasible solution of the dual problem ($\mathcal{D}$). The core set $\mathcal{X}_0$ is initialized at step 4. At each iteration, the algorithm implicitly constructs a "trial" ball with center c^k and radius $(\gamma^k)^{1/2}$. By Lemma 2.1, this ball coincides with MEB($\mathcal{A}$) if and only if u^k is an optimal solution of ($\mathcal{D}$). Otherwise, at least one point in $\mathcal{A}$ lies outside of this ball. Note that δ_k satisfies $\|a^\kappa - c^k\|^2 = (1+\delta_k)\gamma^k$, where $a^\kappa \in \mathcal{A}$ is the furthest point from c^k. It follows that the trial ball encloses $\mathcal{A}$ if its radius is expanded by a factor

of $(1+\delta_k)^{1/2}$, i.e., $\Phi(u^k) \le \Phi(u^*) \le (1+\delta_k)\Phi(u^k)$. Unless the termination criterion is satisfied, the new center c^{k+1} is computed by shifting c^k toward the furthest point a^κ, which is added to the working core set $\mathcal{X}_{k+1}$, and u^{k+1} is updated accordingly to ensure that dual feasibility is maintained. The algorithm continues in an iterative manner by computing a new trial ball corresponding to u^{k+1}.

Algorithm 3.1 is the adaptation of the Frank–Wolfe algorithm to the dual problem $(\mathcal{D})$. At each iteration, the quadratic objective function $\Phi(u)$ of $(\mathcal{D})$ is linearized at the current feasible solution u^k. Since the feasible region of $(\mathcal{D})$ is the unit simplex, the unit vector e^κ, where κ is the index of the furthest point in $\mathcal{A}$ from c^k, solves the linearized subproblem. It is easy to verify that

$$\lambda^k = \arg\max_{\lambda\in[0,1]} \Phi\left((1-\lambda)u^k + \lambda e^\kappa\right).$$

We remark that Algorithm 3.1 uses only the first-order approximation to the objective function Φ. As such, each iteration is fairly cheap, but the number of iterations is usually significantly higher than other algorithms that use second-order information such as interior-point methods. However, such general-purpose algorithms become computationally infeasible for larger problems, since each iteration is usually much more expensive. This observation provides one of our motivations to develop a specialized algorithm for this problem.

3.1. Analysis of the first algorithm. This subsection is devoted to the analysis of Algorithm 3.1.

LEMMA 3.1. *$u^0 \in \mathbb{R}^m$ satisfies $\gamma^0 = \Phi(u^0) \ge (1/3)\Phi(u^*) = (1/3)\gamma_{\mathcal{A}}$, where $u^* \in \mathbb{R}^m$ and $\gamma_{\mathcal{A}}$ are the optimal solution and the optimal value of $(\mathcal{D})$, respectively. Furthermore, $\delta_0 \le 8$.*

Proof. For any vectors $y, z \in \mathbb{R}^n$ and any $\varphi \in \mathbb{R}$, it is easy to verify that

$$\|(1-\varphi)y + \varphi z\|^2 = (1-\varphi)\|y\|^2 + \varphi\|z\|^2 - \varphi(1-\varphi)\|y-z\|^2. \tag{6}$$

Note that

$$\Phi(u^0) = (1/2)\left\|a^\alpha\right\|^2 + (1/2)\left\|a^\beta\right\|^2 - \left\|(1/2)\left(a^\alpha + a^\beta\right)\right\|^2 = (1/4)\left\|a^\alpha - a^\beta\right\|^2, \tag{7}$$

where we used (6) to derive the second equality. The proof is based on establishing that at least one of a^α and a^β is sufficiently away from the center $c_{\mathcal{A}}$ of $\mathrm{MEB}(\mathcal{A})$.

First, suppose that $\left\|a^1 - c_{\mathcal{A}}\right\| \ge (1/\sqrt{3})\rho_{\mathcal{A}}$, where $\rho_{\mathcal{A}}$ is the radius of $\mathrm{MEB}(\mathcal{A})$. Let $\mathcal{H}$ be the hyperplane passing through $c_{\mathcal{A}}$ that is perpendicular to $a^1 - c_{\mathcal{A}}$. Let $\mathcal{H}_+$ denote the closed half-space whose boundary is $\mathcal{H}$ and which does not contain a^1. By Lemma 2.2, $\mathcal{H}_+$ contains a point $a^j \in \mathcal{A}$ such that $\left\|a^j - c_{\mathcal{A}}\right\| = \rho_{\mathcal{A}}$. Therefore, $\left\|a^\alpha - a^1\right\|^2 \ge \left\|a^1 - c_{\mathcal{A}}\right\|^2 + (\rho_{\mathcal{A}})^2 \ge (4/3)\gamma_{\mathcal{A}}$, where $\gamma_{\mathcal{A}} = \Phi(u^*) = (\rho_{\mathcal{A}})^2$ is the optimal value of $(\mathcal{D})$. It follows from (7) that

$$\Phi(u^0) = (1/4)\left\|a^\beta - a^\alpha\right\|^2 \ge (1/4)\left\|a^1 - a^\alpha\right\|^2 \ge (1/3)\Phi(u^*).$$

Suppose now that $\left\|a^1 - c_{\mathcal{A}}\right\| = \theta\rho_{\mathcal{A}}$, where $\theta < 1/\sqrt{3}$. In this case, $\left\|a^1 - a^\alpha\right\| \le \left\|a^1 - c_{\mathcal{A}}\right\| + \|c_{\mathcal{A}} - a^\alpha\|$, which implies that

$$\|c_{\mathcal{A}} - a^\alpha\| \ge \left\|a^1 - a^\alpha\right\| - \left\|a^1 - c_{\mathcal{A}}\right\| \ge (1+\theta^2)^{1/2}\rho_{\mathcal{A}} - \theta\rho_{\mathcal{A}} = [(1+\theta^2)^{1/2} - \theta]\rho_{\mathcal{A}},$$

where we again invoked Lemma 2.2 to obtain a lower bound on $\left\|a^1 - a^\alpha\right\|$. Therefore, one more application of Lemma 2.2 yields

$$\begin{aligned}\Phi(u^0) &= (1/4)\left\|a^\beta - a^\alpha\right\|^2 \\ &\geq (1/4)\left(\left\|a^\alpha - c_{\mathcal{A}}\right\|^2 + (\rho_{\mathcal{A}})^2\right) \\ &\geq (1/4)\left(1+\theta^2+\theta^2 - 2\theta(1+\theta^2)^{1/2}+1\right)\gamma_{\mathcal{A}} \\ &= (1/2)\left(1+\theta^2-\theta(1+\theta^2)^{1/2}\right)\gamma_{\mathcal{A}}.\end{aligned}$$

It is easy to verify that $(1/2)\left(1+\theta^2-\theta(1+\theta^2)^{1/2}\right)$ is a decreasing function of θ. Since $\theta < 1/\sqrt{3}$, it follows that

$$\Phi(u^0) \geq (1/2)\,(1+1/3-2/3)\,\gamma_{\mathcal{A}} = (1/3)\Phi(u^*),$$

which completes the first part of the proof.

Let a^κ be the furthest point in $\mathcal{A}$ from $c^0 = (1/2)(a^\alpha + a^\beta)$. Then,

$$\begin{aligned}\left\|a^\kappa - c^0\right\| &\leq \left\|a^\kappa - a^\alpha\right\| + \left\|a^\alpha - c^0\right\|, \\ &\leq \left\|a^\beta - a^\alpha\right\| + (1/2)\left\|a^\beta - a^\alpha\right\| = (3/2)\left\|a^\beta - a^\alpha\right\|,\end{aligned}$$

where we used the definition of c^0 and the fact that a^β is the furthest point in $\mathcal{A}$ from a^α to derive the second inequality. Therefore, $\delta_0 = (\|a^\kappa - c^0\|^2/\gamma^0) - 1 \leq [4(9/4)(\gamma^0/\gamma^0)] - 1 = 8$, where we used (7). The second part of the assertion follows. □

Lemma 3.1 establishes several properties of the initial feasible solution $u^0 \in \mathbb{R}^m$. The next lemma relates the dual objective function values evaluated at the successive iterates generated by Algorithm 3.1.

LEMMA 3.2. *For each $k = 0, 1, \ldots$, the following relationship is satisfied:*

$$\gamma^{k+1} = \gamma^k\left(1 + \frac{\delta_k^2}{4(1+\delta_k)}\right). \tag{8}$$

Proof. Let a^κ denote the furthest point from c^k. Then, $u^{k+1} = \left(1-\lambda^k\right)u^k + \lambda^k e^\kappa$. Therefore,

$$\begin{aligned}\gamma^{k+1} &= \Phi\left((1-\lambda^k)u^k + \lambda^k e^\kappa\right) \\ &= (1-\lambda^k)\sum_{i=1}^m u_i^k (a^i)^T(a^i) + \lambda^k (a^\kappa)^T(a^\kappa) - \left\|(1-\lambda^k)\left(\sum_{i=1}^m u_i^k a^i\right) + \lambda^k a^\kappa\right\|^2 \\ &= (1-\lambda^k)\left(\sum_{i=1}^m u_i^k (a^i)^T(a^i) - \left\|\sum_{i=1}^m u_i^k a^i\right\|^2\right) + \lambda^k(1-\lambda^k)\left\|\sum_{i=1}^m u_i^k a^i - a^\kappa\right\|^2 \\ &= (1-\lambda^k)\gamma^k + \lambda^k(1-\lambda^k)\left\|a^\kappa - c^k\right\|^2 \\ &= (1-\lambda^k)\gamma^k + \lambda^k(1-\lambda^k)(1+\delta_k)\gamma^k \\ &= \gamma^k\left(1 + \frac{\delta_k^2}{4(1+\delta_k)}\right),\end{aligned}$$

where we used (6) in the third equality, the definitions of c^k and δ_k in the fourth and fifth equalities, respectively, and the definition of λ^k in the last equality. □

We now focus on establishing an upper bound on the number of iterations required to have an iterate u^k with δ_k sufficiently small. To that end, let us define

$$\tau_\nu := \min\left\{k : \delta_k \leq \frac{1}{2^\nu}\right\}, \quad \nu = 0, 1, \ldots . \tag{9}$$

LEMMA 3.3. *τ_ν satisfies the following relationships:*

$$\tau_0 \leq 9, \tag{10a}$$

$$\tau_\nu - \tau_{\nu-1} \leq 12.5(2^\nu) \quad \nu = 1, 2, \ldots . \tag{10b}$$

Proof. Let us first consider τ_0. At each iteration $k < \tau_0$, we have $\delta_k > 1$. By Lemma 3.2,

$$\begin{aligned}\gamma^{k+1} &= \gamma^k\left(1 + \frac{\delta_k^2}{4(1+\delta_k)}\right),\\ &\geq \gamma^k(1 + 1/8),\end{aligned}$$

where we used the fact that $1 + (1/4)(x^2/(1+x))$ is an increasing function of x. Iterating this inequality, we obtain $\gamma^{k+1} \geq (9/8)^{k+1}\gamma^0$. By Lemma 3.1 and the feasibility of u^{k+1}, we have

$$\gamma_{\mathcal{A}} \geq \gamma^{k+1} \geq (9/8)^{k+1}\gamma^0 \geq (9/8)^{k+1}\left(\gamma_{\mathcal{A}}/3\right),$$

which implies that $\tau_0 \leq k + 1 \leq \log(3)/\log(9/8)$ or, equivalently, that $\tau_0 \leq 9$.

Let us now consider $\tau_\nu - \tau_{\nu-1}$ for $\nu = 1, 2, \ldots$. Let $\mu := \tau_{\nu-1}$. At each iteration k with $\delta_k > 1/2^\nu$, we similarly have

$$\gamma^{k+1} = \gamma^k\left(1 + \frac{\delta_k^2}{4(1+\delta_k)}\right) \geq \gamma^k\left(1 + \frac{1}{2^{2+\nu}(2^\nu+1)}\right).$$

At iteration μ, we have $\delta_\mu \leq 1/2^{\nu-1}$. Since the ball centered at c^μ with radius $[(1+\delta_\mu)\gamma^\mu]^{1/2}$ encloses $\mathcal{A}$, it follows that $\gamma^\mu \leq \gamma_{\mathcal{A}} \leq (1+\delta_\mu)\gamma^\mu \leq (1 + (1/2^{\nu-1}))\gamma^\mu$. Together with the repeated application of the inequality above, we have

$$\gamma_{\mathcal{A}} \geq \gamma^{\mu+k} \geq \gamma^\mu\left(1 + \frac{1}{2^{2+\nu}(2^\nu+1)}\right)^k \geq \frac{\gamma_{\mathcal{A}}}{1 + (1/2^{\nu-1})}\left(1 + \frac{1}{2^{2+\nu}(2^\nu+1)}\right)^k,$$

which implies that

$$\begin{aligned}\tau_\nu - \tau_{\nu-1} &\leq \frac{\log\left(1 + \frac{1}{2^{\nu-1}}\right)}{\log\left(1 + \frac{1}{2^{2+\nu}(2^\nu+1)}\right)}\\ &\leq \frac{1}{2^{\nu-1}}\frac{\frac{1}{2^{2+\nu}(2^\nu+1)} + 1}{\frac{1}{2^{2+\nu}(2^\nu+1)}} = \frac{2}{2^\nu} + 8(2^\nu + 1)\\ &\leq 9 + 8(2^\nu) \leq (12.5)2^\nu,\end{aligned}$$

where we used the inequalities $\log(1+x) \leq x$ for $x > -1$ and $\log(1+x) \geq x/(x+1)$ for $x > -1$. ∎

The following lemma establishes an upper bound on the number of iterations to obtain an iterate with $\delta_k \leq \delta$.

LEMMA 3.4. *Let $\delta \in (0,1)$. Algorithm* 3.1 *computes an iterate k satisfying $\delta_k \leq \delta$ in at most $9 + 50/\delta$ iterations.*

Proof. Let σ be an integer such that $1/2^\sigma \leq \delta \leq 2/2^\sigma$. Therefore, after at most τ_σ iterations, Algorithm 3.1 computes an iterate k satisfying $\delta_k \leq \delta$. By Lemma 3.3,

$$\tau_\sigma = \tau_0 + \sum_{\nu=1}^{\sigma} (\tau_\nu - \tau_{\nu-1}) \leq 9 + 12.5 \sum_{\nu=1}^{\sigma} 2^\nu \leq 9 + 25(2^\sigma) \leq 9 + 50/\delta. \qquad \square$$

We now have all of the ingredients to establish the iteration complexity of Algorithm 3.1.

THEOREM 3.1. *Given $\mathcal{A} := \{a^1, \ldots, a^m\} \subset \mathbb{R}^n$ and $\epsilon \in (0,1)$, Algorithm* 3.1 *computes a $(1+\epsilon)$-approximation to MEB($\mathcal{A}$) in at most $9 + 25/\epsilon$ iterations.*

Proof. Let u^η denote the final iterate computed by Algorithm 3.1, and let $\gamma^\eta = \Phi(u^\eta)$. Then, the trial ball centered at c^η with radius $[(1+\delta_\eta)\gamma^\eta]^{1/2}$ encloses $\mathcal{A}$. Note that u^η is a feasible solution of $(\mathcal{D})$, and $\delta_\eta \leq (1+\epsilon)^2 - 1$ by the termination criterion. Therefore, $(\gamma^\eta)^{1/2} \leq \rho_{\mathcal{A}} \leq [(1+\delta_\eta)\gamma^\eta]^{1/2} \leq (1+\epsilon)(\gamma^\eta)^{1/2}$, which implies that the ball centered at c^η with radius $[(1+\delta_\eta)\gamma^\eta]^{1/2}$ is a $(1+\epsilon)$-approximation to MEB($\mathcal{A}$).

By Lemma 3.4, Algorithm 3.1 computes such an iterate with $\delta \leq (1+\epsilon)^2 - 1$ in at most $9 + 50/(2\epsilon + \epsilon^2) \leq 9 + 25/\epsilon$ iterations. $\square$

Theorem 3.1 establishes that Algorithm 3.1 converges in $O(1/\epsilon)$ iterations. The next result presents the overall complexity.

THEOREM 3.2. *Given $\mathcal{A} := \{a^1, \ldots, a^m\} \subset \mathbb{R}^n$ and $\epsilon \in (0,1)$, Algorithm* 3.1 *computes a $(1+\epsilon)$-approximation to MEB($\mathcal{A}$) in at most $O(mn/\epsilon)$ arithmetic operations.*

Proof. The computation of the initial feasible solution u^0 requires two furthest point computations, which can be performed in $O(mn)$ operations. At each iteration, the dominating work is the computation of the furthest point from the center of the current trial ball, which also requires $O(mn)$ operations (note that γ^k can be updated using (8) in $O(1)$ operations). The result follows from Theorem 3.1. $\square$

We remark that the overall complexity of Algorithm 3.1 is linear in the number of points m and also linear in the dimension n. As such, the worst-case running time asymptotically matches the currently best known bound due to [33]. In particular, Theorem 3.2 suggests that Algorithm 3.1 is especially well-suited for large instances of the minimum enclosing ball problem where a moderately small value of ϵ (such as 10^{-3}) would be satisfactory.

We close this section by establishing that Algorithm 3.1 explicitly computes a core set of size $O(1/\epsilon)$, which also asymptotically matches the currently best known bound.

THEOREM 3.3. *Given $\mathcal{A} := \{a^1, \ldots, a^m\} \subset \mathbb{R}^n$ and $\epsilon \in (0,1)$, let η denote the index of the final iterate computed by Algorithm* 3.1. *Then, $\mathcal{X}_\eta \subseteq \mathcal{A}$ is an ϵ-core set of $\mathcal{A}$. Furthermore, $|\mathcal{X}_\eta| = O(1/\epsilon)$.*

Proof. Let u^η denote the final iterate returned by Algorithm 3.1, and let $\gamma^\eta = \Phi(u^\eta)$. Clearly, the restriction of u^η to its positive entries is a feasible solution of the dual formulation of the minimum enclosing ball problem for $\mathcal{X}_\eta$. Therefore, $\gamma^\eta \leq (\rho_{\mathcal{X}_\eta})^2 \leq (\rho_{\mathcal{A}})^2$, where $\rho_{\mathcal{X}_\eta}$ is the radius of MEB($\mathcal{X}_\eta$). However, $\gamma_{\mathcal{A}} = (\rho_{\mathcal{A}})^2 \leq (1+\delta_\eta)\gamma^\eta \leq (1+\epsilon)^2\gamma^\eta$ by Theorem 3.1. Combining these inequalities, we obtain $\rho_{\mathcal{X}_\eta} \leq \rho_{\mathcal{A}} \leq (1+\epsilon)(\gamma^\eta)^{1/2} \leq (1+\epsilon)\rho_{\mathcal{X}_\eta}$ as desired.

Note that $|\mathcal{X}_\eta|$ is precisely equal to the number of positive components of u^η. However, the initial solution u^0 has only two positive components. Each iteration can add at most one positive component to u^k. Therefore, $|\mathcal{X}_\eta| \leq 11 + 25/\epsilon = O(1/\epsilon)$ by Theorem 3.1. $\square$

4. The second algorithm. In this section, we describe our second algorithm, which is a modification of Algorithm 3.1.

Algorithm 4.1 The second algorithm that computes a $(1+\epsilon)$-approximation to MEB($\mathcal{A}$).

Require: Input set of points $\mathcal{A} = \{a^1, \ldots, a^m\} \subset \mathbb{R}^n, \epsilon > 0$.

$\alpha \leftarrow \arg\max_{i=1,\ldots,m} \left\|a^i - a^1\right\|^2, \quad \beta \leftarrow \arg\max_{i=1,\ldots,m} \left\|a^i - a^\alpha\right\|^2;$
$u_i^0 \leftarrow 0, \quad i = 1, \ldots, m;$
$u_\alpha^0 \leftarrow 1/2, \quad u_\beta^0 \leftarrow 1/2;$
$\mathcal{X}_0 \leftarrow \{a^\alpha, a^\beta\};$
$c^0 \leftarrow \sum_{i=1}^m u_i^0 a^i;$
$\gamma^0 \leftarrow \Phi(u^0);$
$\kappa \leftarrow \arg\max_{i=1,\ldots,m} \left\|a^i - c^0\right\|^2, \quad \xi \leftarrow \arg\min_{i:a^i \in \mathcal{X}_0} \left\|a^i - c^0\right\|^2;$
$\delta_0^+ \leftarrow \left(\left\|a^\kappa - c^0\right\|^2 / \gamma^0\right) - 1, \quad \delta_0^- \leftarrow 1 - \left(\left\|a^\xi - c^0\right\|^2 / \gamma^0\right);$
$\delta_0 \leftarrow \max\{\delta_0^+, \delta_0^-\};$
$k \leftarrow 0;$
While $\delta_k > (1+\epsilon)^2 - 1$, do
loop
  if $\delta_k > \delta_k^-$, **then**
    $\lambda^k \leftarrow \delta_k / [2(1+\delta_k)];$
    $k \leftarrow k+1;$
    $u^k \leftarrow (1-\lambda^{k-1})u^{k-1} + \lambda^{k-1} e^\kappa;$
    $c^k \leftarrow (1-\lambda^{k-1})c^{k-1} + \lambda^{k-1} a^\kappa;$
    $\mathcal{X}_k \leftarrow \mathcal{X}_{k-1} \cup \{a^\kappa\};$
  else
    $\lambda^k \leftarrow \min\left\{\frac{\delta_k^-}{2(1-\delta_k^-)}, \frac{u_\xi^k}{1-u_\xi^k}\right\};$
    if $\lambda^k = u_\xi^k/(1-u_\xi^k)$, **then**
      $\mathcal{X}_{k+1} \leftarrow \mathcal{X}_k \backslash \{a^\xi\};$
    else
      $\mathcal{X}_{k+1} \leftarrow \mathcal{X}_k;$
    end if
    $k \leftarrow k+1;$
    $u^k \leftarrow (1+\lambda^{k-1})u^{k-1} - \lambda^{k-1} e^\xi;$
    $c^k \leftarrow (1+\lambda^{k-1})c^{k-1} - \lambda^{k-1} a^\xi;$
  end if
  $\gamma^k \leftarrow \Phi(u^k);$
  $\kappa \leftarrow \arg\max_{i=1,\ldots,m} \left\|a^i - c^k\right\|^2, \quad \xi \leftarrow \arg\min_{i:a^i \in \mathcal{X}_k} \left\|a^i - c^k\right\|^2;$
  $\delta_k^+ \leftarrow \left(\left\|a^\kappa - c^k\right\|^2 / \gamma^k\right) - 1, \quad \delta_k^- \leftarrow 1 - \left(\left\|a^\xi - c^k\right\|^2 / \gamma^k\right);$
  $\delta_k \leftarrow \max\{\delta_k^+, \delta_k^-\};$
end loop
Output $c^k, \mathcal{X}_k, u^k, \sqrt{(1+\delta_k)\gamma^k}$.

Algorithm 4.1 starts off with the same initial solution u^0 as the one computed by Algorithm 3.1. At each iteration, the furthest point in $\mathcal{A}$ from the center c^k of the trial ball is computed as in Algorithm 3.1. In contrast, each iteration of Algorithm 4.1 also includes the computation of the *closest* point to c^k among all points in $\mathcal{X}_k \subseteq \mathcal{A}$.

Geometrically, the parameter δ_k^- is the largest number such that the current ball shrunken by a factor of $(1-\delta)^{1/2}$ does not contain any points in $\mathcal{X}_k$ for any $\delta > \delta_k^-$. Algebraically, the step performed by Algorithm 4.1 in this case corresponds to moving away from the vertex of the unit simplex that minimizes the linear approximation to $\Phi(u)$ at u^k, where the minimization is over the vertices $\{e^j : u_j^k > 0\}$. The feasible solution u^k is updated in different ways based on these two computations. If $\delta_k = \delta_k^+$, then Algorithm 4.1 uses the exact same update as in Algorithm 3.1. Otherwise, the new center c^{k+1} is obtained by moving the current center c^k *away* from the closest point $a^\xi \in \mathcal{X}_k$. Therefore, Algorithm 4.1 is obtained by incorporating "away" steps into Algorithm 3.1. For "away" steps, it is easy to verify that

$$\lambda^k = \arg \max_{\lambda \in \left[0, u_\xi^k/(1-u_\xi^k)\right]} \Phi\left((1+\lambda)u^k - \lambda e^\xi\right). \tag{11}$$

Note that the range of λ is chosen to ensure that the dual feasibility constraint $u^{k+1} \geq 0$ is satisfied.

4.1. Analysis of the second algorithm. The analysis of Algorithm 4.1 is very similar to that of Algorithm 3.1. As in [39], we call iteration k a *plus*-iteration if $\delta_k = \delta_k^+$. If $\delta_k = \delta_k^-$ and $\lambda^k = (\delta_k^-)/[2(1-\delta_k^-)]$, then we call it a *minus*-iteration. The working core set remains unchanged at a minus-iteration. Finally, if $\delta_k = \delta_k^-$ and $\lambda^k = u_\xi^k/(1-u_\xi^k)$, we then call it a *drop*-iteration, since the ξth component of u^k drops to 0 and a^ξ is removed from the working core set.

Our analysis mimics the analysis of [39] for a similar algorithm that computes an approximation to the minimum-volume enclosing ellipsoid of a finite set of points. The next lemma establishes a lower bound on the improvement at each plus- or minus-iteration.

LEMMA 4.1. *At each plus- or minus-iteration,*

$$\gamma^{k+1} \geq \gamma^k \left(1 + \frac{\delta_k^2}{4(1+\delta_k)}\right), \quad k = 0, 1, \ldots. \tag{12}$$

Proof. At a plus-iteration, the result directly follows from Lemma 3.2. At a minus-iteration, a similar application of (6) reveals that

$$\gamma^{k+1} = \Phi\left((1+\lambda^k)u^k - \lambda^k e^\xi\right) = \gamma^k \left(1 + \frac{(\delta_k^-)^2}{4(1-\delta_k^-)}\right).$$

The result easily follows from the observation that

$$\frac{(\delta_k^-)^2}{4(1-\delta_k^-)} \geq \frac{(\delta_k^-)^2}{4(1+\delta_k^-)}$$

and that $\delta_k^- = \delta_k$ at a minus-iteration. ∎

Lemma 4.1 establishes that Algorithm 4.1 makes at least as much improvement as Algorithm 3.1 at each plus- or minus-iteration. At a drop-iteration, it is easy to show that $\gamma^{k+1} \geq \gamma^k$. However, we can no longer find a positive lower bound on $\gamma^{k+1} - \gamma^k \geq 0$. Using similar reasoning as in [39], each drop-iteration can be paired with the most recent plus-iteration k at which u_ξ^k was increased from 0, except for the αth and βth components, which were positive at the initial solution and may be decreased to zero for the first time. Therefore, we can double the iteration count (and add two iterations to account for the initial positive components of u^0) in the

analysis of Algorithm 3.1 to establish that Algorithm 4.1 can compute a $(1+\epsilon)$-approximation to MEB($\mathcal{A}$) in at most twice as many iterations as that required by Algorithm 3.1. Note that this does not affect the asymptotic iteration bound of Algorithm 3.1. Furthermore, each iteration still requires $O(mn)$ operations, which implies that the asymptotic overall complexity of Algorithm 4.1 also remains the same as that of Algorithm 3.1. Finally, the asymptotic bound on the size of the core set is also unaffected. However, we remark that Algorithm 4.1 has the potential to compute even smaller core sets than those returned by Algorithm 3.1 due to the possible inclusion of minus- and drop-iterations. We summarize these results in the following theorem.

THEOREM 4.1. *Given $\mathcal{A} := \{a^1, \ldots, a^m\} \subset \mathbb{R}^n$ and $\epsilon \in (0,1)$, Algorithm* 4.1 *computes a $(1+\epsilon)$-approximation to MEB($\mathcal{A}$) in $O(mn/\epsilon)$ operations. Furthermore, upon termination, $\mathcal{X}_\eta \subseteq \mathcal{A}$ is an ϵ-core set and $|\mathcal{X}_\eta| = O(1/\epsilon)$, where η is the index of the final iterate computed by Algorithm* 4.1.

4.2. Linear convergence of the second algorithm. Despite the fact that Algorithm 4.1 appears to be a simple modification of Algorithm 3.1, it turns out that these two algorithms actually exhibit different characteristics. In particular, we establish that Algorithm 4.1 enjoys linear convergence, while a similar rate of convergence cannot, in general, be expected from Algorithm 3.1.

As observed in [41, 20], the search directions of Algorithm 3.1 always point toward the extreme points of the unit simplex. Therefore, the angle between these directions and the gradient of the objective function gets increasingly closer to the right angle in the situation when the optimal solution lies on the boundary of the unit simplex and is not an extreme point. For the minimum enclosing ball problem, an optimal solution of the dual problem will almost always lie in a lower-dimensional face of the unit simplex, except for the trivial cases such as a single input point or an input set sampled from the surface of a ball. It follows that Algorithm 3.1 is, in general, expected to exhibit a sublinear rate of convergence. In fact, this result has been formalized in [41] (see also [20, Theorem 3]) for the Frank–Wolfe algorithm under even stronger assumptions than those satisfied by the dual formulation of the minimum enclosing ball problem.

In an attempt to circumvent this drawback of Algorithm 3.1, Algorithm 4.1 works with an enlarged set of search directions by including those directions pointing away from the extreme points of the unit simplex. Such a general algorithm that incorporates "away" steps into the Frank–Wolfe algorithm was first proposed by Wolfe [41], and its convergence properties have been investigated by several authors. For the general problem of maximizing a concave function over a polytope, Wolfe [41] sketched and Guélat and Marcotte [20] detailed the proof of linear convergence under the assumptions of Lipschitz continuity of the gradient of the objective function, strong concavity of the objective function, and strict complementarity. More recently, Ahipaşaoğlu, Sun, and Todd [1] established the linear convergence of such an algorithm for the problem of maximizing a concave function over the unit simplex under a slightly different set of assumptions. Unfortunately, none of these previous results is directly applicable to our case, since either set of these assumptions implies the uniqueness of the optimal solution, which is not, in general, satisfied by the dual formulation of the minimum enclosing ball problem.

We therefore use an argument similar to that of [1] to establish the linear convergence of Algorithm 4.1. We work with a perturbation of the primal formulation $(\mathcal{P}_2)$ and show that the distance from an optimal primal-dual solution of the perturbed problem to the set of optimal primal-dual solutions of $(\mathcal{P}_2)$ and $(\mathcal{D})$ satisfies

a Lipschitz condition using the stability results of Robinson [34] for general nonlinear programming problems.

Let us define the following perturbation of $(\mathcal{P}_2)$:

$$\begin{array}{lll}(\mathcal{P}(z(u,\delta))) & \min_{c,\gamma} & \gamma \\ & \text{subject to} & \\ & & (a^i)^T a^i - 2(a^i)^T c + c^T c \le \gamma + z_i(u,\delta), \quad i = 1,\dots,m,\end{array}$$

where $u \in \mathbb{R}^m$ lies on the unit simplex, $\delta \ge 0$, and $z(u,\delta)$ is given by

$$z_i(u,\delta) := \begin{cases} \delta\Phi(u) & \text{if } u_i = 0, \\ (a^i)^T a^i - 2(a^i)^T c(u) + c(u)^T c(u) - \Phi(u) & \text{else,} \end{cases} \quad ; \quad i = 1,\dots,m,$$

where

$$c(u) := \sum_{i=1}^m u_i a^i.$$

Let $z^k := z(u^k, \delta_k),\ k = 0, 1, \dots$, where $u^k \in \mathbb{R}^m$ denotes the kth iterate and δ_k is the corresponding measure as computed by Algorithm 4.1. By a definition of δ_k,

$$(a^i)^T a^i - 2(a^i)^T c^k + (c^k)^T c^k - \Phi(u^k) \le \delta_k \Phi(u^k), \quad i = 1,\dots,m,$$

and

$$(a^i)^T a^i - 2(a^i)^T c^k + (c^k)^T c^k - \Phi(u^k) \ge -\delta_k \Phi(u^k), \quad \text{if } u_i^k > 0,$$

which implies that $|z_i^k| \le \delta_k \Phi(u^k)$ for $i = 1,\dots,m$. We remark that the latter inequality above is not necessarily satisfied by the kth iterate computed by Algorithm 3.1. Furthermore,

$$(u^k)^T z^k = \sum_{i:u_i^k > 0} u_i^k (a^i)^T a^i - 2(c^k)^T (c^k) + (c^k)^T c^k - \Phi(u^k) = 0, \tag{13}$$

where we used the definitions of c^k and $\Phi(u)$ together with the fact that u^k lies on the unit simplex. Using the fact that $c(u^k) = c^k$, it follows that $(c^k, \Phi(u^k))$ is a feasible solution of $(\mathcal{P}(z^k))$. The next lemma establishes that $(c^k, \Phi(u^k))$ is actually an optimal solution.

LEMMA 4.2. *For all $k = 0, 1, \dots$, $(c^k, \Phi(u^k))$ is an optimal solution of $(\mathcal{P}(z^k))$.*

Proof. The feasibility of $(c^k, \Phi(u^k))$ follows from the discussions preceding the lemma. Since $(\mathcal{P}(z^k))$ is a convex optimization problem and $(c^k, \Phi(u^k))$ satisfies the optimality conditions along with u^k as the Lagrange multipliers, the result follows. ∎

Let $\Xi(z(u,\delta))$ denote the optimal value of $(\mathcal{P}(z(u,\delta)))$. Note that Ξ is a convex function of $z(u,\delta)$, and if u^* is any Lagrange multiplier corresponding to the optimal solution of $\mathcal{P}(0)$ (equivalently, of $(\mathcal{P}_2)$), then u^* is a subgradient of Ξ at 0. Therefore, for all $k = 0, 1, \dots$,

$$\begin{array}{rcl}\Phi(u^k) = \Xi(z^k) & \ge & \Xi(0) + (u^*)^T z^k \\ & = & \Phi(u^*) + (u^* - u^k)^T z^k \\ & \ge & \Phi(u^*) - \|u^k - u^*\| \|z^k\|,\end{array} \tag{14}$$

where we used Lemma 4.2 and (13).

Let Δ denote the diameter of the input set $\mathcal{A}$, i.e., the maximum distance between any pair of points in $\mathcal{A}$. Since $(1/4)\Delta^2 \leq \Phi(u^*) \leq \Delta^2$, we have, for all k,

$$|z_i^k| \leq \delta_k \Phi(u^k) \leq \delta_k \Phi(u^*) \leq \delta_k \Delta^2,$$

which implies that $\|z^k\| \leq \sqrt{m}\Delta^2\delta_k$.

We will next use the stability results of Robinson [34] to establish an upper bound on $\|u^k - u^*\|$. We need to verify that all of the assumptions are satisfied for the unperturbed problem $(\mathcal{P}(0))$. Since the problem is convex and Slater's constraint qualification is satisfied, the constraints are regular at any feasible solution. Furthermore, let (c^*, γ^*) be the unique optimal solution of $(\mathcal{P}(0))$, and let u^* be any corresponding Lagrange multiplier (i.e., any optimal solution of $(\mathcal{D})$). Then, the Lagrangian function $\mathcal{L} : \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}$ for the problem $(\mathcal{P}(0))$ is given by

$$\mathcal{L}((c,\gamma),u) = \gamma + \sum_{i=1}^{m} u_i \left((a^i)^T a^i - 2(a^i)^T c + c^T c - \gamma \right).$$

By taking derivatives with respect to the primal variables $(c, \gamma) \in \mathbb{R}^n \times \mathbb{R}$, we obtain

$$\nabla_{(c,\gamma)} \mathcal{L}((c,\gamma),u) = \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \sum_{i=1}^{m} u_i \begin{bmatrix} -2a^i + 2c \\ -1 \end{bmatrix},$$

$$\nabla^2_{(c,\gamma)} \mathcal{L}((c,\gamma),u) = \sum_{i=1}^{m} u_i \begin{bmatrix} 2I & 0 \\ 0 & 0 \end{bmatrix},$$

where $I \in \mathbb{R}^{n \times n}$ denotes the identity matrix. Note that any direction $d \in \mathbb{R}^{n+1}$ orthogonal to the gradient of the objective function of $(\mathcal{P}(0))$ is of the form $d = [(d')^T, 0]^T$, where $d' \in \mathbb{R}^n$. Therefore, for any such direction d,

$$d^T \nabla^2_{(c,\gamma)} \mathcal{L}(c^*, \gamma^*, u^*) d = 2(d')^T d' = 2\|d\|^2,$$

since u^* lies on the unit simplex, which implies that Robinson's second-order sufficient condition is satisfied (see Definition 2.1 in [34]) by the optimal solution (c^*, γ^*) of $(\mathcal{P}(0))$ along with any dual optimal solution u^*. Therefore, by Theorem 4.2 in [34], there exists a dual optimal solution u^* and a positive constant ℓ such that

$$\|u^k - u^*\| \leq \ell \|z^k\| \leq \ell\sqrt{m}\Delta^2\delta_k \tag{15}$$

for all sufficiently small δ_k. Combining this inequality with (14), we obtain

$$\Phi(u^*) - \Phi(u^k) \leq m\ell\Delta^4(\delta_k)^2 \tag{16}$$

for all sufficiently small δ_k.

Suppose now that $\delta_k \leq 1/2$. Since $\Phi(u^k) \leq \Phi(u^*) \leq (1+\delta_k)\Phi(u^k) \leq (3/2)\Phi(u^k)$, it follows that

$$\Phi(u^k) \geq (2/3)\Phi(u^*) \geq (1/6)\Delta^2.$$

At each plus- or minus-iteration, by Lemma 4.1, we obtain

$$\Phi(u^{k+1}) \geq \Phi(u^k)\left(1 + \frac{\delta_k^2}{4(1+\delta_k)}\right) \geq \Phi(u^k) + \frac{\delta_k^2 \Delta^2}{36}. \tag{17}$$

Combining (16) and (17), at each plus- or minus-iteration, we obtain

$$\Phi(u^*) - \Phi(u^{k+1}) \le \Phi(u^*) - \Phi(u^k) - \frac{\delta_k^2 \Delta^2}{36} \le \left(1 - \frac{1}{36m\ell\Delta^2}\right)(\Phi(u^*) - \Phi(u^k)) \tag{18}$$

for all sufficiently small δ_k. This establishes the linear convergence of Algorithm 4.1.

THEOREM 4.2. *Given $\mathcal{A} := \{a^1, \ldots, a^m\} \subset \mathbb{R}^n$, Algorithm 4.1 computes iterates u^k such that $\Phi(u^*) - \Phi(u^k)$ is nonincreasing. Asymptotically, this gap is decreased by at least a factor of $(1 - 1/(36m\ell\Delta^2))$ at each plus- or minus-iteration. There exist constants $\bar{\tau}$ and ϑ that depend on the input data such that Algorithm 4.1 computes a $(1+\epsilon)$-approximation to MEB($\mathcal{A}$) in $\bar{\tau} + \vartheta \log(1/\epsilon)$ operations for $\epsilon \in (0, 1)$.*

Proof. Lemma 4.1 and the following discussions imply that $\Phi(u^*) - \Phi(u^k)$ is a nonincreasing sequence. The asymptotic linear convergence follows from (18). Therefore, we need only to establish the last statement.

Let $\tau := \max\{\tau_1, \tau^*\}$, where τ_1 is defined as in (9) and τ^* is the smallest value of k such that the inequality (15) is satisfied. After iteration τ, the sequence $\Phi(u^*) - \Phi(u^k)$ satisfies the relationship (18). By the termination criterion of Algorithm 4.1, it suffices to compute an iterate k_* such that $\Phi(u^{k_*}) \le \Phi(u^*) \le (1+\delta_{k_*})\Phi(u^{k_*}) \le (1+\epsilon)^2\Phi(u^{k_*})$. This implies that the final iterate satisfies $\Phi(u^*) - \Phi(u^{k_*}) \le [(1+\epsilon)^2 - 1]\Phi(u^{k_*})$. Since $\Phi(u^k) \ge (1/6)\Delta^2$ for all $k \ge \tau$, it follows that the termination criterion is satisfied if $\Phi(u^*) - \Phi(u^{k_*}) \le (1/6)[(1+\epsilon)^2 - 1]\Delta^2$. By (18), $\Phi(u^*) - \Phi(u^{k+1}) \le (1 - (1/\bar{\mu}))(\Phi(u^*) - \Phi(u^k)) \le (1 - (1/\bar{\mu}))(\Delta^2 - (1/6)\Delta^2) = (5/6)(1 - (1/\bar{\mu}))\Delta^2$ at each plus- or minus-iteration for all $k \ge \tau$, where $\bar{\mu} := 36m\ell\Delta^2$. Therefore, once Algorithm 4.1 computes iterate τ, we have

$$\Phi(u^*) - \Phi(u^{\tau+\hat{k}}) \le \frac{5}{6}\left(1 - \frac{1}{\bar{\mu}}\right)^{\hat{k}} \Delta^2$$

after $\hat{k}$ plus- or minus-iterations. Therefore, if

$$\frac{5}{6}\left(1 - \frac{1}{\bar{\mu}}\right)^{\hat{k}} \Delta^2 \le \frac{1}{6}(2\epsilon + \epsilon^2)\Delta^2,$$

then the termination criterion is satisfied after $\hat{k}$ plus- or minus-iterations. It follows that $\hat{k}$ satisfies

$$\log 5 + \hat{k} \log\left(1 - \frac{1}{\bar{\mu}}\right) \le \log\epsilon + \log(\epsilon + 2).$$

Using the inequality $\log(1 + x) \le x$ for all $x > -1$, a sufficient condition in order for the above inequality to be satisfied is given by

$$\log 5 - \frac{\hat{k}}{\bar{\mu}} \le \log 2 + \log\epsilon,$$

which implies that $\tau' + \bar{\mu}\log(1/\epsilon)$ plus- or minus-iterations will suffice, where $\tau' = \bar{\mu}\log(5/2)$. By the argument following Lemma 4.1, we can double the iteration count and add two iterations to account for the drop-iterations, which completes the proof. ∎

We remark that Theorem 4.2 establishes a polynomial convergence result for Algorithm 4.1 even if ϵ is part of the input data. In addition, it implies that the

convergence is "fast" once inequality (15) is satisfied. However, the bound on the number of iterations depends on the data as it is not known a priori when the linear convergence will kick in. As such, it does not provide a better global complexity bound than that of Theorem 4.1. Nevertheless, the results of this section will shed some light into the usually better practical performance of Algorithm 4.1 in section 6.

5. Extensions. In this section, we establish that the algorithmic frameworks of sections 3 and 4 can be used to compute an approximation to the minimum enclosing ball of more general input sets. While the cost of each iteration of the corresponding algorithms may depend on the input set, the iteration complexity and the asymptotic size of the core set remain unchanged. Therefore, the existence of an ϵ-core set of size $O(1/\epsilon)$ extends to more general sets including those with uncountably many points.

We remark that the analysis of both of the algorithms heavily relies on the structure of the dual optimization formulation $(\mathcal{D})$ of the minimum enclosing ball problem of a finite set of points. In this section, we argue that the same algorithmic framework can be applied to much more general input sets with minor modifications. We employ similar arguments as in [43], where a Frank–Wolfe-type algorithm for the problem of computing the minimum-volume enclosing ellipsoid of a finite set of ellipsoids is studied. Given a possibly infinite set of points, the primal optimization formulation $(\mathcal{P}_2)$ can be extended to a semi-infinite optimization problem with a linear objective function and infinitely many convex quadratic constraints. The main idea is to approximate the given input set using only a carefully selected finite subset of points and then to refine this approximation by adding more points if necessary. This leads to an approximation of the primal formulation with only a finite number of constraints, and this approximation is refined by adding more constraints. In the dual formulation, we therefore start with a finite number of variables and add more variables if necessary.

Let $\mathcal{A} \subset \mathbb{R}^n$ be an arbitrary compact input set, and let us first consider Lemma 3.1, which establishes the quality of the initial feasible solution computed by each of the two algorithms. The initial working core set $\mathcal{X}_0$ provides the first approximation to the given input set with only two points. Let $\Phi_0(\cdot)$ denote the objective function of the dual formulation of the minimum enclosing ball problem for $\mathcal{X}_0$, and let $\gamma_{\mathcal{A}}$ denote the optimal value of the aforementioned semi-infinite primal formulation. The result of Lemma 3.1 continues to hold, since the proof relies on Lemma 2.2, which remains true for arbitrary compact input sets. The proof of Lemma 2.2 is based on the argument that an enclosing ball of smaller radius can be constructed by moving the center away from the half-space in the direction of the normal vector of the bounding hyperplane if the hypothesis of Lemma 2.2 is not satisfied by that half-space. Therefore, we still have $\Phi_0(u^0) \geq (1/3)\gamma_{\mathcal{A}}$, which implies that the quality of the initial solution is independent of the input set.

Similarly, let $\Phi_k(\cdot)$ denote the objective function of the dual formulation of the minimum enclosing ball problem for $\mathcal{X}_k \subset \mathcal{A}$. At iteration k in each algorithm, $\mathcal{X}_k$ provides the current finite approximation to $\mathcal{A}$. Let $c^k \in \mathbb{R}^n$ denote the current center. Each algorithm computes the furthest point in $\mathcal{A}$ from c^k. In Algorithm 3.1, $\mathcal{X}_{k+1}$ is obtained by adding this point to $\mathcal{X}_k$. Unless the furthest point in $\mathcal{A}$ already belongs to $\mathcal{X}_k$, the dual formulation for $\mathcal{X}_{k+1}$ differs from that for $\mathcal{X}_k$ in only one variable. Therefore, $[(u^k)^T, 0]^T$ is a feasible solution for the new dual formulation that satisfies $\Phi_{k+1}([(u^k)^T, 0]^T) = \Phi_k(u^k)$, which implies that the improvement in each iteration still obeys the relation given by Lemma 3.2, with γ^{k+1} replaced by $\Phi_{k+1}(u^{k+1})$ and γ^k by $\Phi_k(u^k)$. Note that the dimension of u^{k+1} is one more than that of u^k in this case. It follows that the upper bound on the number of iterations required by Algorithm 3.1 to

achieve a prescribed accuracy as well as the bound on the core set remain unchanged for a general compact input set $\mathcal{A} \subset \mathbb{R}^n$.

The preceding argument establishes the same improvement result at a plus-iteration of Algorithm 4.1 for a general input set $\mathcal{A}$. Since $\mathcal{X}_k$ is finite, the computation of the closest point in $\mathcal{X}_k$ is straightforward independently of the input set. At a minus-iteration, the dimension of the dual formulation remains the same. Therefore, Lemma 4.1 still applies. At a drop-iteration, we can reverse the argument employed at a plus-iteration, since the number of dual variables actually decreases in this case. We conclude that the iteration complexity of Algorithm 4.1 and the upper bound on the size of the core set also remain unchanged for a general compact input set $\mathcal{A} \subset \mathbb{R}^n$. On the other hand, our analysis that leads to the linear convergence of Algorithm 4.1 is not likely to be extended to more general input sets, since it explicitly relies on the stability results for nonlinear programming problems with a finite number of constraints.

We give another perspective on the extension of the two algorithms to more general input sets. Let $\mathcal{X}_\eta \subset \mathcal{A}$ denote the finite set computed by either one of the two algorithms upon termination on a general input set $\mathcal{A}$. Then, each algorithm would geometrically behave exactly the same way on the input set $\mathcal{X}_\eta$ as it would on the original input set $\mathcal{A}$. However, the termination criterion is satisfied for the whole set $\mathcal{A}$. Clearly, the set $\mathcal{X}_\eta \subset \mathcal{A}$ is not known a priori and is sequentially generated by each algorithm. Furthermore, the cost of each iteration is likely to be higher for a general input set $\mathcal{A}$ in comparison with that for $\mathcal{X}_\eta$. Therefore, the main work involved in each algorithm is the extraction of the finite set $\mathcal{X}_\eta$ from $\mathcal{A}$.

In order to transform this conceptual algorithmic framework into a practical algorithm, we need to ensure that each operation required by either algorithm can be carried out efficiently for a given input set. Note that both of the algorithms in this paper compute the initial feasible solution in a similar fashion. This computation entails finding the furthest point in the input set from a fixed point. In addition, similar furthest point computations are performed at each iteration of both of the algorithms. Therefore, the extent of input sets which are amenable to these algorithms highly depends on the efficiency with which such computations can be performed.

We now specify several input sets for which similar algorithmic frameworks can be applied.

5.1. Set of balls. Let $\mathcal{A} = \{\mathcal{B}_1, \ldots, \mathcal{B}_m\} \subset \mathbb{R}^n$ be a set of m balls. Given $\mathcal{B}_{c,\rho}$ and $x \in \mathbb{R}^n$, the furthest point in $\mathcal{B}_{c,\rho}$ from x is given by $x^* = c + \rho(c - x)/\|c - x\|$, which can be computed in $O(n)$ operations. Therefore, each iteration of Algorithm 3.1 still requires $O(mn)$ operations, which implies that Algorithm 3.1 computes a $(1+\epsilon)$-approximation to $\mathrm{MEB}(\mathcal{A})$ in $O(mn/\epsilon)$ operations and returns an ϵ-core set of size $O(1/\epsilon)$. In addition to computing the furthest point at each iteration, Algorithm 4.1 also requires the computation of the closest point in a finite set. The size of this set is bounded above by $O(1/\epsilon)$, which implies that each iteration can be performed in $O(mn + n/\epsilon)$ operations. Therefore, Algorithm 4.1 can compute a $(1+\epsilon)$-approximation to $\mathrm{MEB}(\mathcal{A})$ in $O(mn/\epsilon + n/\epsilon^2)$ operations and returns an ϵ-core set of size $O(1/\epsilon)$.

5.2. Set of ellipsoids. Let $\mathcal{A} = \{\mathcal{E}_1, \ldots, \mathcal{E}_m\} \subset \mathbb{R}^n$ be a set of m ellipsoids given by $\mathcal{E}_i := \{x \in \mathbb{R}^n : (x - c^i)^T Q^i (x - c^i) \leq 1\}$, where $c^i \in \mathbb{R}^n$ and $Q^i \in \mathbb{R}^{n \times n}$ is symmetric and positive definite for $i = 1, \ldots, m$. The furthest point in an ellipsoid from a given point can be computed using a tight semidefinite programming relaxation with a fixed number of constraints in $O(n^{O(1)})$ operations in the real number model of com-

putation [43], where $O(1)$ denotes a universal constant greater than three. Therefore, Algorithm 3.1 computes a $(1+\epsilon)$-approximation to MEB($\mathcal{A}$) in $O(mn^{O(1)}/\epsilon)$ operations and returns an ϵ-core set of size $O(1/\epsilon)$. Similarly, Algorithm 4.1 can compute a $(1+\epsilon)$-approximation to MEB($\mathcal{A}$) in $O(mn^{O(1)}/\epsilon + n^{O(1)}/\epsilon^2)$ operations and returns an ϵ-core set of size $O(1/\epsilon)$.

5.3. Set of half ellipsoids. $\mathcal{H}$ is said to be a half ellipsoid if it is given by the intersection of an ellipsoid with a half-space. Let $\mathcal{A} = \{\mathcal{H}_1, \ldots, \mathcal{H}_m\}$, where $\mathcal{H}_i := \{x \in \mathbb{R}^n : (x - c^i)^T Q^i (x - c^i) \leq 1, \quad (f^i)^T x \leq \omega^i\}$, where $c^i \in \mathbb{R}^n$, $f^i \in \mathbb{R}^n, \omega^i \in \mathbb{R}$, and $Q^i \in \mathbb{R}^{n \times n}$ is symmetric and positive definite for $i = 1, \ldots, m$. Sturm and Zhang [36] established that the maximization of any quadratic function over a half ellipsoid can be cast as a semidefinite programming problem with a fixed number of constraints similarly to quadratic optimization over an ellipsoid. Therefore, the asymptotic overall complexity bounds of Algorithms 3.1 and 4.1 are identical to those for the case of a set of ellipsoids. In particular, both algorithms return an ϵ-core set of size $O(1/\epsilon)$.

5.4. Set of intersections of a pair of similar ellipsoids. Two n-dimensional ellipsoids $\mathcal{E}_1$ and $\mathcal{E}_2$ are said to be similar if they both admit a representation using the same semidefinite matrix. This implies that the length and the alignment of the corresponding axes are the same. Let $\mathcal{A} = \{\mathcal{T}_1, \ldots, \mathcal{T}_m\}$, where $\mathcal{T}_i := \{x \in \mathbb{R}^n : (x - c^i)^T Q^i (x - c^i) \leq 1, \quad (x - h^i)^T Q^i (x - h^i) \leq 1\}$, where $c^i \in \mathbb{R}^n, h^i \in \mathbb{R}^n$, and $Q^i \in \mathbb{R}^{n \times n}$ is symmetric and positive definite for $i = 1, \ldots, m$. It follows from the results of [36] that any quadratic optimization problem over the intersection of a pair of similar ellipsoids can be decomposed into two quadratic optimization problems over two half ellipsoids. Therefore, the asymptotic complexity bounds of Algorithms 3.1 and 4.1 are identical to those for the case of a set of half ellipsoids. Similarly, both algorithms return an ϵ-core set of size $O(1/\epsilon)$.

5.5. Further extensions. We have described several classes of more general input sets for which an approximate minimum enclosing ball can be computed in polynomial time (for fixed ϵ) using the appropriate extensions of Algorithms 3.1 and 4.1. Obviously, the results can be extended to input sets that are composed of a combination of elements from each of the above classes. In particular, it is remarkable that the existence of an ϵ-core set of size $O(1/\epsilon)$ extends to much more general classes of input sets including those with uncountably many points.

Similarly to the discussion in [43], the extent of input sets to which similar algorithmic frameworks can be applied largely depends on the efficiency of the furthest point computation required at each iteration of each of the two algorithms. It is well known that the maximization of a convex quadratic function over certain sets (such as polytopes defined by inequalities) is computationally intractable. Therefore, our algorithmic framework does not yield a polynomial-time algorithm for an input set of polytopes. In summary, the discovery of polynomial-time routines for quadratic optimization over other classes of input sets may lead to further efficient generalizations of our algorithms.

6. Computational results. In this section, we report the results of our computational experiments. We implemented Algorithms 3.1 and 4.1 in MATLAB. For the purposes of comparison, we also implemented the first-order algorithm of Bădoiu and Clarkson [2] (henceforth the BC algorithm). Their simple algorithm starts by setting any arbitrary point $a^i \in \mathcal{A}$ as the initial center c^1. At iteration k, let a^{j_k} denote the furthest point from c^k, $k = 1, 2, \ldots$. The center is updated according to the following

TABLE 1
Computational results on instances with $m \gg n$ ($\epsilon = 10^{-3}$).

		Time			Core set size			Iterations		
n	m	A1	A2	BC	A1	A2	BC	A1	A2	BC
10	500	0.06	0.03	0.12	4.2	3.9	5.2	168.7	44.5	435.5
10	1000	0.15	0.03	0.14	4.6	3.8	5.4	330.7	41.6	344.4
20	5000	1.7	0.36	3.11	5.9	5.2	7	246.8	46	464.2
20	10000	4.46	0.58	4.65	4.9	4.1	5.8	319.2	36.3	334.4
30	30000	27	6.45	24.59	8.6	6.8	9.1	446.4	103.6	409
50	50000	71.62	16.87	68.78	10.5	9.5	11.8	429.8	98.4	415.1
100	100000	287.99	77.74	268.11	15.9	14.5	16.6	451.7	119	422.6

relation:

$$c^{k+1} = [1 - 1/(k+1)]c^k + [1/(k+1)]a^{j_k}, \quad k = 1, 2, \ldots .$$

Bădoiu and Clarkson establish that $1/\epsilon^2$ such updates suffice in order to obtain a $(1+\epsilon)$-approximation to MEB($\mathcal{A}$). Note that each iteration requires $O(mn)$ operations, which yields an overall complexity bound of $O(mn/\epsilon^2)$.

Similarly to Algorithms 3.1 and 4.1, it is easy to verify that the BC algorithm also generates a sequence of feasible solutions for the dual formulation of the minimum enclosing ball problem. Therefore, in order to have a fair and meaningful comparison, we employed the same termination criterion that we used for Algorithms 3.1 and 4.1 rather than running the BC update for $1/\epsilon^2$ times.

In contrast with Algorithms 3.1 and 4.1, the objective functions evaluated at the iterates generated by the BC algorithm are not monotonically increasing in general. Therefore, the analysis of the BC algorithm uses entirely different tools [2].

The computational experiments were carried out on a Pentium IV processor with a clock speed of 2.80 GHz and 512 MB RAM running under Linux. We used MATLAB version 7.3.0.298 (R2006b) in our experiments.

We used three data sets in our experiments. The first data set is restricted to instances with $m \gg n$ and was randomly generated as in [1], with sizes (n, m) varying from $(10, 500)$ to $(100, 100000)$. For each fixed (n, m), ten different data sets were generated, and the results are reported in terms of the averages over these data sets in Table 1, which is divided into four sets of columns. The first set of columns reports the size (n, m). The next three sets of columns present the CPU time, core set size, and the number of iterations, respectively. Each one of these three sets is further divided into three columns labeled A1, A2, and BC corresponding to Algorithm 3.1, Algorithm 4.1, and the BC algorithm, respectively. In all of our experiments, we set $\epsilon = 10^{-3}$.

As illustrated by Table 1, each of the three algorithms is capable of quickly computing an approximation to the minimum enclosing ball of the given input set. In particular, all three algorithms terminated under eight minutes even on the largest instances. In terms of CPU time, Algorithm 4.1 has significantly better performance than Algorithm 3.1 and the BC algorithm, both of which have similar running times. All three algorithms computed very small core sets of similar sizes. Algorithm 4.1 always returned the smallest core sets for each input set. The core sets computed by Algorithm 3.1 and the BC algorithm have similar sizes with the former being slightly better than the latter. In terms of the number of iterations, Algorithm 4.1 once again significantly outperforms the other two algorithms. Unlike Algorithm 3.1, the number

TABLE 2
Computational results on instances with $n \gg m$ ($\epsilon = 10^{-3}$).

		Time			Core set size			Iterations		
n	m	A1	A2	BC	A1	A2	BC	A1	A2	BC
10000	100	7.56	7.62	29.9	90.6	90.4	90.8	117.4	118.2	476.6
10000	1000	149.39	148.16	321.25	198.4	197	202	241	238.8	524.2
25000	1000	541.47	539.06	957.02	266.6	265.6	272.4	303.2	301.8	541.2

TABLE 3
Vertices of the unit simplex ($m = n = 1000$).

	Time			Core set size			Iterations		
ϵ	A1	A2	BC	A1	A2	BC	A1	A2	BC
1	.24	.25	.19	2	2	2	0	0	1
.1	.83	.83	.82	11	11	11	9	9	10
.01	6.57	6.58	7.27	101	101	101	99	99	100
.001	63.89	64.07	71.36	1000	1000	1000	998	998	999

of iterations of the BC algorithm seems to be independent of the dimensions of the input set.

A close examination of Table 1 reveals that Algorithm 4.1 resulted in reductions of 73% to 88% in terms of running time and of 74% to 90% in terms of the number of iterations in comparison with the other two algorithms. These results seem to indicate that the linear convergence of Algorithm 4.1 may be responsible for the improved performance. Furthermore, due to allowing points to be dropped from the working core set, the sizes of the core sets computed by Algorithm 4.1 are about 10% to 30% smaller than those returned by the other two algorithms.

The second data set consists of instances with $n \gg m$. In particular, we generated random instances with (n, m) varying from $(10000, 100)$ to $(25000, 1000)$. The averaged results are presented in Table 2, which is organized similarly to Table 1. The results indicate that all three algorithms compute core sets of similar sizes. Algorithm 3.1 and Algorithm 4.1 exhibit similar performances in terms of running time and the number of iterations due to the fact that "away" steps are performed relatively infrequently on such instances. On the other hand, the running time and the number of iterations of the BC algorithm are considerably larger than either of our two algorithms. Once again, note that the number of iterations of the BC algorithm seems to be relatively insensitive to m and n, which suggests a stronger relationship with $1/\epsilon$ in comparison with our algorithms.

The final data set we considered is the vertices of the unit simplex. Bădoiu and Clarkson [3] establish a tight upper bound of $\lceil 1/\epsilon \rceil$ on the size of the core set for such an input set under the assumption that $n \geq \lfloor 1/\epsilon \rfloor$. In an attempt to assess the performances of the three algorithms on such a data set, we considered the vertices of the unit simplex with $n = 1000$ using $\epsilon \in \{1, .1, .01, .001\}$. The results of this experiment are presented in Table 3, which is organized similarly to Table 1.

As illustrated by Table 3, all three algorithms have similar performances on the vertices of the unit simplex in $\mathbb{R}^n$, with $n = 1000$. Note that both the size of the core set and the number of iterations grow proportionally to $1/\epsilon$. These results are in agreement with the tight core set bound of [3]. This example illustrates that the asymptotic bounds on the core set size and the number of iterations for Algorithms 3.1 and 4.1, in general, cannot be improved. However, all three algorithms computed the exact minimum enclosing ball for $\epsilon = 10^{-3}$ (and for any $\epsilon \geq 10^{-3}$). Therefore, this

example illustrates that the upper bound of $\lceil 1/\epsilon \rceil$ on the size of the core set is no longer tight for $n \leq \lfloor 1/\epsilon \rfloor$.

We do not compare our algorithms with other exact or approximate algorithms, since such computational studies have been performed in earlier literature. For instance, it is well known that the minimum enclosing ball problem can be formulated as an instance of second-order cone programming and interior-point methods can achieve very high accuracy (e.g., 10^{-8}) in small- and medium-scale instances. However, each iteration requires the computation and factorization of an $(n+1) \times (n+1)$ matrix, which can be performed in $O(n^3)$ and $O(mn^2)$ operations, respectively [24]. Therefore, such an approach is not computationally feasible for large instances as illustrated by the results of [42, 44]. Similarly, exact algorithms [16] perform well on small- and medium-scale instances, but the performance degrades significantly for large-scale instances [24]. Since our focus is on applications with large-scale instances in which a moderate accuracy suffices, our computational results indicate that our algorithms are capable of solving such instances in a reasonable amount of time.

7. Concluding remarks. In this paper, we proposed and analyzed two algorithms that compute an approximation to the minimum enclosing ball of a given finite set of points. Both algorithms exploit the special structure of the dual formulation of the problem and can geometrically be viewed as generating a sequence of trial balls until a ball with desired properties is computed. Each of the two algorithms is especially well-suited for the large-scale instances of the minimum enclosing ball problem for which a moderate approximation suffices. Both algorithms can compute a small core set whose size depends only on the approximation parameter. The second algorithm asymptotically exhibits linear convergence, which further contributes to its efficiency. We have discussed how our algorithms can be extended to more general input sets without sacrificing the iteration complexity and hence the size of the core set. In particular, we established that several more general classes of input sets admit small and finite core sets. Our computational experiments reveal that both of our algorithms are capable of quickly computing a good approximation to the minimum enclosing ball of a finite set of points. Algorithm 4.1, which is obtained by incorporating "away" steps into Algorithm 3.1, seems to exhibit a significantly better performance than other first-order algorithms. The sizes of the core sets computed by our algorithms are usually fairly small. The example that consists of the vertices of the unit simplex illustrates that our analysis, in general, cannot be improved.

While the discovery of efficient algorithms such as interior-point methods revolutionized convex optimization, the computational cost of each iteration of such algorithms quickly becomes prohibitive as the size of the problems increases. Therefore, it seems desirable to design specialized algorithms for large-scale problems that exploit the underlying special structure of the problem. We have developed two such algorithms for the minimum enclosing ball problem in this paper. We intend to continue our work on developing specialized algorithms for other classes of large-scale structured optimization problems in the near future.

Acknowledgments. I gratefully acknowledge the insightful comments and suggestions by the Associate Editor and two anonymous referees, which contributed significantly to the improvement of the manuscript. In particular, section 4.2 was added based on the comments of an anonymous referee, and the alternative perspective on the extension of the algorithms to more general input sets in section 5 was suggested by another anonymous referee.

REFERENCES

[1] D. Ahipaşaoğlu, P. Sun, and M. J. Todd, *Linear convergence of a modified Frank-Wolfe algorithm for computing minimum-volume enclosing ellipsoids*, Optim. Methods Softw., 23 (2008), pp. 5–19.

[2] M. Bădoiu and K. L. Clarkson, *Smaller core-sets for balls*, in Proceedings of the 14th Annual Symposium on Discrete Algorithms, 2003, pp. 801–802.

[3] M. Bădoiu and K. L. Clarkson, *Optimal core-sets for balls*, Comput. Geom. Theory Appl., 40 (2008), pp. 14–22.

[4] M. Bădoiu, S. Har-Peled, and P. Indyk, *Approximate clustering via core-sets*, in Proceedings of the 34th Annual ACM Symposium on Theory of Computing, 2002, pp. 250–257.

[5] L. Blum, M. Shub, and S. Smale, *On a theory of computation and complexity over the real numbers: NP-completeness, recursive functions, and universal machines*, Bull. Amer. Math. Society (N.S.), 21 (1989), pp. 1–46.

[6] L. M. Blumenthal and G. E. Wahlin, *On the spherical surface of smallest radius enclosing a bounded subset of n-dimensional Euclidean space*, Bull. Amer. Math. Soc. (N.S.), 47 (1941), pp. 771–777.

[7] R. K. Chakraborty and P. K. Chaudhuri, *Note on geometrical solutions for some minimax location problems*, Transportation Sci., 15 (1981), pp. 164–166.

[8] J. A. Chatelon, D. W. Hearn, and T. J. Lowe, *A subgradient algorithm for certain minimax and minisum location problems*, Math. Program., 15 (1978), pp. 130–145.

[9] G. Chrystal, *On the problem to construct the minimum circle enclosing n given points in the plane*, in Proc. Edinb. Math. Soc., 3 (1885), pp. 30–33.

[10] K. L. Clarkson, *Coresets, sparse greedy approximation and the Frank-Wolfe algorithm*, in Proceedings of the 19th Annual Symposium on Discrete Algorithms, 2008, pp. 922–931.

[11] D. J. Elzinga and D. W. Hearn, *The minimum covering sphere problem*, Management Sci., 19 (1972), pp. 96–104.

[12] K. Fischer and B. Gärtner, *The smallest enclosing ball of balls: Combinatorial structure and algorithms*, Internat. J. Comput. Geom. Appl., 14 (2004), pp. 341–378.

[13] K. Fischer, B. Gärtner, and M. Kutz, *Fast smallest-enclosing-ball computation in high dimensions*, in Algorithms–ESA, Lect. Notes in Comput. Sci. 2832, G. Di Battista and U. Zwick. eds., Springer, Berlin/Heidelberg, 2003, pp. 630–641.

[14] R. L. Francis, *Some aspects of a minimax location problem*, Oper. Res., 15 (1967), pp. 1163–1169.

[15] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval Res. Logist., 3 (1956), pp. 95–110.

[16] B. Gärtner, *Fast and robust smallest enclosing balls*, in Proceedings of the 7th Annual European Symposium on Algorithms (ESA), Lect. Notes in Comput. Sci. 1643, J. Nesetril, ed., Springer, New York, 1999, pp. 325–338.

[17] B. Gärtner and S. Schönherr, *An efficient, exact, and generic quadratic programming solver for geometric optimization*, in Proceedings of the 16th Annual Symposium on Computational Geometry, 2000, pp. 110–118.

[18] A. Goel, P. Indyk, and K. R. Varadarajan, *Reductions among high-dimensional proximity problems*, in Proceedings of the 12th ACM-SIAM Symposium on Discrete Algorithms, 2001, pp. 769–778.

[19] M. Grötschel, L. Lovász, and A. Schrijver, *Geometric Algorithms and Combinatorial Optimization*, Algorithms Combin. 2, Springer, New York, 1988.

[20] J. Guélat and P. Marcotte, *Some comments on Wolfe's away steps*, Math. Program., 35 (1986), pp. 110–119.

[21] C. H. Guo, M. Y. Lu, J. T. Sun, and Y. C. Lu, *A new algorithm for computing the minimal enclosing sphere in feature space*, in Fuzzy Systems and Knowledge Discovery, Lect. Notes in Comput. Sci. 3614, L. Wang and Y. Jin, eds., Springer, Berlin/Heidelberg, 2005, pp. 196–204.

[22] D. W. Hearn and J. Vijay, *Efficient algorithms for the (weighted) minimum circle problem*, Oper. Res., 30 (1981), pp. 777–795.

[23] S. K. Jacobsen, *An algorithm for the minimax Weber problem*, European J. Oper. Res., 6 (1981), pp. 144–148.

[24] P. Kumar, J. S. B. Mitchell, and E. A. Yildirim, *Approximate minimum enclosing balls in high dimensions using core-sets*, ACM J. Exp. Algorithmics, 8 (2003), article no. 1.1.

[25] C. L. Lawson, *The smallest covering cone or sphere*, SIAM Rev., 7 (1965), pp. 415–416.

[26] J. Matoušek and B. Gärtner, *Understanding and Using Linear Programming*, Universitext, Springer, New York, 2006.

[27] N. Megiddo, *Linear time algorithms for linear programming in* $\mathbb{R}^3$ *and related problems*, SIAM J. Comput., 12 (1983), pp. 759–776.
[28] K. P. K. Nair and R. Chandrasekaran, *Optimal location of a single service center of certain types*, Naval Res. Logist., 18 (1971), pp. 503–510.
[29] F. Nielsen and R. Nock, *Approximating smallest enclosing balls*, in Computational Science and Its Applications, Lect. Notes in Comput. Sci. 3045, A. Laganá, M. Gavrilov, V. Kumar, Y. Mun, C. Tan, and O. Gervasi, eds., Springer, Berlin/Heidelberg, 2004, pp. 147–157.
[30] F. Nielsen and R. Nock, *A fast deterministic smallest enclosing disk approximation algorithm*, Inform. Process. Lett., 93 (2005), pp. 263–268.
[31] B. J. Oommen, *An efficient geometric solution to the minimum spanning circle problem*, Oper. Res., 35 (1987), pp. 80–86.
[32] S. H. Pan and X. S. Li, *An efficient algorithm for the smallest enclosing ball problem in high dimensions*, Appl. Math. Comput., 172 (2006), pp. 49–61.
[33] R. Panigrahy, *Minimum Enclosing Polytope in High Dimensions*, manuscript, 2006.
[34] S. M. Robinson, *Generalized equations and their solutions, part* ii: *Applications to nonlinear programming*, Math. Program. Study, 19 (1982), pp. 200–221.
[35] M. I. Shamos, *Computational Geometry*, Ph.D. thesis, Yale University, New Haven, CT, 1978.
[36] J. F. Sturm and S. Z. Zhang, *On cones of nonnegative quadratic functions*, Math. Oper. Res., 28 (2003), pp. 246–267.
[37] J. J. Sylvester, *A question in the geometry of situation*, Q. J. Pure Appl. Math., 1 (1857).
[38] J. J. Sylvester, *On Poncelet's approximate linear valuation of Surd forms*, Philos. Mag., 20 (1860), pp. 203–222. Fourth Series.
[39] M. J. Todd and E. A. Yildirim, *On Khachiyan's algorithm for the computation of minimum volume enclosing ellipsoids*, Discrete Appl. Math., 155 (2007), pp. 1731–1744.
[40] E. Welzl, *Smallest enclosing disks (balls and ellipsoids)*, in New Results and New Trends in Computer Science, Lect. Notes in Comput. Sci. 555, H. Maurer, ed., Springer-Verlag, 1991, pp. 359–370.
[41] P. Wolfe, *Convergence theory in nonlinear programming*, in Integer and Nonlinear Programming, J. Abadie, ed., North-Holland, Amsterdam, 1970, pp. 1–36.
[42] S. Xu, R. M. Freund, and J. Sun, *Solution methodologies for the smallest enclosing circle problem*, Comput. Optim. Appl., 25 (2003), pp. 283–292.
[43] E. A. Yildirim, *On the minimum volume covering ellipsoid of ellipsoids*, SIAM J. Optim., 17 (2006), pp. 621–641.
[44] G. Zhou, K. C. Toh, and B. Sun, *Efficient algorithms for the smallest enclosing ball problem*, Comput. Optim. Appl., 30 (2005), pp. 147–160.

SIAM J. OPTIM.
Vol. 19, No. 3, pp. 1392–1396

IDENTIFICATION AND ELIMINATION OF INTERIOR POINTS FOR THE MINIMUM ENCLOSING BALL PROBLEM*

S. DAMLA AHIPAŞAOĞLU† AND E. ALPER YILDIRIM‡

Abstract. Given $\mathcal{A} := \{a^1, \ldots, a^m\} \subset \mathbb{R}^n$, we consider the problem of reducing the input set for the computation of the minimum enclosing ball of $\mathcal{A}$. In this note, given an approximate solution to the minimum enclosing ball problem, we propose a simple procedure to identify and eliminate points in $\mathcal{A}$ that are guaranteed to lie in the interior of the minimum-radius ball enclosing $\mathcal{A}$. Our computational results reveal that incorporating this procedure into two recent algorithms proposed by Yıldırım lead to significant speed-ups in running times especially for randomly generated large-scale problems. We also illustrate that the extra overhead due to the elimination procedure remains at an acceptable level for spherical or almost spherical input sets.

Key words. minimum enclosing balls, input set reduction, approximation algorithms

AMS subject classifications. 90C25, 90C46, 65K05

DOI. 10.1137/080727208

1. Introduction.

Given $\mathcal{A} := \{a^1, \ldots, a^m\} \subset \mathbb{R}^n$, we denote the unique minimum enclosing ball of $\mathcal{A}$ by MEB($\mathcal{A}$), i.e.,

$$\text{MEB}(\mathcal{A}) = \mathcal{B}_{c^*,\rho^*} := \{x \in \mathbb{R}^n : \|x - c^*\| \leq \rho^*\},$$

where $c^* \in \mathbb{R}^n$ is the optimal center, $\rho^* \in \mathbb{R}$ is the optimal radius, and $\|\cdot\|$ denotes the Euclidean norm. Given $\epsilon > 0$, a ball $\mathcal{B}_{c,\rho}$ is said to be a $(1+\epsilon)$-approximate solution to MEB($\mathcal{A}$) if

$$\rho \leq \rho^*, \quad \mathcal{A} \subset \mathcal{B}_{c,(1+\epsilon)\rho}. \tag{1}$$

In this note, given a $(1+\epsilon)$-approximate solution $\mathcal{B}_{c,\rho}$ to MEB($\mathcal{A}$), we propose a simple condition that should be satisfied by each point in $\mathcal{A}$ that lies on the boundary of MEB($\mathcal{A}$). Furthermore, we derive an upper bound on the Euclidean distance between c and c^*.

2. Main result.

LEMMA 2.1. *Given $\mathcal{A} := \{a^1, \ldots, a^m\} \subset \mathbb{R}^n$ and $\epsilon > 0$, let $\mathcal{B}_{c,\rho}$ be a $(1+\epsilon)$-approximate solution to MEB($\mathcal{A}$). Then,*

$$\|c - c^*\| \leq (2\epsilon + \epsilon^2)^{1/2}\rho. \tag{2}$$

Furthermore, each point $a^i \in \mathcal{A}$ on the boundary of MEB($\mathcal{A}$) satisfies

$$\|a^i - c\| \geq (1 - (2\epsilon + \epsilon^2)^{1/2})\rho. \tag{3}$$

*Received by the editors June 13, 2008; accepted for publication (in revised form) July 21, 2008; published electronically November 21, 2008.

http://www.siam.org/journals/siopt/19-3/72720.html

†School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY 14853 (dse8@cornell.edu). This author was supported in part by NSF grant DMS-0513337 and ONR grant N00014-08-1-0036.

‡Department of Industrial Engineering, Bilkent University, 06800 Bilkent, Ankara, Turkey (yildirim@bilkent.edu.tr). This author was supported in part by TÜBİTAK (Turkish Scientific and Technological Research Council) grant 107M411.

Proof. Suppose that $c \neq c^*$. Consider the hyperplane $\mathcal{H}$ passing through c^* perpendicular to $c^* - c$. Let $\mathcal{H}_+$ denote the closed halfspace bounded by $\mathcal{H}$ and not containing c. Then, by [2, Lemma 2.2], there exists a point $a^j \in \mathcal{H}_+ \cap \mathcal{A}$ such that $\|a^j - c^*\| = \rho^*$. Therefore, $\|c - a^j\|^2 \geq \|c - c^*\|^2 + \|c^* - a^j\|^2$, which implies that

$$\begin{aligned}\|c - c^*\|^2 &\leq \|c - a^j\|^2 - \|c^* - a^j\|^2,\\ &\leq (1+\epsilon)^2\rho^2 - (\rho^*)^2,\\ &\leq (1+\epsilon)^2\rho^2 - \rho^2,\\ &= (2\epsilon + \epsilon^2)\rho^2,\end{aligned}$$

where we used (1) to derive the second and third inequalities. This establishes (2).

Let a^i be any point on the boundary of MEB($\mathcal{A}$). Then, $\|a^i - c^*\| \leq \|a^i - c\| + \|c - c^*\|$, which implies that

$$\begin{aligned}\|a^i - c\| &\geq \rho^* - \|c - c^*\|,\\ &\geq \rho - (2\epsilon + \epsilon^2)^{1/2}\rho,\\ &= (1 - (2\epsilon + \epsilon^2)^{1/2})\rho,\end{aligned}$$

where we used (1) and (2) to derive the second inequality. This completes the proof. ∎

3. Computational results. Recently, Yıldırım [2] proposed two first-order algorithms that can compute a $(1+\epsilon)$-approximate solution to the minimum enclosing ball of a finite input set $\mathcal{A}$ of points for any given $\epsilon > 0$. Each algorithm generates a sequence of approximate minimum enclosing balls $\mathcal{B}_{c^k,\rho^k}$, which converge to MEB($\mathcal{A}$) in the limit. Each such ball is a $(1+\epsilon^k)$-approximate solution to MEB($\mathcal{A}$) for a certain $\epsilon^k > 0$, and the algorithm terminates when $\epsilon^k \leq \epsilon$. Both of these algorithms extract a small core set $\mathcal{X} \subseteq \mathcal{A}$ and can be extended to much more general input sets without sacrificing the small core set result.

Lemma 2.1 can be easily incorporated into both of the algorithms in [2] in an attempt to eliminate interior points in $\mathcal{A}$ (with respect to MEB($\mathcal{A}$)) thereby reducing the size of the input set. This elimination procedure does not affect the minimum enclosing ball and may decrease the computational cost of each iteration due to the reduction in the input size.

In order to assess the implications of Lemma 2.1 in practice, we have performed computational tests in which the simple elimination procedure proposed in this note was incorporated into each of the two algorithms in [2]. In our experiments, we checked the boundary condition (3) at an approximate minimum enclosing ball generated throughout either algorithm only if the right-hand side of (3) is sufficiently bounded away from zero. This strategy eliminates the computational cost of checking the boundary condition at an iterate where it would be unlikely to remove a large subset of input points. At iterate k, (3) is checked in our computational experiments only if $1 - (2\epsilon^k + (\epsilon^k)^2)^{1/2} > 0.55$, where 0.55 is a threshold value that was found to work well empirically.

The computational experiments were carried out on a 3.40 GHz Pentium IV processor with 1.0 GB RAM using MATLAB version R2006b on four different data sets. The first two data sets were randomly generated using different procedures outlined below. The last two sets consist of spherical or almost spherical input sets.

3.1. Random input sets. The first data set was randomly generated as in [2] with sizes (n, m) varying from $(10, 500)$ to $(100, 100000)$, while the second one was

TABLE 1
Computational results for the first data set ($\epsilon = 10^{-3}$).

		CPU time						Reduced input size	
n	m	A1	A1E	Speed-up	A2	A2E	Speed-up	A1E	A2E
10	500	0.0594	0.0541	1.10	0.0219	0.0156	1.40	124.2	99.8
10	1000	0.0694	0.0469	1.48	0.0297	0.0203	1.46	202.4	200.7
20	5000	2.2016	0.5078	4.34	0.3594	0.2172	1.65	420.4	330.3
20	10000	3.9844	0.5484	7.27	0.5641	0.1484	3.80	147.8	158.2
30	30000	14.1031	0.8516	16.56	2.8281	0.5562	5.08	121.1	107.3
50	50000	48.9359	5.3875	9.08	12.0109	4.1469	2.90	695.8	400.9
100	100000	141.6518	35.0223	4.04	62.692	30.5357	2.05	1626.2	1650.1

TABLE 2
Computational results for the second data set ($\epsilon = 10^{-3}$).

		CPU time						Reduced input size	
n	m	A1	A1E	Speed-up	A2	A2E	Speed-up	A1E	A2E
10	500	0.2016	0.1953	1.03	0.0250	0.0094	2.66	12.7	12.2
10	1000	0.2018	0.1469	1.37	0.0484	0.025	1.94	15.4	15
20	5000	3.0062	0.3281	9.16	0.475	0.1109	4.28	38.4	37
20	10000	5.0328	0.3312	15.20	0.9188	0.1812	5.07	42	40.9
30	30000	24.5359	1.2594	19.48	3.9656	0.9094	4.36	85.5	79.7
50	50000	52.8751	4.1865	12.63	13.0204	3.8463	3.39	202.2	213.4
100	100000	267.05	27.9984	9.54	56.1344	20.7188	2.71	430.9	423.8

generated using the standard normal distribution with the same sizes (n, m). We used $\epsilon = 10^{-3}$ for both data sets. For each fixed (n, m), ten different problem instances were generated for each data set. The computational results are reported in terms of averages over these instances in Table 1 and Table 2, each of which is divided into three sets of columns. The first set of columns reports the size (n, m). The second set of columns presents the results regarding the CPU time and is further divided into two parts, the first of which is devoted to the computational results related to [2, Algorithm 3.1] (an adaptation of the Frank–Wolfe algorithm to the minimum enclosing ball problem), while the second one displays those results using [2, Algorithm 4.1] (an adaptation of the Frank–Wolfe algorithm with *away steps* to the minimum enclosing ball problem). In the first part, A1 and A1E denote the CPU times in seconds using [2, Algorithm 3.1] without and with the elimination procedure, respectively, and speed-up denotes the resulting speed-up factor in running time due to the elimination procedure measured in terms of the ratio of A1 to A1E. Similarly, A2 and A2E denote the CPU times in seconds using [2, Algorithm 4.1] without and with the elimination procedure, respectively, and speed-up denotes the resulting speed-up factor in running time measured in terms of the ratio of A2 to A2E. The last set of columns reports the number of remaining input points upon termination using each algorithm with the elimination procedure.

As illustrated by Table 1 and Table 2, the incorporation of the elimination procedure into each of the two algorithms results in significant savings in running times especially for large instances where $m \gg n$. The procedure described in Lemma 2.1 identifies and eliminates 75% to 99% of the data points in our experiments, and the running times may improve by more than a factor of 19 on some instances. It is also worth noticing that the speed-up factors obtained from Algorithm 3.1 are generally considerably larger than those obtained with Algorithm 4.1. This may be due to the reason that the asymptotical linear convergence property of Algorithm 4.1 [2] already

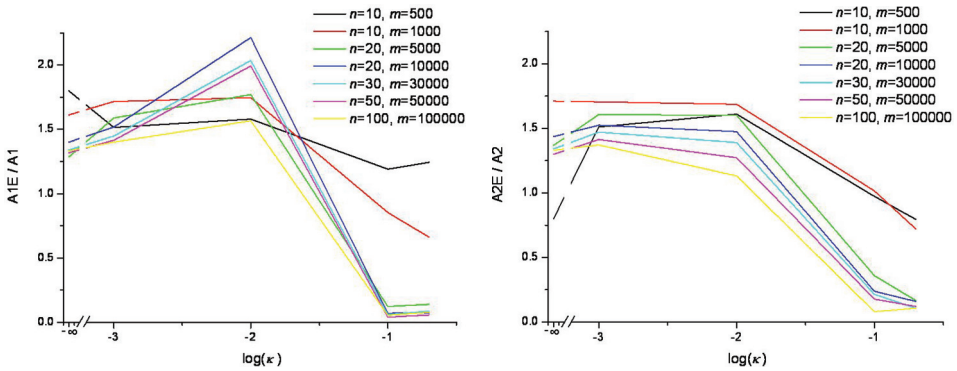


FIG. 1. *Experimental results for almost spherical input sets.*

results in significantly better performance compared to that of Algorithm 3.1, which may not leave much room for further improvement. Finally, we remark that the elimination procedure does not seem to have a noticeable effect on the core set sizes and on the number of iterations for either of the two algorithms.

3.2. Spherical and almost spherical input sets. In an attempt to assess the extent of extra overhead due to the elimination procedure, we considered data sets where all points lie on (or almost on) the unit sphere centered at the origin. An input set $\mathcal{A}$ is said to lie on a κ-approximate unit sphere centered at the origin, denoted by $\mathcal{S}_\kappa$, if $\mathcal{A} \subset \mathcal{S}_\kappa := \{x \in \mathbb{R}^n : 1 - \kappa \leq \|x\| \leq 1 + \kappa\}$. For an input set $\mathcal{A} \subset \mathcal{S}_\kappa$ where $\kappa \geq 0$ is small, the elimination procedure will keep testing input points for removal at each iteration but will be unable to remove a substantial subset of the input set. In the extreme case where $\kappa = 0$, none of the input points can be removed, since there would be no interior point. This extra overhead will necessarily result in an increase in the running time of an algorithm that uses the elimination procedure. We generated random input sets $\mathcal{A} \subset \mathcal{S}_\kappa$, where $\kappa \in \{0, 0.001, 0.01, 0.1, 0.2\}$, with sizes (n, m) varying from $(10, 500)$ to $(100, 100000)$ as in our experiments with the first two data sets. For each choice of experimental parameters, the computational results averaged over ten data sets are illustrated in Figure 1. The horizontal axis in each graph corresponds to κ using the logarithmic scale, while the vertical axis in the graph on the left (on the right) corresponds to the "slow-down" factor measured in terms of the ratio of the running time of Algorithm 3.1 (Algorithm 4.1) with the elimination procedure to the running time of the same algorithm without the elimination. A close examination of these two graphs reveals that the slow-down factors usually remain at an acceptable level especially for the faster Algorithm 4.1. Note that the elimination procedure leads to an extra overhead of at most 70% on all instances for Algorithm 4.1. A comparison of the slow-down and speed-up factors stemming from our experiments seems to justify the use of the elimination procedure, especially since spherical input sets would not likely be encountered in practical applications.

We also tested the two algorithms on data sets which consist of the vertices of the unit simplex where $n \in \{1000, 2500, 5000\}$. Note that each point in such an input set lies on the boundary of the minimum enclosing ball, and it is known that each point should be in the core set if $\epsilon \leq 1/n$ [1]. We tested each of the two algorithms with and without the elimination procedure using $\epsilon = 1/n$. The computational results are reported in Table 3, which is organized in a similar manner to that of Table 1. Note

TABLE 3
Computational results for the vertices of the unit simplex ($\epsilon = 1/n$).

n	m	CPU time					
		A1	A1E	A1E/A1	A2	A2E	A2E/A2
1000	1000	39.5312	53.625	1.357	40.078	54.219	1.352
2500	2500	251.75	336.7188	1.338	252.234	339.391	1.346
5000	5000	988.2812	1301.5312	1.317	983.25	1289.547	1.311

that the increase in the running time of each algorithm due to the inclusion of the elimination procedure is only around 35% for large spherical instances.

4. Concluding remarks. In this paper, we have described a procedure that identifies and eliminates data points that cannot lie on the boundary of the minimum enclosing ball of a finite set of points. This procedure can be easily incorporated into any iterative algorithm that generates a sequence of approximate minimum enclosing balls converging to the minimum enclosing ball of a given input set. Our computational results demonstrate the resulting significant improvements in the practical performance of the two algorithms proposed in [2] especially for randomly generated input sets. The extra overhead of the elimination procedure remains at an acceptable level for spherical or almost spherical input sets.

Furthermore, the same elimination procedure can also be incorporated into algorithms that can compute an approximate minimum enclosing ball of more general input sets such as a set of balls or ellipsoids for which the algorithms in [2] can still be applied. Such input sets can be viewed as an infinite set of points, and condition (3) essentially means that all input points that lie in the interior of a ball of a certain radius centered at the current approximate center c can be safely removed without affecting the optimal solution. In this case, an element of a more general input set (such as a ball or ellipsoid) can be completely removed if the furthest point on that element from the current approximate center c already lies in the interior of the aforementioned ball centered at c, which readily implies that every point on that element should necessarily violate (3). This may lead to considerable savings in the computation of minimum enclosing balls of more general input sets arising from practical applications.

Acknowledgments. We thank Mike Todd for encouraging us to prepare this manuscript. We gratefully acknowledge the thoughtful comments of two anonymous referees and the Associate Editor.

REFERENCES

[1] M. BĂDOIU AND K. L. CLARKSON, *Optimal core-sets for balls*, Comput. Geom., 40 (2008), pp. 14–22.

[2] E. A. YILDIRIM, *Two algorithms for the minimum enclosing ball problem*, SIAM J. Optim., 19 (2008), pp. 1368–1391.

SIAM J. OPTIM.
Vol. 19, No. 3, pp. 1397–1416

ITERATIVE MINIMIZATION SCHEMES FOR SOLVING THE SINGLE SOURCE LOCALIZATION PROBLEM*

AMIR BECK†, MARC TEBOULLE‡, AND ZAHAR CHIKISHEV§

Abstract. We consider the problem of locating a single radiating source from several noisy measurements using a maximum likelihood (ML) criteria. The resulting optimization problem is nonconvex and nonsmooth, and thus finding its global solution is in principle a hard task. Exploiting the special structure of the objective function, we introduce and analyze two iterative schemes for solving this problem. The first algorithm is a very simple explicit fixed-point-based formula, and the second is based on solving at each iteration a nonlinear least squares problem, which can be solved globally and efficiently after transforming it into an equivalent quadratic minimization problem with a single quadratic constraint. We show that the nonsmoothness of the problem can be avoided by choosing a specific "good" starting point for both algorithms, and we prove the convergence of the two schemes to stationary points. We present empirical results that support the underlying theoretical analysis and suggest that, despite of its nonconvexity, the ML problem can effectively be solved globally using the devised schemes.

Key words. single source location problem, Weiszfield algorithm, nonsmooth and nonconvex minimization, fixed-point methods, nonlinear least squares, generalized trust region, semidefinite relaxation

AMS subject classifications. 90C26, 90C22, 90C90

DOI. 10.1137/070698014

1. Introduction.

1.1. The source localization problem.

Consider the problem of locating a single radiating source from noisy range measurements collected using a network of passive sensors. More precisely, consider an array of m sensors, and let $\mathbf{a}_j \in \mathbb{R}^n$ denote the coordinates of the jth sensor.[1] Let $\mathbf{x} \in \mathbb{R}^n$ denote the unknown source's coordinate vector, and let $d_j > 0$ be a noisy observation of the range between the source and the jth sensor:

$$d_j = \|\mathbf{x} - \mathbf{a}_j\| + \varepsilon_j, \quad j = 1, \ldots, m, \tag{1.1}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_m)^T$ denotes the unknown noise vector. Such observations can be obtained, for example, from the time-of-arrival measurements in a constant-velocity propagation medium. The source localization problem is the following.

The source localization problem: Given the observed range measurements $d_j > 0$, find a "good" approximation of the source $\mathbf{x}$ satisfying (1.1).

*Received by the editors July 22, 2007; accepted for publication (in revised form) June 17, 2008; published electronically November 21, 2008. This research was partially supported by the Israel Science Foundation under ISF grant 489/06.

http://www.siam.org/journals/siopt/19-3/69801.html

†Department of Industrial Engineering and Management, Technion—Israel Institute of Technology, Haifa 32000, Israel (becka@ie.technion.ac.il).

‡School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel (teboulle@post.tau.ac.il).

§Department of Mathematics, Technion—Israel Institute of Technology, Haifa 32000, Israel (nosound@tx.technion.ac.il).

[1] In practical applications $n = 2$ or 3.

The source localization problem has received significant attention in the signal processing literature and specifically in the field of mobile phones localization [12, 5, 13]. It is also worth mentioning that the interest in wireless localization problems has increased since the first ruling of the Federal Communications Commission (FCC) for the detection of emergency calls in the United States in 1996 [17]. Currently, a high percentage of Enhanced 911 (E911) calls originate from mobile phones. Due to the unknown location of the wireless E911 calls, these calls do not receive the same quality of emergency assistance that fixed network 911 calls enjoy. To deal with this problem, the FCC issued an order on July 12, 1996, requiring all wireless service providers to report accurate mobile station location information to the E911 operator.

In addition to emergency management, mobile position information is also useful in mobile advertising, asset tracking, fleet management, location-sensitive billing [12], interactive map consultation, and monitoring of the mentally impaired [5].

1.2. The maximum likelihood criteria. In this paper we adopt the maximum-likelihood (ML) approach for solving the source localization problem (1.1); see, e.g., [4]. When $\boldsymbol{\varepsilon}$ follows a Gaussian distribution with a covariance matrix proportional to the identity matrix, the source $\mathbf{x}$ is the ML estimate that is the solution of the problem:

$$\text{(ML):} \quad \min_{\mathbf{x}\in\mathbb{R}^n} \left\{ f(\mathbf{x}) \equiv \sum_{j=1}^{m} (\|\mathbf{x}-\mathbf{a}_j\| - d_j)^2 \right\}. \tag{1.2}$$

Note that, in addition to the statistical interpretation, the latter problem is a least squares problem in the sense that it minimizes the squared sum of the errors.

An alternative approach for estimating the source location $\mathbf{x}$ is by solving the following least squares (LS) problem in the squared domain:

$$\text{(LS):} \quad \min_{\mathbf{x}\in\mathbb{R}^n} \sum_{j=1}^{m} \left(\|\mathbf{x}-\mathbf{a}_j\|^2 - d_j^2\right)^2. \tag{1.3}$$

Despite of its nonconvexity, the LS problem can be solved globally and efficiently by transforming it into a problem of minimizing a quadratic function subject to a single quadratic constraint [1] (more details will be given in section 3.2). However, the LS approach has two major disadvantages compared to the ML approach: first, the LS formulation lacks the statistical interpretation of the ML problem. Second, as demonstrated by the numerical simulations in section 4, the LS estimate provides less accurate solutions than those provided by the the ML approach.

The ML problem, like the LS problem, is nonconvex. However, as opposed to the LS problem for which a global solution can be computed efficiently [1], the ML problem seems to be a difficult problem to solve efficiently. A possible reason for the increased difficulty of the ML problem is its nonsmoothness. One approach for approximating the solution of the ML problem is via semidefinite relaxation (SDR) [4, 1]. We also note that the source localization problem formulated as (ML) can be viewed as a special instance of sensor network localization problems in which several sources are present; see, for example, the recent work in [3]; for this class of problems, semidefinite programming-based algorithms have been developed.

In this paper we depart from the SDR techniques and seek other efficient approaches to solve the ML problem. This is achieved by exploiting the special structure of the objective function which allows us to devise fixed-point-based iterative schemes

for solving the nonsmooth and nonconvex ML problem (1.2). The first scheme admits a very simple explicit iteration formula given by

$$\mathbf{x}^{k+1} = \mathcal{M}_1\left(\mathbf{x}^k, \mathbf{a}\right) \text{ (where } \mathbf{a} \equiv (\mathbf{a}_1, \ldots, \mathbf{a}_m)),$$

while the second iterative scheme is of the form

$$\mathbf{x}^{k+1} \in \operatorname*{argmin}_{\mathbf{x}} \mathcal{M}_2\left(\mathbf{x}, \mathbf{x}^k, \mathbf{a}\right)$$

and requires the solution of an additional subproblem which will be shown to be efficiently solved. The main goals of this paper are to introduce the building mechanism of these two schemes, to develop and analyze their convergence properties, and to demonstrate their computational viability for solving the ML problem (1.2), as well as their effectiveness when compared with the LS and SDR approaches.

1.3. Paper layout. In the next section, we present and analyze the first scheme, which is a simple fixed-point-based method. The second algorithm, which is based on solving a sequence of least squares problems of a similar structure to that of (1.3), is presented and analyzed in section 3. The construction of both methods is motivated by two different interpretations of the well-known Weiszfeld method for the Fermat–Weber location problem [16]. For both schemes, we show that the nonsmoothness of the problem can be avoided by choosing a specific "good" starting point. Empirical results presented in section 4 provide a comparison between the two devised algorithms, as well as a comparison to different approaches such as LS and SDR. In particular, the numerical results suggest that, despite its nonconvexity, the ML problem can, for all practical purposes, be globally solved using the devised methods.

1.4. Notation. Throughout the paper, the following notation is used: vectors are denoted by boldface lowercase letters, e.g., $\mathbf{y}$, and matrices by boldface uppercase letters, e.g., $\mathbf{A}$. The ith component of a vector $\mathbf{y}$ is written as y_i. Given two matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} \succ \mathbf{B}$ ($\mathbf{A} \succeq \mathbf{B}$) means that $\mathbf{A} - \mathbf{B}$ is positive definite (semidefinite). The directional derivative of a function $f : \mathbb{R}^n \to \mathbb{R}$ at $\bar{\mathbf{x}}$ in the direction $\mathbf{v}$ is defined (if it exists) by

$$f'(\mathbf{x}; \mathbf{v}) \equiv \lim_{t \to 0^+} \frac{f(\bar{\mathbf{x}} + t\mathbf{v}) - f(\bar{\mathbf{x}})}{t}. \tag{1.4}$$

The α-level set of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined by $\mathrm{Lev}(f, \alpha) = \{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) \leq \alpha\}$. The collection of m sensors $\{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$ is denoted by $\mathcal{A}$.

2. A simple fixed-point algorithm. In this section we introduce a simple fixed-point algorithm that is designed to solve the ML problem (1.2). The algorithm is inspired by the celebrated Weiszfeld algorithm for the Fermat–Weber problem, which is briefly recalled in section 2.1. In section 2.2 we introduce and analyze the fixed-point scheme designed to solve the ML problem.

2.1. A small detour: Weiszfeld algorithm for the Fermat–Weber problem. As was already mentioned, the ML problem (1.2) is nonconvex and nonsmooth, and thus finding its exact solution is in principle a difficult task. We propose a fixed-point scheme motivated by the celebrated Weiszfeld algorithm [16, 7] for solving the Fermat–Weber location problem:

$$\min_{\mathbf{x}} \left\{ s(\mathbf{x}) \equiv \sum_{j=1}^{m} \omega_j \|\mathbf{x} - \mathbf{a}_j\| \right\}, \tag{2.1}$$

where $\omega_j > 0$ and $\mathbf{a}_j \in \mathbb{R}^n$ for $j = 1, \ldots, m$. Of course, the Fermat–Weber problem is much easier to analyze and solve than the ML problem (1.2) since it is a well-structured nonsmooth convex minimization problem. This problem has been extensively studied in the location theory literature; see, for instance, [11]. Our objective here is to mimic the Weiszfeld algorithm [16] to obtain an algorithm for solving the nonsmooth and nonconvex ML problem (1.2). The Weiszfeld method is a very simple fixed-point scheme that is designed to solve the Fermat–Weber problem. One way to derive it is to write the first order global optimality conditions for the convex problem (2.1)

$$\nabla s(\mathbf{x}) = \sum_{j=1}^{m} \omega_j \frac{\mathbf{x} - \mathbf{a}_j}{\|\mathbf{x} - \mathbf{a}_j\|} = 0 \ \forall \mathbf{x} \notin \mathcal{A}$$

as

$$\mathbf{x} = \frac{\sum_{j=1}^{m} \omega_j \frac{\mathbf{a}_j}{\|\mathbf{x}-\mathbf{a}_j\|}}{\sum_{j=1}^{m} \frac{\omega_j}{\|\mathbf{x}-\mathbf{a}_j\|}},$$

which naturally calls for the iterative scheme

$$\mathbf{x}^{k+1} = \frac{\sum_{j=1}^{m} \omega_j \frac{\mathbf{a}_j}{\|\mathbf{x}^k-\mathbf{a}_j\|}}{\sum_{j=1}^{m} \frac{\omega_j}{\|\mathbf{x}^k-\mathbf{a}_j\|}}. \tag{2.2}$$

For the convergence analysis of the Weiszfeld algorithm (2.2) and modified versions of the algorithm, see, e.g., [10, 15], and references therein.

2.2. The simple fixed-point algorithm: Definition and analysis. Similarly to the Weiszfeld method, our starting point for constructing a fixed-point algorithm to solve the ML problem is by writing the optimality conditions. Assuming that $\mathbf{x} \notin \mathcal{A}$ we have that $\mathbf{x}$ is a stationary point for problem (ML) if and only if

$$\nabla f(\mathbf{x}) = 2 \sum_{j=1}^{m} (\|\mathbf{x} - \mathbf{a}_j\| - d_j) \frac{\mathbf{x} - \mathbf{a}_j}{\|\mathbf{x} - \mathbf{a}_j\|} = \mathbf{0}, \tag{2.3}$$

which can be written as

$$\mathbf{x} = \frac{1}{m} \left\{ \sum_{j=1}^{m} \mathbf{a}_j + \sum_{j=1}^{m} d_j \frac{\mathbf{x} - \mathbf{a}_j}{\|\mathbf{x} - \mathbf{a}_j\|} \right\}.$$

The latter relation calls for the following fixed-point algorithm, which we term the *standard fixed point (SFP) scheme*.

ALGORITHM SFP.

$$\mathbf{x}^{k+1} = \frac{1}{m} \left\{ \sum_{j=1}^{m} \mathbf{a}_j + \sum_{j=1}^{m} d_j \frac{\mathbf{x}^k - \mathbf{a}_j}{\|\mathbf{x}^k - \mathbf{a}_j\|} \right\}, \quad k \geq 0. \tag{2.4}$$

Like in the Weiszfeld algorithm, the SFP scheme is not well defined if $\mathbf{x}^k \in \mathcal{A}$ for some k. In what follows we will show that by carefully selecting the initial vector $\mathbf{x}^0$

we can *guarantee* that the iterates are not in the sensors set $\mathcal{A}$, therefore establishing that the method is well defined. At this juncture, it is interesting to notice that the approach we suggest here for dealing with the points of nonsmoothness that occur at $\mathbf{x}^k \in \mathcal{A}$ is quite different from the common approaches for handling the nonsmoothness. For example, in order to avoid the nondifferentiable points of the Fermat–Weber objective function, several modifications of the Weiszfeld method were proposed; see, e.g., [10, 15], and references therein. However, there do not seem to have been any attempts in the literature to choose good initial starting points to avoid the nonsmoothness difficulty. A constructive procedure for choosing a good starting point for the SFP method will be given at the end of this section.

Before proceeding with the analysis of the SFP method, we record the fact that, much like the Weiszfeld algorithm (see [7]), the SFP scheme is a gradient method with a fixed step size.

PROPOSITION 2.1. *Let* $\{\mathbf{x}^k\}$ *be the sequence generated by the SFP method* (2.4)*, and suppose that* $\mathbf{x}^k \notin \mathcal{A}$ *for all* $k \geq 0$*. Then*

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{1}{2m}\nabla f\left(\mathbf{x}^k\right). \tag{2.5}$$

Proof. The proof follows by a straightforward calculation, using the gradient of f computed in (2.3). ∎

A gradient method does not necessarily converge without additional assumptions (e.g., assuming that ∇f is Lipschitz continuous and/or using a line search [2]). Nevertheless, we show below that scheme (2.4) *does* converge.

By Proposition 2.1 the SFP method can be compactly written as

$$\mathbf{x}^{k+1} = T\left(\mathbf{x}^k\right), \tag{2.6}$$

where $T : \mathbb{R}^n \setminus \mathcal{A} \to \mathbb{R}^n$ is the operator defined by

$$T(\mathbf{x}) = \mathbf{x} - \frac{1}{2m}\nabla f(\mathbf{x}). \tag{2.7}$$

In the convergence analysis of the SFP method, we will also make use of the auxiliary function:

$$h(\mathbf{x}, \mathbf{y}) \equiv \sum_{j=1}^{m} \|\mathbf{x} - \mathbf{a}_j - d_j r_j(\mathbf{y})\|^2 \quad \forall \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^n \setminus \mathcal{A}, \tag{2.8}$$

where

$$r_j(\mathbf{y}) \equiv \frac{\mathbf{y} - \mathbf{a}_j}{\|\mathbf{y} - \mathbf{a}_j\|}, \quad j = 1, \ldots, m.$$

Note that for every $\mathbf{y} \notin \mathcal{A}$, the following relations hold for every $j = 1, \ldots, m$:

$$\|r_j(\mathbf{y})\| = 1, \tag{2.9}$$

$$(\mathbf{y} - \mathbf{a}_j)^T r_j(\mathbf{y}) = \|\mathbf{y} - \mathbf{a}_j\|. \tag{2.10}$$

In Lemma 2.1 below, we prove several key properties of the auxiliary function h defined in (2.8).

LEMMA 2.1.
(a) $h(\mathbf{x},\mathbf{x}) = f(\mathbf{x})$ *for every* $\mathbf{x} \notin \mathcal{A}$.
(b) $h(\mathbf{x},\mathbf{y}) \geq f(\mathbf{x})$ *for every* $\mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^n \setminus \mathcal{A}$.
(c) *If* $\mathbf{y} \notin \mathcal{A}$, *then*

$$T(\mathbf{y}) = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} h(\mathbf{x},\mathbf{y}). \tag{2.11}$$

Proof. (a) For every $\mathbf{x} \notin \mathcal{A}$,

$$\begin{aligned} f(\mathbf{x}) &= \sum_{j=1}^{m} (\|\mathbf{x}-\mathbf{a}_j\| - d_j)^2 \\ &= \sum_{j=1}^{m} (\|\mathbf{x}-\mathbf{a}_j\|^2 - 2d_j\|\mathbf{x}-\mathbf{a}_j\| + d_j^2) \\ &\overset{(2.9),(2.10)}{=} \sum_{j=1}^{m} (\|\mathbf{x}-\mathbf{a}_j\|^2 - 2d_j(\mathbf{x}-\mathbf{a}_j)^T r_j(\mathbf{x}) + d_j^2\|r_j(\mathbf{x})\|^2) = h(\mathbf{x},\mathbf{x}), \end{aligned}$$

where the last equation follows from (2.8).

(b) Using the definition of f and h given in (1.2) and (2.8), respectively, and the fact (2.9), a short computation shows that for every $\mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^n \setminus \mathcal{A}$,

$$\begin{aligned} h(\mathbf{x},\mathbf{y}) - f(\mathbf{x}) &= 2\sum_{j=1}^{m} d_j \left(\|\mathbf{x}-\mathbf{a}_j\| - (\mathbf{x}-\mathbf{a}_j)^T r_j(\mathbf{y})\right) \\ &\geq 0, \end{aligned}$$

where the last inequality follows from the Cauchy–Schwarz inequality and using again (2.9).

(c) For any $\mathbf{y} \in \mathbb{R}^n \setminus \mathcal{A}$, the function $\mathbf{x} \mapsto h(\mathbf{x},\mathbf{y})$ is strictly convex on $\mathbb{R}^n$ and consequently admits a unique minimizer $\mathbf{x}^*$ satisfying

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

Using the definition of h given in (2.8), the latter identity can be explicitly written as

$$\sum_{j=1}^{m} (\mathbf{x}^* - \mathbf{a}_j - d_j r_j(\mathbf{y})) = \mathbf{0},$$

which by simple algebraic manipulation can be shown to be equivalent to $\mathbf{x}^* = \mathbf{y} - \frac{1}{2m}\nabla f(\mathbf{y})$, establishing that $\mathbf{x}^* = T(\mathbf{y})$. ∎

Using Lemma 2.1 we are now able to prove the monotonicity property of the operator T with respect to f.

LEMMA 2.2. *Let* $\mathbf{y} \notin \mathcal{A}$. *Then*

$$f(T(\mathbf{y})) \leq f(\mathbf{y}),$$

and equality holds if and only if $T(\mathbf{y}) = \mathbf{y}$.

Proof. By (2.11) and the strict convexity of the function $\mathbf{x} \mapsto h(\mathbf{x},\mathbf{y})$, one has

$$h(T(\mathbf{y}),\mathbf{y}) < h(\mathbf{x},\mathbf{y}) \text{ for every } \mathbf{x} \neq T(\mathbf{y}).$$

In particular, if $T(\mathbf{y}) \neq \mathbf{y}$, then

$$h(T(\mathbf{y}), \mathbf{y}) < h(\mathbf{y}, \mathbf{y}) = f(\mathbf{y}), \tag{2.12}$$

where the last equality follows from Lemma 2.1(a). By Lemma 2.1(b), $h(T(\mathbf{y}), \mathbf{y}) \geq f(T(\mathbf{y}))$, which, combined with (2.12), establishes the desired strict monoton icity. ∎

Theorem 2.1 given below states the basic convergence results for the SFP method. In the proof, we exploit the boundedness of the level sets of the objective function f, which is recorded in the following lemma.

LEMMA 2.3. *The level sets of f are bounded.*

Proof. The proof follows immediately from the fact that $f(\mathbf{x}) \to \infty$ as $\|\mathbf{x}\| \to \infty$. ∎

THEOREM 2.1 (convergence of the SFP method). *Let $\{\mathbf{x}^k\}$ be generated by* (2.4) *such that $\mathbf{x}^0$ satisfies*

$$f\left(\mathbf{x}^0\right) < \min_{j=1,\ldots,m} f(\mathbf{a}_j). \tag{2.13}$$

Then

(a) *$\mathbf{x}^k \notin \mathcal{A}$ for every $k \geq 0$;*

(b) *for every $k \geq 0$, $f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k)$, and equality is satisfied if and only if $\mathbf{x}^{k+1} = \mathbf{x}^k$.*

(c) *the sequence of function values $\{f(\mathbf{x}^k)\}$ converges;*

(d) *the sequence $\{\mathbf{x}^k\}$ is bounded;*

(e) *every convergent subsequence $\{\mathbf{x}^{k_l}\}$ satisfies $\mathbf{x}^{k_l+1} - \mathbf{x}^{k_l} \to \mathbf{0}$;*

(f) *any limit point of $\{\mathbf{x}^k\}$ is a stationary point of f.*

Proof. (a) and (b) The proof follows by induction on k using Lemma 2.2.

(c) The proof readily follows from the monotonicity and lower boundedness (by zero) of the sequence $\{f(\mathbf{x}^k)\}$.

(d) By (b), all of the iterates $\mathbf{x}^k$ are in the level set $\mathrm{Lev}(f, f(\mathbf{x}^0))$ which, by Lemma 2.3, establishes the boundedness of the sequence $\{\mathbf{x}^k\}$.

(e) and (f) Let $\{\mathbf{x}^{k_l}\}$ be a convergent subsequence of $\{\mathbf{x}^k\}$ with limit point $\mathbf{x}^*$. Since $f(\mathbf{x}^{k_l}) \leq f(\mathbf{x}^0) < \min_{j=1,\ldots,m} f(\mathbf{a}_j)$, it follows by the continuity of f that $f(\mathbf{x}^*) \leq f(\mathbf{x}^0) < \min_{j=1,\ldots,m} f(\mathbf{a}_j)$, proving that $\mathbf{x}^* \notin \mathcal{A}$. By (2.6)

$$\mathbf{x}^{k_l+1} = T\left(\mathbf{x}^{k_l}\right). \tag{2.14}$$

Therefore, since the subsequence $\{\mathbf{x}^{k_l}\}$ and its limit point $\mathbf{x}^*$ are not in $\mathcal{A}$, by the continuity of ∇f on $\mathbb{R}^n \setminus \mathcal{A}$, we conclude that the subsequence $\{\mathbf{x}^{k_l+1}\}$ converges to a vector $\bar{\mathbf{x}}$ satisfying

$$\bar{\mathbf{x}} = T(\mathbf{x}^*). \tag{2.15}$$

To prove (e), we need to show that $\bar{\mathbf{x}} = \mathbf{x}^*$. Since both $\mathbf{x}^*$ and $\bar{\mathbf{x}}$ are limit points of $\{\mathbf{x}^k\}$ and since the sequence of function values converges (by (c)), then the continuity of f over $\mathbb{R}^n$ implies that $f(\mathbf{x}^*) = f(\bar{\mathbf{x}})$. Invoking Lemma 2.2 for $\mathbf{y} = \mathbf{x}^*$, we conclude that $\bar{\mathbf{x}} = \mathbf{x}^*$, proving claim (e). Part (f) follows from the observation that the equality $\mathbf{x}^* = T(\mathbf{x}^*)$ is equivalent (by the definition of T) to $\nabla f(\mathbf{x}^*) = \mathbf{0}$. ∎

Remark 2.1. It is easy to find a vector $\mathbf{x}^0$ satisfying condition (2.13). For example, Procedure INIT, that will be described at the end of this section, produces a point satisfying (2.13).

Combining claims (c) and (f) of Theorem 2.1, we immediately obtain convergence of the sequence of function values.

COROLLARY 2.1. *Let $\{\mathbf{x}^k\}$ be the sequence generated by the SFP algorithm satisfying* (2.13). *Then $f(\mathbf{x}^k) \to f^*$, where f^* is the function value at a stationary point of f.*

We were able to prove the convergence of the function values of the sequence. The situation is more complicated for the sequence itself, where we were able only to show that all limit points are stationary points. We can prove convergence of the sequence itself if we assume that all stationary points of the objective function are isolated.[2] The proof of this claim strongly relies on the following lemma from [8].

LEMMA 2.4 (see [8, Lemma 4.10]). *Let $\mathbf{x}^*$ be an isolated limit point of a sequence $\{\mathbf{x}^k\}$ in $\mathbb{R}^n$. If $\{\mathbf{x}^k\}$ does not converge, then there is a subsequence $\{\mathbf{x}^{k_l}\}$ which converges to $\mathbf{x}^*$ and an $\epsilon > 0$ such that $\|\mathbf{x}^{k_l+1} - \mathbf{x}^{k_l}\| \geq \epsilon$.*

We can now use the above lemma to prove a convergence result under the assumption that all stationary points of f are isolated.

THEOREM 2.2 (convergence of the sequence). *Let $\{\mathbf{x}^k\}$ be generated by* (2.4) *such that $\mathbf{x}^0$ satisfies* (2.13). *Suppose further that all stationary points of f are isolated. Then the sequence $\{\mathbf{x}^k\}$ converges to a stationary point.*

Proof. Let $\mathbf{x}^*$ be a limit point of $\{\mathbf{x}^k\}$ (its existence follows from the boundedness of the sequence proved in Theorem 2.1(d)). By our assumption $\mathbf{x}^*$ is an isolated point. Suppose in contradiction that the sequence does not converge. Then by Lemma 2.4 there exists a subsequence $\{\mathbf{x}^{k_l}\}$ that converges to $\mathbf{x}^*$ satisfying $\|\mathbf{x}^{k_l+1} - \mathbf{x}^{k_l}\| \geq \epsilon$. However, this is in contradiction to (e) of Theorem 2.1. We thus conclude that $\{\mathbf{x}^k\}$ converges to a stationary point. ∎

The analysis of the SFP method relies on the validity of condition (2.13) on the starting point $\mathbf{x}^0$. We will now show that, thanks to the special structure of the objective function (ML), we can compute such a point through a simple procedure. This is achieved by establishing the following result.

LEMMA 2.5. *Let $\mathcal{A} \equiv \{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$ be the given set of m sensors, and let*

$$g_j(\mathbf{x}) = \sum_{i=1, i\neq j}^{m} (\|\mathbf{x} - \mathbf{a}_i\| - d_i)^2, \quad j = 1, \ldots, m.$$

Then for every $j = 1, \ldots, m$ the following apply:

(i) *If $\nabla g_j(\mathbf{a}_j) \neq \mathbf{0}$, then $f'(\mathbf{a}_j; -\nabla g_j(\mathbf{a}_j)) < 0$. Otherwise, if $\nabla g_j(\mathbf{a}_j) = \mathbf{0}$, then $f'(\mathbf{a}_j; \mathbf{v}) < 0$ for every $\mathbf{v} \neq \mathbf{0}$. In particular, there exists a descent direction from every sensor point.*

(ii) *Every $\bar{\mathbf{x}} \in \mathcal{A}$ is not a local optimum for the ML problem* (1.2).

Proof. (i) For convenience, for every $j = 1, \ldots, m$ we denote

$$f_j(\mathbf{x}) = (\|\mathbf{x} - \mathbf{a}_j\| - d_j)^2 \tag{2.16}$$

so that the objective function of problem (ML) can be written as

$$f(\mathbf{x}) = f_j(\mathbf{x}) + g_j(\mathbf{x}) \tag{2.17}$$

for every $\mathbf{x} \in \mathbb{R}^n$ and $j = 1, \ldots, m$. Note that f is not differentiable for every $\mathbf{x} \in \mathcal{A}$. Nonetheless, the directional derivative of f at $\mathbf{x}$ in the direction $\mathbf{v} \in \mathbb{R}^n$ always exists

[2] We say that $\mathbf{x}^*$ is an isolated stationary point of f, if there are no other stationary points in some neighborhood of $\mathbf{x}^*$.

and is given by

$$f'(\bar{\mathbf{x}};\mathbf{v}) = \begin{cases} \nabla f(\bar{\mathbf{x}})^T\mathbf{v}, & \bar{\mathbf{x}} \notin \mathcal{A}, \\ \nabla g_j(\mathbf{a}_j)^T\mathbf{v} - 2d_j\|\mathbf{v}\|, & \bar{\mathbf{x}} = \mathbf{a}_j. \end{cases} \tag{2.18}$$

Indeed, the above formula for $\bar{\mathbf{x}} \notin \mathcal{A}$ is obvious. In the other case, suppose then that $\bar{\mathbf{x}} = \mathbf{a}_j$ for some $j \in \{1,\ldots,m\}$. Noting that g_j is differentiable at $\mathbf{a}_j$, we have $g_j'(\mathbf{a}_j;\mathbf{v}) = \nabla g_j(\mathbf{a}_j)^T\mathbf{v}$, and using definition (2.16) for f_j, we get $f_j'(\mathbf{a}_j;\mathbf{v}) = -2d_j\|\mathbf{v}\|$, and hence with (2.17), we obtain the desired formula (2.18) for $f'(\mathbf{a}_j;\mathbf{v})$. Finally, if $\nabla g_j(\mathbf{a}_j) \neq \mathbf{0}$, then using (2.18) we have

$$f'(\mathbf{a}_j; -\nabla g_j(\mathbf{a}_j)) = -\|\nabla g_j(\mathbf{a}_j)\|^2 - 2d_j\|\nabla g_j(\mathbf{a}_j)\| < 0.$$

Otherwise, if $\nabla g_j(\mathbf{a}_j) = \mathbf{0}$, then for every $\mathbf{v} \neq \mathbf{0}$ we have

$$f'(\mathbf{a}_j;\mathbf{v}) = -2d_j\|\mathbf{v}\| < 0.$$

(ii) By part (i) there exists a descent direction from every sensor point $\bar{\mathbf{x}} \in \mathcal{A}$. Therefore, none of the sensor points can be a local optimum for problem (ML). ∎

Using the descent directions provided by Lemma 2.5, we can compute a point $\bar{\mathbf{x}}$ satisfying

$$f(\bar{\mathbf{x}}) < \min_{j=1,\ldots,m} f(\mathbf{a}_j)$$

by the following procedure.

PROCEDURE INIT.

1. $t = 1$.
2. **Set** k to be an index for which $f(\mathbf{a}_k) = \min_{j=1,\ldots,m} f(\mathbf{a}_j)$.
3. **Set**

$$\mathbf{v}_0 = \begin{cases} -\nabla g_k(\mathbf{a}_k), & \nabla g_k(\mathbf{a}_k) \neq \mathbf{0}, \\ \mathbf{e}, & \nabla g_k(\mathbf{a}_k) = \mathbf{0}, \end{cases} \tag{2.19}$$

   where $\mathbf{e}$ is the vector of all ones.[3]
4. **While** $f(\mathbf{a}_k + t\mathbf{v}_0) \geq f(\mathbf{a}_k)$, **set** $t = t/2$. **End**
5. The output of the algorithm is $\mathbf{a}_k + t\mathbf{v}_0$.

The validity of this procedure stems from the fact that, by Lemma 2.5, the direction $\mathbf{v}_0$ defined in (2.19) is always a descent direction.

One of the advantages of the SFP scheme is its simplicity. However, the SFP method, being a gradient method, does have the tendency to converge to local minima. In the next section we will present a second and more involved algorithm to solve the ML problem. As we shall see in the numerical examples presented in section 4, the empirical performance of this second iterative scheme is significantly better than that of the SFP, both with respect to the number of required iterations and with respect to the probability of getting stuck in a local/nonglobal point.

3. A sequential weighted least squares algorithm. In this section we study a different method for solving the ML problem (1.2), which we call the sequential weighted least squares (SWLS) algorithm. The SWLS algorithm is also motivated by the construction of the Weiszfeld method, but from a different viewpoint; see section 3.1. Each iteration of the method consists of solving a nonlinear least squares problem, whose solution is found by the approach discussed in section 3.2. The convergence analysis of the SWLS algorithm is given in section 3.3.

[3] We could have chosen any other nonzero vector.

3.1. The SWLS algorithm. To motivate the SWLS algorithm, let us first go back to the Weiszfeld scheme for solving the classical Fermat–Weber location problem, whereby we rewrite the iterative scheme (2.2) in the following equivalent, but different, way:

$$\mathbf{x}^{k+1} = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} \left\{ \sum_{j=1}^{m} \omega_j \frac{\|\mathbf{x}-\mathbf{a}_j\|^2}{\|\mathbf{x}^k-\mathbf{a}_j\|} \right\}. \tag{3.1}$$

The strong convexity of the objective function in (3.1) (recall that $\omega_j > 0$ for all j) implies that $\mathbf{x}^{k+1}$ is uniquely defined as a function of $\mathbf{x}^k$. Therefore, the Weiszfeld method (2.2) for solving problem (2.1) can also be written as

$$\mathbf{x}^{k+1} = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} q\left(\mathbf{x}, \mathbf{x}^k\right),$$

where

$$q(\mathbf{x},\mathbf{y}) \equiv \sum_{j=1}^{m} \omega_j \frac{\|\mathbf{x}-\mathbf{a}_j\|^2}{\|\mathbf{y}-\mathbf{a}_j\|} \text{ for every } \mathbf{x}\in\mathbb{R}^n, \mathbf{y}\in\mathbb{R}^n\setminus\mathcal{A}.$$

The auxiliary function q was essentially constructed from the objective function s of the Fermat–Weber location problem, by replacing the norm terms $\|\mathbf{x}-\mathbf{a}_j\|$ with $\frac{\|\mathbf{x}-\mathbf{a}_j\|^2}{\|\mathbf{y}-\mathbf{a}_j\|}$, i.e., with $s(\mathbf{x}) = q(\mathbf{x},\mathbf{x})$. Mimicking this observation for the ML problem under study, we will use an auxiliary function in which each norm term $\|\mathbf{x}-\mathbf{a}_j\|$ in the objective function (1.2) is replaced with $\frac{\|\mathbf{x}-\mathbf{a}_j\|^2}{\|\mathbf{y}-\mathbf{a}_j\|}$, resulting in the following auxiliary function:

$$g(\mathbf{x},\mathbf{y}) \equiv \sum_{i=1}^{m} \left(\frac{\|\mathbf{x}-\mathbf{a}_i\|^2}{\|\mathbf{y}-\mathbf{a}_i\|} - d_i \right)^2, \quad \mathbf{x}\in\mathbb{R}^n, \mathbf{y}\in\mathbb{R}^n\setminus\mathcal{A}. \tag{3.2}$$

The general step of the algorithm for solving problem (ML), the SWLS method, is now given by

$$\mathbf{x}^{k+1} \in \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} g\left(\mathbf{x}, \mathbf{x}^k\right)$$

or more explicitly by the following algorithm.

ALGORITHM SWLS.

$$\mathbf{x}^{k+1} \in \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} \sum_{j=1}^{m} \left(\frac{\|\mathbf{x}-\mathbf{a}_j\|^2}{\|\mathbf{x}^k-\mathbf{a}_j\|} - d_j \right)^2. \tag{3.3}$$

The name SWLS stems from the fact that at each iteration k we are required to solve the following weighted least squares (WLS) version of the LS problem (1.3):

$$\text{(WLS):} \quad \min_{\mathbf{x}} \sum_{j=1}^{m} \omega_j^k \left(\|\mathbf{x}-\mathbf{c}_j\|^2 - \beta_j^k \right)^2, \tag{3.4}$$

with

$$\mathbf{c}_j = \mathbf{a}_j, \qquad \beta_j^k = d_j \|\mathbf{x}^k - \mathbf{a}_j\|, \qquad \omega_j^k = \frac{1}{\|\mathbf{x}^k - \mathbf{a}_j\|^2}. \tag{3.5}$$

Note that the SWLS algorithm as presented above is not defined for iterations in which $\mathbf{x}^k \in \mathcal{A}$. In our random numerical experiments (cf. section 4) this situation never occurred; i.e., $\mathbf{x}^k$ did not belong to $\mathcal{A}$ for every k. However, from a theoretical point of view this issue must be resolved. Similarly to the methodology advocated in the convergence analysis of the SFP method, our approach for avoiding the sensor points $\mathcal{A}$ is by choosing a "good enough" initial vector. In section 3.3, we introduce a simple condition on the initial vector $\mathbf{x}^0$ under which the algorithm is well defined and proven to converge.

3.2. Solving the WLS subproblem. We will now show how the WLS subproblem (3.4) can be solved globally and efficiently by transforming it into a problem of minimizing a quadratic function subject to a single quadratic constraint. This derivation is a straightforward extension of the solution technique devised in [1] and is briefly described here for completeness.

For a given fixed k (for simplicity we omit the index k below), we first transform (3.4) into a constrained minimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n,\alpha\in\mathbb{R}} \left\{ \sum_{j=1}^{m} \omega_j \left(\alpha - 2\mathbf{c}_j^T\mathbf{x} + \|\mathbf{c}_j\|^2 - \beta_j\right)^2 : \|\mathbf{x}\|^2 = \alpha \right\}, \tag{3.6}$$

which can also be written as (using the substitution $\mathbf{y} = (\mathbf{x}^T, \alpha)^T$)

$$\min_{\mathbf{y}\in\mathbb{R}^{n+1}} \left\{ \|\mathbf{A}\mathbf{y} - \mathbf{b}\|^2 : \mathbf{y}^T\mathbf{D}\mathbf{y} + 2\mathbf{f}^T\mathbf{y} = 0 \right\}, \tag{3.7}$$

where

$$\mathbf{A} = \begin{pmatrix} -2\sqrt{\omega_1}\mathbf{c}_1^T & \sqrt{\omega_1} \\ \vdots & \vdots \\ -2\sqrt{\omega_m}\mathbf{c}_m^T & \sqrt{\omega_m} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} \sqrt{\omega_1}\left(\beta_1 - \|\mathbf{c}_1\|^2\right) \\ \vdots \\ \sqrt{\omega_m}\left(\beta_m - \|\mathbf{c}_m\|^2\right) \end{pmatrix}$$

and

$$\mathbf{D} = \begin{pmatrix} \mathbf{I}_n & \mathbf{0}_{n\times 1} \\ \mathbf{0}_{1\times n} & 0 \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} 0 \\ -0.5 \end{pmatrix}.$$

Note that (3.7) belongs to the class of problems consisting of minimizing a quadratic function subject to a single quadratic constraint. Problems of this type are called generalized trust region subproblems (GTRS). GTRS problems possess necessary and sufficient optimality conditions from which efficient solution methods can be derived; see, e.g., [6, 9].

The SWLS scheme is of course more involved than the simpler SFP scheme. However, as explained above, the additional computations required in SWLS to solve the subproblem can be done efficiently and are worthwhile, since the SWLS algorithm usually possesses a much larger region of convergence to the global minimum than the SFP scheme, which in turn implies that it has the tendency of avoiding local minima and a greater chance of hitting the global minimum. This will be demonstrated on the numerical examples given in section 4.

TABLE 1
Number of runs (out of 10000*) for which Assumption* 2 *is satisfied for* $\mathbf{x}^0 = \mathbf{x}_{\rm LS}$.

σ	1e-3	1e-2	1e-1	1e+0
N_σ	10000	10000	9927	6281

3.3. Convergence analysis of the SWLS method. In this section we provide an analysis of the SWLS method. We begin by presenting our underlying assumptions in section 3.3.1, and in section 3.3.2 we prove the convergence results of the method.

3.3.1. Underlying assumptions. The following assumption will be made throughout this section.

Assumption 1. The matrix

$$\mathbf{A} = \begin{pmatrix} 1 & \mathbf{a}_1^T \\ 1 & \mathbf{a}_2^T \\ \vdots & \vdots \\ 1 & \mathbf{a}_m^T \end{pmatrix}$$

is of full column rank.

This assumption is equivalent to saying that $\mathbf{a}_1, \ldots, \mathbf{a}_m$ do not reside in a lower-dimensional affine space (i.e., a line if $n = 2$ and a plane if $n = 3$).

To guarantee the well definiteness of the SWLS algorithm (i.e., $\mathbf{x}^k \notin \mathcal{A}$ for all k), we will make the following assumption on the initial vector $\mathbf{x}^0$.

Assumption 2. $\mathbf{x}^0 \in \mathcal{R}$. where

$$\mathcal{R} := \left\{ \mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) < \frac{\min_j \{d_j\}^2}{4} \right\}. \tag{3.8}$$

A similar assumption was made for the SFP method (see condition (2.13)). Note that for the true source location $\mathbf{x}_{\rm true}$ one has $f(\mathbf{x}_{\rm true}) = \sum_{j=1}^m \varepsilon_j^2$. Therefore, $\mathbf{x}_{\rm true}$ satisfies Assumption 2 if the errors ε_j are smaller in some sense than the range measurements d_j. This is a very reasonable assumption since in real applications the errors ε_i are often an order of magnitude smaller than d_i. Now, if the initial point $\mathbf{x}^0$ is good enough in the sense that it is close to the true source location, then Assumption 2 will be satisfied. We have observed through numerical experiments that the solution to the LS problem (1.3) often satisfies Assumption 2 as the following example demonstrates.

Example 3.1. Consider the source localization problem with $m = 5$ and $n = 2$. We performed Monte Carlo runs, where in each run the sensor locations $\mathbf{a}_j$ and the source location $\mathbf{x}$ were randomly generated from a uniform distribution over the square $[-20, 20] \times [-20, 20]$. The observed distances d_j are given by (1.1) with ε_j being independently generated from a normal distribution with mean zero and standard deviation σ. In our experiments σ takes on four different values: $1, 10^{-1}, 10^{-2}$, and 10^{-3}. For each σ, N_σ denotes the number of runs for which the condition $f(\mathbf{x}_{\rm LS}) < \frac{\min_j d_j^2}{4}$ holds, and the results are given in Table 1. Clearly, Assumption 2 fails only for high noise levels.

The following simple and important property will be used in our analysis.

LEMMA 3.1. *Let* $\mathbf{x} \in \mathcal{R}$. *Then*

$$\|\mathbf{x} - \mathbf{a}_j\| > d_j/2, \quad j = 1, \ldots, m. \tag{3.9}$$

Proof. Suppose in contradiction that there exists j_0 for which $\|\mathbf{x}-\mathbf{a}_{j_0}\| \le d_{j_0}/2$. Then

$$f(\mathbf{x}) = \sum_{j=1}^{m} (\|\mathbf{x}-\mathbf{a}_j\| - d_j)^2 \ge (\|\mathbf{x}-\mathbf{a}_{j_0}\| - d_{j_0})^2 \ge \frac{d_{j_0}^2}{4} \ge \frac{\min\{d_j\}^2}{4},$$

which contradicts $\mathbf{x} \in \mathcal{R}$. ∎

A direct consequence of Lemma 3.1 is that any element in $\mathcal{R}$ cannot be one of the sensors.

COROLLARY 3.1. *If* $\mathbf{x} \in \mathcal{R}$, *then* $\mathbf{x} \notin \mathcal{A}$.

3.3.2. Convergence analysis of the SWLS method. We begin with the following result which plays a key role in the forthcoming analysis.

LEMMA 3.2. *Let* δ *be a positive number, and let* $t > \delta/2$. *Then*

$$\left(\frac{s^2}{t} - \delta\right)^2 \ge 2(s-\delta)^2 - (t-\delta)^2 \tag{3.10}$$

for every $s > \sqrt{\frac{\delta t}{2}}$, *and equality is satisfied if and only if* $s = t$.

Proof. Rearranging (3.10) one has to prove

$$A(s,t) \equiv \left(\frac{s^2}{t} - \delta\right)^2 - 2(s-\delta)^2 + (t-\delta)^2 \ge 0.$$

Some algebra shows that the expression $A(s,t)$ can be written as follows:

$$A(s,t) = \frac{1}{t}(s-t)^2 \left(\left(\frac{s}{\sqrt{t}} + \sqrt{t}\right)^2 - 2\delta\right). \tag{3.11}$$

Using the conditions $t > \delta/2$ and $s > \sqrt{\frac{\delta t}{2}}$, we obtain

$$\left(\frac{s}{\sqrt{t}} + \sqrt{t}\right)^2 - 2\delta > \left(\sqrt{\frac{\delta}{2}} + \sqrt{\frac{\delta}{2}}\right)^2 - 2\delta = 0. \tag{3.12}$$

Therefore, from (3.11) and (3.12) it readily follows that $A(s,t) \ge 0$ and that equality holds if and only if $s = t$. ∎

Thanks to Lemma 3.2, we establish the next result which is essential in proving the monotonicity of the SWLS method.

LEMMA 3.3. *Let* $\mathbf{y} \in \mathcal{R}$. *Then the function* $g(\mathbf{x},\mathbf{y})$ *given in* (3.2) *is well defined on* $\mathbb{R}^n \times \mathcal{R}$, *and with*

$$\mathbf{z} \in \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} g(\mathbf{x},\mathbf{y}), \tag{3.13}$$

the following properties hold:

(a) $f(\mathbf{z}) \le f(\mathbf{y})$, *and the equality is satisfied if and only if* $\mathbf{z} = \mathbf{y}$;

(b) $\mathbf{z} \in \mathcal{R}$.

Proof. By Corollary 3.1, any $\mathbf{y} \in \mathcal{R}$ implies $\mathbf{y} \notin \mathcal{A}$, and hence the function g given by (cf. (3.2))

$$g(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{m} \left(\frac{\|\mathbf{x}-\mathbf{a}_i\|^2}{\|\mathbf{y}-\mathbf{a}_i\|} - d_i\right)^2$$

is well defined on $\mathbb{R}^n \times \mathcal{R}$. Now, by (3.13) and $\mathbf{y} \in \mathcal{R}$ we have

$$g(\mathbf{z},\mathbf{y}) \leq g(\mathbf{y},\mathbf{y}) = f(\mathbf{y}) < \frac{\min\{d_j\}^2}{4}. \tag{3.14}$$

In particular,

$$\left(\frac{\|\mathbf{z}-\mathbf{a}_j\|^2}{\|\mathbf{y}-\mathbf{a}_j\|} - d_j\right)^2 < \frac{d_j^2}{4}, \quad j = 1,\ldots,m,$$

from which it follows that

$$\frac{\|\mathbf{z}-\mathbf{a}_j\|^2}{\|\mathbf{y}-\mathbf{a}_j\|} \geq \frac{d_j}{2}, \quad j = 1,\ldots,m. \tag{3.15}$$

Invoking Lemma 3.2, whose conditions are satisfied by (3.15) and Lemma 3.1, we obtain

$$\left(\frac{\|\mathbf{z}-\mathbf{a}_j\|^2}{\|\mathbf{y}-\mathbf{a}_j\|} - d_j\right)^2 \geq 2(\|\mathbf{z}-\mathbf{a}_j\| - d_j)^2 - (\|\mathbf{y}-\mathbf{a}_j\| - d_j)^2.$$

Summing over $j = 1,\ldots,m$, we obtain

$$\sum_{j=1}^m \left(\frac{\|\mathbf{z}-\mathbf{a}_j\|^2}{\|\mathbf{y}-\mathbf{a}_j\|} - d_j\right)^2 \geq 2\sum_{j=1}^m (\|\mathbf{z}-\mathbf{a}_j\| - d_j)^2 - \sum_{j=1}^m (\|\mathbf{y}-\mathbf{a}_j\| - d_j)^2.$$

Therefore, together with (3.14), we get

$$f(\mathbf{y}) \geq g(\mathbf{z},\mathbf{y}) \geq 2f(\mathbf{z}) - f(\mathbf{y}),$$

showing that $f(\mathbf{z}) \leq f(\mathbf{y})$. Now, assume that $f(\mathbf{y}) = f(\mathbf{z})$. Then by Lemma 3.2 it follows that the following set of equalities is satisfied:

$$\|\mathbf{y}-\mathbf{a}_j\| = \|\mathbf{z}-\mathbf{a}_j\|, \quad j = 1,\ldots,m, \tag{3.16}$$

which after squaring and rearranging reads as

$$(\|\mathbf{y}\|^2 - \|\mathbf{z}\|^2) - 2\mathbf{a}_j^T(\mathbf{y}-\mathbf{z}) = 0, \quad j = 1,\ldots,m.$$

Therefore,

$$\begin{pmatrix} 1 & \mathbf{a}_1^T \\ 1 & \mathbf{a}_2^T \\ \vdots & \vdots \\ 1 & \mathbf{a}_m^T \end{pmatrix} \begin{pmatrix} \|\mathbf{y}\|^2 - \|\mathbf{z}\|^2 \\ -2(\mathbf{y}-\mathbf{z}) \end{pmatrix} = 0.$$

Thus, by Assumption 1, $\mathbf{z} = \mathbf{y}$, and the proof of (a) is completed. To prove (b), using (a) and (3.14), we get

$$f(\mathbf{z}) \leq f(\mathbf{y}) < \min_{j=1,\ldots,m} \frac{d_j^2}{4},$$

proving that $\mathbf{z} \in \mathcal{R}$. ∎

We are now ready to prove the main convergence results for the SWLS method.

THEOREM 3.1 (convergence of the SWLS method). *Let $\{\mathbf{x}^k\}$ be the sequence generated by the SWLS method. Suppose that Assumptions* 1 *and* 2 *hold true. Then*

(a) $\mathbf{x}^k \in \mathcal{R}$ *for* $k \geq 0$*;*
(b) *for every* $k \geq 0$, $f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k)$ *and equality holds if and only if* $\mathbf{x}^{k+1} = \mathbf{x}^k$*;*
(c) *the sequence of function values* $\{f(\mathbf{x}^k)\}$ *converges;*
(d) *the sequence* $\{\mathbf{x}^k\}$ *is bounded;*
(e) *every convergent subsequence* $\{\mathbf{x}^{k_l}\}$ *satisfies* $\mathbf{x}^{k_l+1} - \mathbf{x}^{k_l} \to \mathbf{0}$*;*
(f) *any limit point of* $\{\mathbf{x}^k\}$ *is a stationary point of* f*.*

Proof. (a) and (b) The proof follows by induction on k using Lemma 3.3.

(c) The proof follows from the fact that $\{f(\mathbf{x}^k)\}$ is bounded below (by zero) and is a nonincreasing sequence.

(d) By (b), all of the iterates $\mathbf{x}^k$ are in the level set $\mathrm{Lev}(f, f(\mathbf{x}^0))$ which, by Lemma 2.3, establishes the boundedness of the sequence $\{\mathbf{x}^k\}$.

(e) Let $\{\mathbf{x}^{k_l}\}$ be a convergent subsequence, and denote its limit by $\mathbf{x}^*$. By claims (a) and (b), we have for every k that

$$f\left(\mathbf{x}^k\right) \leq f\left(\mathbf{x}^0\right) < \min_{j=1,\ldots,m} \frac{d_j^2}{4},$$

which combined with the continuity of f implies $\mathbf{x}^* \in \mathcal{R}$ and hence $\mathbf{x}^* \notin \mathcal{A}$, by Corollary 3.1. Now, recall that

$$\mathbf{x}^{k_l+1} \in \operatorname{argmin} g\left(\mathbf{x}, \mathbf{x}^{k_l}\right).$$

To prove the convergence of $\{\mathbf{x}^{k_l+1}\}$ to $\mathbf{x}^*$, we will show that every subsequence converges to $\mathbf{x}^*$. Let $\{\mathbf{x}^{k_{l_p}+1}\}$ be a convergent subsequence, and denote its limit by $\mathbf{y}^*$. Since

$$\mathbf{x}^{k_{l_p}+1} \in \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} g\left(\mathbf{x}, \mathbf{x}^{k_{l_p}}\right),$$

the following holds:

$$g\left(\mathbf{x}, \mathbf{x}^{k_{l_p}}\right) \geq g\left(\mathbf{x}^{k_{l_p}+1}, \mathbf{x}^{k_{l_p}}\right) \text{ for every } \mathbf{x} \in \mathbb{R}^n.$$

Taking the limits of both sides in the last inequality and using the continuity of the function f, we have

$$g(\mathbf{x}, \mathbf{x}^*) \geq g(\mathbf{y}^*, \mathbf{x}^*) \text{ for every } \mathbf{x} \in \mathbb{R}^n,$$

and hence

$$\mathbf{y}^* \in \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} g(\mathbf{x}, \mathbf{x}^*). \tag{3.17}$$

Since the sequence of function values converges, it follows that $f(\mathbf{x}^*) = f(\mathbf{y}^*)$. Invoking Lemma 3.3 with $\mathbf{y} = \mathbf{x}^*$ and $\mathbf{z} = \mathbf{y}^*$, we obtain $\mathbf{x}^* = \mathbf{y}^*$, establishing claim (e).

(f) To prove the claim, note that (3.17) and $\mathbf{x}^* = \mathbf{y}^*$ imply that

$$\mathbf{x}^* \in \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} g(\mathbf{x}, \mathbf{x}^*).$$

Thus, by the first order optimality conditions we obtain the following:

$$0 = \nabla_{\mathbf{x}} g(\mathbf{x}, \mathbf{x}^*)|_{\mathbf{x}=\mathbf{x}^*} = 4 \sum_{j=1}^{m} \left(\|\mathbf{x}^* - \mathbf{a}_j\| - d_j\right) \frac{\mathbf{x}^* - \mathbf{a}_j}{\|\mathbf{x}^* - \mathbf{a}_j\|} = 2\nabla f(\mathbf{x}^*). \qquad \square$$

As a direct consequence of Theorem 3.1, we obtain the following convergence in function values.

COROLLARY 3.2. *Let $\{\mathbf{x}^k\}$ be the sequence generated by the algorithm. Then $f(\mathbf{x}^k) \to f^*$, where f^* is the function value at some stationary point $\mathbf{x}^*$ of f.*

As was shown for the SFP algorithm, global convergence of the sequence generated by the SWLS algorithm can also be established under the same condition, i.e., assuming that f admits isolated stationary points.

THEOREM 3.2 (convergence of the sequence). *Let $\{\mathbf{x}^k\}$ be generated by* (3.3) *such that Assumptions* 1 *and* 2 *hold. Suppose further that all stationary points of f are isolated. Then the sequence $\{\mathbf{x}^k\}$ converges to a stationary point.*

Proof. The proof is the same as the proof of Theorem 2.2. □

4. Numerical examples. In this section we present numerical simulations illustrating the performance of the SFP and SWLS schemes, as well as numerical comparisons with the LS approach and with the SDR of the ML problem. The simulations were performed in MATLAB, and the semidefinite programs were solved by SeDuMi [14].

Before describing the numerical results, for the reader's convenience, we first recall the SDR proposed in [4], which will be used in our numerical experiments comparisons. The first stage is to rewrite problem (ML) given in (1.2) as

$$\begin{array}{ll} \min_{\mathbf{x},\mathbf{g}} & \sum_{j=1}^{m}(g_j - d_j)^2 \\ \text{s.t.} & g_j^2 = \|\mathbf{x} - \mathbf{a}_j\|^2, \quad j = 1, \ldots, m. \end{array}$$

Making the change of variables

$$\mathbf{G} = \begin{pmatrix} \mathbf{g} \\ 1 \end{pmatrix} \begin{pmatrix} \mathbf{g}^T & 1 \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix} \begin{pmatrix} \mathbf{x}^T & 1 \end{pmatrix},$$

problem (1.2) becomes

$$\begin{array}{ll} \min_{\mathbf{X},\mathbf{G}} & \sum_{j=1}^{m}(G_{jj} - 2d_j G_{m+1,j} + d_j^2) \\ \text{s.t.} & G_{jj} = \mathrm{Tr}(\mathbf{C}_j \mathbf{X}), \quad j = 1, \ldots, m, \\ & \mathbf{G} \succeq \mathbf{0}, \ \ \mathbf{X} \succeq \mathbf{0}, \\ & G_{m+1,m+1} = X_{n+1,n+1} = 1, \\ & \mathrm{rank}(\mathbf{X}) = \mathrm{rank}(\mathbf{G}) = 1, \end{array}$$

where

$$\mathbf{C}_j = \begin{pmatrix} \mathbf{I} & -\mathbf{a}_j \\ -\mathbf{a}_j^T & \|\mathbf{a}_j\|^2 \end{pmatrix}, \quad j = 1, \ldots, m.$$

Dropping the rank constraints in the above problem, we obtain the desired SDR of problem (1.2). The SDR is not guaranteed to provide an accurate solution to the ML problem, but it can always be considered as an approximation of the ML problem.

In the first example, we show that the SWLS scheme usually possesses a larger region of convergence to the global minimum than the scheme SFP. This last property is further demonstrated in the second example, which compares the SFP and SWLS methods and also demonstrates the superiority of the SWLS scheme. The last example

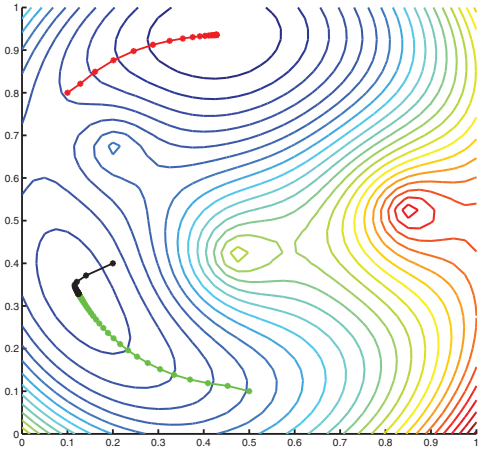


FIG. 1. *The SFP method for three initial points.*

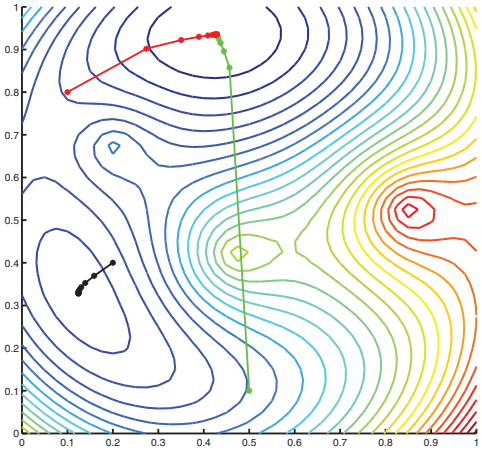


FIG. 2. *The SWLS method for three initial points.*

illustrates the attractiveness of the solution obtained by the SWLS method over the SDR and the LS approaches.

Example 4.1 (region of convergence of the SFP and SWLS methods). In this example we show typical behaviors of the SFP and SWLS methods. Consider an instance of the source localization problem in the plane ($n = 2$) with three sensors ($m = 3$) in the locations (0.466,0.418), (0.846,0.525) and (0.202,0.672). Figures 1 and 2 describe the results produced by the iterative schemes SFP and SWLS, respectively, for three initial trial points. The global minimum is $(0.4285, 0.9355)$, and there exists one additional local minimum at $(0.1244, 0.3284)$. As demonstrated in Figure 2, the SWLS method might converge to a local minimum; however, it seems to have a greater chance than the SFP algorithm to avoid local minima; for example, the SWLS converged to the (relatively far) global minimum from the initial starting point (0.5,0.1), while the SFP converged to the local minimum. We estimated the probability to converge to the global minimum by invoking both methods for 1681 initial starting

TABLE 2
Comparison between the SFP and SWLS methods.

m	#tight	$\#(f(\hat{\mathbf{x}}_{\mathrm{SFP}}) > f(\hat{\mathbf{x}}_{\mathrm{SWLS}}))$	Iter – SFP	Iter – SWLS
3	314	152	207(500.2)	26.2 (5)
4	325	96	124(192.6)	29.9(1.8)
5	259	83	93.6(96.2)	30.9(3.1)
10	278	23	66.5 (35.3)	31.6 (1.3)

points, which are the nodes of a 41×41 grid over the square $[0,1] \times [0,1]$. The SFP method converged to the global minimum in 45.87% of the runs, while the SWLS methods converged to the global minimum in 83.28% of the runs. Thus, the SWLS method has a much wider region of convergence to the global minimum. This was our observation in many other examples that we ran, which suggests that the SWLS method has the tendency to converge to the global minimum.

Remark 4.1. As shown in Proposition 2.1, the SFP scheme is just a gradient method with a fixed step size. Thanks to Lemma 2.5, which as shown in section 2.2 can be used in order to avoid the nonsmoothness, we can of course use more sophisticated smooth unconstrained minimization methods. Indeed, we also tested a gradient method with an Armijo step-size rule and a trust region method [8], which uses second order information. Our observation was that, while these methods usually possess an improved rate of convergence in comparison to the SFP method, they essentially have the same region of convergence to the global minimum as the SFP algorithm.

Example 4.2 (comparison of the SFP and SWLS methods). We performed Monte Carlo runs, where in each run the sensor locations $\mathbf{a}_j$ and the true source location were randomly generated from a uniform distribution over the square $[-1000, 1000] \times [-1000, 1000]$. The observed distances d_j are given by (1.1) with ε_j being generated from a normal distribution with mean zero and standard deviation 20. Both the SFP and SWLS methods were employed with (the same) initial point, which was also uniformly randomly generated from the square $[-1000, 1000] \times [-1000, 1000]$. The stopping rule for both the SWLS and SFP methods was $\|\nabla f(\mathbf{x}^k)\| < 10^{-5}$.

The results of the runs are summarized in Table 2. For each value of m, 1000 realizations were generated. The numbers in the first column are the number of sensors, and in the second column we give the number of runs out of 1000 in which the SDR of the ML problem was tight; that is, the matrix which is the optimal solution of the SDR has rank one. We have also compared the SWLS solution with the SDR solution for these "tight" runs (about a quarter of the runs). In all of these runs, the SWLS and SDR solutions coincided; i.e., the SWLS method produced the exact ML solution. The third column contains the number of runs out of 1000 in which the solution produced by the SFP method was worse than the SWLS method. In all of the remaining runs, the two methods converge to the same point; thus, there were no runs in which the SWLS produced worse results. The last two columns contain the mean and standard deviation of the number of iterations of each of the methods in the form "mean (standard deviation)."

As can be clearly seen from the table, the SWLS method requires much less iterations than the SFP method, and in addition it is more robust in the sense that the number of iterations are more or less constant. In contrast, the standard deviations of the number of iterations of the SFP method are quite large. For example, the huge standard deviation 500.2 in the first row stems from the fact that in some of the runs the SFP algorithm required thousands of iterations!

TABLE 3
Mean squared position error of the SDR, LS and SWLS methods.

σ	SDR	LS	SWLS
$1e-3$	$2.4e-6$	$2.7e-6$	$\mathbf{1.5e-6}$
$1e-2$	$2.2e-4$	$1.6e-4$	$\mathbf{1.3e-4}$
$1e-1$	$2.2e-2$	$1.9e-2$	$\mathbf{1.3e-2}$
$1e+0$	$2.2e+0$	$2.7e+0$	$\mathbf{2.0e+0}$

From the above examples we conclude that the SWLS method does tend to converge to the global minimum. Of course, we can always construct an example in which the method converges to a local minimum (as was demonstrated in Example 4.1), but it seems that for random instances this convergence to a nonglobal solution is not likely.

We should also note that we also compared the SFP and SWLS methods with the initial point chosen as the solution of the LS problem (1.3). For this choice of the initial point, the SFP and SWLS methods always converged to the same location point[4] (which is probably the global minimum); however, with respect to the number of iterations, the SWLS method was still significantly superior to the SFP algorithm. We have also compared the SWLS solution with the SDR solution for the runs in which the SDR solution is tight (about a quarter of the runs (cf. column 1 in Table 2)). In all of these runs, the SWLS and SDR solutions coincided; i.e., the SWLS method produced the exact ML solution.

The last example shows the attractiveness of the SWLS method over the LS and SDR approaches.

Example 4.3 (comparison with the LS and SDR estimates). Here we compare the solution of (1.3) and the solution of the SDR with the SWLS solution. The stopping rule for the SWLS method was $\|\nabla f(\mathbf{x}_k)\| < 10^{-5}$. We generated 100 random instances of the source localization problem with five sensors, where in each run the sensor locations $\mathbf{a}_j$ and the source location $\mathbf{x}$ were randomly generated from a uniform distribution over the square $[-10, 10] \times [-10, 10]$. The observed distances d_j are given by (1.1) with ε_j being independently generated from a normal distribution with mean zero and standard deviation σ. In our experiments σ takes four different values: $1, 10^{-1}, 10^{-2}$, and 10^{-3}. The numbers in the three right columns of Table 3 are the average of the squared position error $\|\hat{\mathbf{x}} - \mathbf{x}\|^2$ over 100 realizations, where $\hat{\mathbf{x}}$ is the solution by the corresponding method. The best result for each possible value of σ is marked in boldface. From the table, it is clear that the SWLS algorithm outperforms the LS and SDR methods for all four values of σ.

REFERENCES

[1] A. BECK, P. STOICA, AND J. LI, *Exact and approximate solutions of source localization problems*, IEEE Trans. Signal Process., 56 (2008), pp. 1770–1778.

[2] D. P. BERTSEKAS, *Nonlinear Programming*, 2nd ed., Athena Scientific, Belmont, MA, 1999.

[3] P. BISWAS, T. C. LIAN, T. C. WANG, AND Y. YE, *Semidefinite programming based algorithms for sensor network localization*, ACM Trans. Sen. Netw., 2 (2006), pp. 188–220.

[4] K. W. CHEUNG, W. K. MA, AND H. C. SO, *Accurate approximation algorithm for TOA-based maximum likelihood mobile location using semidefinite programming*, in Proceedings of the ICASSP, Vol. 2, 2004, pp. 145–148.

[4] Numerically we used the criteria that two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ are "the same" if $\|\mathbf{x}_1 - \mathbf{x}_2\| \leq 10^{-8}$.

[5] K. W. Cheung, H. C. So, W. K. Ma, and Y. T. Chan, *Least squares algorithms for time-of-arrival-based mobile location*, IEEE Trans. Signal Process., 52 (2004), pp. 1121–1228.
[6] C. Fortin and H. Wolkowicz, *The trust region subproblem and semidefinite programming*, Optim. Methods Softw., 19 (2004), pp. 41–67.
[7] H. W. Kuhn, *A note on Fermat's problem*, Math. Program., 4 (1973), pp. 98–107.
[8] J. J. Moré and D. C. Sorensen, *Computing a trust region step*, SIAM J. Sci. Stat. Comput., 4 (1983), pp. 553–572.
[9] J. J. Moré, *Generalizations of the trust region subproblem*, Optim. Methods Softw., 2 (1993), pp. 189–209.
[10] L. M. Ostresh, *On the convergence of a class of iterative methods for solving the Weber location problem*, Oper. Res., 26 (1978), pp. 597–609.
[11] J. G. Morris, R. F. Love, and G. O. Wesolowsky, *Facilities Location: Models and Methods*, North–Holland, New York, 1988.
[12] A. H. Sayed, A. Tarighat, and N. Khajehnouri, *Network-based wireless location*, IEEE Signal Proces. Mag., 22 (2005), pp. 24–40.
[13] P. Stoica and J. Li, *Source localization from range-difference measurements*, IEEE Signal Process. Mag., 23 (2006), pp. 63–65, 69.
[14] J. F. Sturm, *Using SeDuMi* 1.02*, a Matlab toolbox for optimization over symmetric cones*, Optim. Methods Softw., 11–12 (1999), pp. 625–653.
[15] Y. Vardi and C. H. Zhang, *A modified Weiszfield algorithm for the Fermat-Weber location problem*, Math. Program. Ser. A, 90 (2001), pp. 559–566.
[16] E. Weiszfeld, *Sur le point pour lequel la somme des distances de n points donnés est minimum*, Tohoku Math. J., 43 (1937), pp. 355–386.
[17] http://www.fcc.gov/911/enhanced/.

SIAM J. OPTIM.
Vol. 19, No. 3, pp. 1417–1432

A SIMPLIFIED APPROACH TO SEMISMOOTH NEWTON METHODS IN FUNCTION SPACE*

ANTON SCHIELA†

Abstract. We present an alternative approach to the analysis of Newton's method for function space problems involving semismooth Nemyckii operators. The simple main idea is to apply a local continuity result to appropriately chosen finite differences. In this respect it runs in parallel to the theory of Fréchet differentiable Nemyckii operators. This leads to a concise proof of superlinear convergence under relaxed conditions, compared to previous results. Moreover, extensions of this technique allow one to prove sharpened bounds on the rate of convergence and study semismooth Newton methods in the presence of compactness.

Key words. continuity of Nemyckii operators, Newton methods in function space, optimal control

AMS subject classifications. 49M15, 49K20, 46N40

DOI. 10.1137/060674375

1. Introduction. Newton's method is a standard algorithm for solving nonlinear systems of equations and optimization problems, both in finite- and in infinite dimensional normed spaces. For a nonlinear system $G(x) = 0$ classical assumptions in the analysis of Newton's method are Lipschitz continuous differentiability of G in a neighborhood of a solution x_* and invertibility of $G'(x)$. Many nonlinear problems in function space are formulated with the help of a pointwise nonlinear function. The corresponding operators in function space are called *Nemyckii-* or superposition operators. Due to their practical importance Nemyckii operators have been analyzed thoroughly, and many standard results have been established (for a thorough exposition, see [2]), such as continuity and differentiability.

For a large class of problems the requirement of continuous differentiability is too strong, because the pointwise nonlinear functions that appear there are not differentiable in the classical sense, but are only semismooth. In the last few years semismooth Newton methods in function space have been studied (cf., e.g., [8, 14, 12, 13, 6]) with great success, particularly in the field of PDE-constrained optimal control, where control constraints can be modeled by semismooth functions. In [12, 13] Newton methods for problems involving semismooth Nemyckii operators were analyzed, and superlinear convergence was studied in the presence of some smoothing operator that maps L_q to L_p for $q < p$. The corresponding proofs (cf. also [6]) rely on sophisticated splittings of the domain of definition into several subdomains and different estimates there.

Here we consider an alternative approach to the semismoothness of Nemyckii operators, which runs closely in parallel to the corresponding theory for Fréchet differentiability. This approach yields several benefits. First, we can replace global Lipschitz continuity conditions as used in [12] by weaker growth conditions and thus extend the theoretical framework for semismooth Newton methods. Second, the semismoothness of Nemyckii operators can be derived as a simple consequence of a *local* continuity result and Hölder's inequality. Further, we derive sharpened bounds on convergence

*Received by the editors November 8, 2006; accepted for publication (in revised form) August 1, 2008; published electronically November 21, 2008.
http://www.siam.org/journals/siopt/19-3/67437.html

†Zuse Institute Berlin, Takustr. 7, 14195 Berlin, Germany (schiela@zib.de). This author's research was supported by the DFG Research Center MATHEON Mathematics for key technologies.

rates and analyze semismooth Newton methods in the presence of compactness. Finally, we illustrate our results with a simple example from optimal control.

Although this work contains a couple of new *results* about semismooth Newton methods, this is not its main emphasis. We rather want to point out new and simpler *techniques* for its analysis and its convergence theory. We hope that this makes the theory of semismooth Newton methods more accessible, and that new ideas may emerge in this framework.

2. Theoretical framework and outline. Let X be a normed space, Y a linear space, and $G : X \to Y$ a nonlinear mapping. Consider the equation

$$G(x) = 0, \tag{1}$$

and assume that there is an $x_* \in X$ that solves this equation; i.e., $G(x_*) = 0$.

A minimal approach to Newton's method. Consider Newton's method for the solution of (1), taking a general point of view. For this purpose let $U(x_*)$ be a neighborhood of x_*, and for all $x \in U(x_*)$ let $G'(x)(\cdot) : X \to Y$ be invertible linear operators. They will play *the role* of derivatives of $G(x)$. Let $x_k \in U(x_*)$, and define a Newton step via the update formula

$$x_{k+1} := x_k - G'(x_k)^{-1} G(x_k). \tag{2}$$

The following theorem covers local superlinear convergence of the corresponding Newton iteration. Because existence of x_* is assumed, which is characteristic for semismooth Newton theorems, completeness of X is not required.

THEOREM 2.1. *For the nonlinear equation* (1) *define*

$$\Theta(x) := \frac{\left\| G'(x)^{-1} \big(G'(x)(x - x_*) - (G(x) - G(x_*)) \big) \right\|_X}{\|x - x_*\|_X}. \tag{3}$$

For $x_k \in U(x_*)$ the Newton step (2) *satisfies*

$$\|x_{k+1} - x_*\|_X = \Theta(x_k) \|x_k - x_*\|_X .$$

If the consistency condition $\lim_{x \to x_*} \Theta(x) = 0$ holds, then Newton's method converges locally superlinearly to x_*.

Proof. By our assumptions the Newton step $x_k \to x_{k+1}$ is well defined in $U(x_*)$. By its definition and by $G(x_*) = 0$ we have

$$\begin{aligned} \|x_{k+1} - x_*\|_X &= \left\| x_k - x_* - G'(x_k)^{-1} G(x_k) \right\|_X \\ &= \left\| G'(x_k)^{-1} \big(G'(x_k)(x_k - x_*) - (G(x_k) - G(x_*)) \big) \right\|_X \\ &= \Theta(x_k) \|x_k - x_*\|_X . \end{aligned}$$

If $\lim_{x \to x_*} \Theta(x) = 0$, then there is a ball around x_*, where $\Theta(x) < 0.5$, and

$$\|x_{k+1} - x_*\|_X < 0.5 \|x_k - x_*\|_X .$$

In this case the Newton sequence remains in this ball by induction, converges to x_*, and thus converges superlinearly by definition of the quantity $\Theta(x)$. ∎

Classically, $G'(x)$ is taken as the Fréchet derivative of G at x, which, however, need not exist in many practically relevant cases, such as semismoothness. We will view $G'(x)$ as an *algorithmic construct* rather than an analytic object. Any choice of $G'(x)$ can be used for which the *consistency condition* $\lim_{x \to x_*} \Theta(x) = 0$ holds. This choice is far from being unique.

Baire–Carathéodory functions and their Nemyckii operators. Consider a measurable set $\Omega \subset \mathbb{R}^d$ equipped with a positive measure μ such that $\mu(\Omega) < \infty$, two separable Banach spaces $\mathcal{X}, \mathcal{Y}$ (usually $\mathcal{X}, \mathcal{Y} = \mathbb{R}^n$), and a function

$$\psi : \mathcal{X} \times \Omega \to \mathcal{Y},$$

$$(x, t) \mapsto \psi(x, t).$$

For $1 \le p \le \infty$ consider the Bochner–Lebesgue space $L_p(\Omega, \mathcal{X})$ of p-integrable functions $v : \Omega \to \mathcal{X}$, and denote its norm by $\|\cdot\|_{L_p}$. As usual, $L_\infty(\Omega, \mathcal{X})$ is the space of essentially bounded functions, and for $p = \infty$ we set $p^{-1} = 0$. Let $D \subset L_p(\Omega, \mathcal{X})$. If it is well defined, then the operator

$$\Psi : D \to L_s(\Omega, \mathcal{Y}),$$

$$x \mapsto \Psi(x) : \Psi(x)(t) = \psi(x(t), t) \qquad \text{a.e.}$$

is called the *Nemyckii* operator from D to $L_s(\Omega, \mathcal{Y})$ corresponding to ψ. To be well defined Ψ must necessarily map measurable functions to measurable functions. Note that there are several concepts of measurability in Banach spaces that, however, coincide in separable spaces. As a sufficient condition ψ is usually assumed to be a *Carathéodory* function; i.e., ψ is continuous in x and measurable in t. However, since pointwise limits of measurable functions in separable spaces are measurable (cf., e.g., [9, 21.4]), this class can be extended substantially to functions ψ that are *pointwise limits* of sequences ψ_n of Carathéodory functions. Inserting a measurable function x into Ψ, the sequence of measurable functions $\Psi_n(x)$ converges pointwise to $\Psi(x)$, which is consequently measurable. This is the class of *Baire–Carathéodory* functions (cf. [2])—an extension that is essential for our theory.

So far Ψ is a mapping between spaces of measurable functions. To assert that $\Psi : L_p(\Omega, \mathcal{X}) \to L_s(\Omega, \mathcal{Y})$ we have to impose a *growth condition* on ψ, which reads (cf. [16, section 26.3]) as follows:

$$|\psi(x, t)|_{\mathcal{Y}} \le a(t) + b\,|x|_{\mathcal{X}}^{p/s} \qquad \text{for some } a \in L_s(\Omega, \mathbb{R}), b \in \mathbb{R}. \tag{4}$$

Hence, the behavior of ψ for large x restricts the choice of spaces on which a corresponding Nemyckii operator Ψ can be defined.

Semilinear operator equations. As a convenient framework, consider nonlinear operators $G(x)$ of the following form, which we will call *semilinear* operators:

$$G(x) := \mathrm{T}\,x + F(x). \tag{5}$$

Let us gather our notation introduced so far and fix our theoretical framework by stating the following set of basic assumptions.

Basic Assumptions 2.2. *Let $\Omega \subset \mathbb{R}^d$ be a measurable set, equipped with a positive measure μ with $\mu(\Omega) < \infty$. Let $\mathcal{X}$ and $\mathcal{Y}$ be separable Banach spaces.*

Let X and Y be linear spaces and assume that there are $1 \le p, q \le \infty$ such that $X \subset L_p(\Omega, \mathcal{X})$ and $L_q(\Omega, \mathcal{Y}) \subset Y$. Assume that $(X, \|\cdot\|_{L_p})$ is a normed space.

Let $T : X \to Y$ be a linear operator and $F : L_p(\Omega, \mathcal{X}) \to L_q(\Omega, \mathcal{Y})$ a Nemyckii operator, corresponding to a Carathéodory function $f(x, t)$.

Assume that there is $x_* \in X$ such that $G(x_*) = 0$. Let

$$f'(\cdot, \cdot)(\cdot) : (\mathcal{X} \times \Omega) \times \mathcal{X} \to \mathcal{Y},$$

$$(x, t, v) \mapsto f'(x, t)(v)$$

be a function that is a Baire–Carathéodory function in (x,t) *and linear in* v.

Assume that there is a neighborhood $U(x_*)$ *of* x_* *such that for every* $x \in U(x_*)$ *the corresponding* linear *Nemyckii operator* $F'(x)(\cdot)$ *maps* X *to* Y *and the linear operator* $G'(x) := T + F'(x)$,

$$G'(x)(\cdot) : X \to Y,$$
$$v \mapsto Tv + F'(x)v,$$

has an inverse $G'(x)^{-1} : Y \to X$.

Remark 2.3. In many applications T is a linear differential operator, which is defined on some Sobolev space $X := W^{\sigma,r}(\Omega) \hookrightarrow L_p(\Omega)$. Since we do not need the completeness of X, we are free to equip $W^{\sigma,r}(\Omega)$ with an L_p-norm. This captures the nonlinearity of F best. Y can be chosen as a dual space that contains $L_q(\Omega)$.

In such a setting $G(x) = 0$ may be a semilinear PDE, but our formulation also includes systems of PDEs and algebraic equations, which arise in control-constrained optimal control (cf. section 6 below).

Linearity of T is assumed for simplicity of presentation. Of course, our framework extends straightforwardly to nonlinear operators $T(x)$ that satisfy suitable smoothness conditions such as semismoothness.

Basic Assumptions 2.2 define our framework and assert that $G : X \to Y$, $G'(x) : X \to Y$, and that the Newton steps (2) are well defined on a neighborhood of x_*, but they are not sufficient to show (superlinear) convergence of the corresponding iteration. To this end we have to establish a relation between f and f'. We will do this conveniently in terms of the function

$$\psi_*(x,t;\alpha) := \begin{cases} \dfrac{f'(x,t)(x - x_*(t)) - (f(x,t) - f(x_*(t),t))}{|x - x_*(t)|_{\mathcal{X}}^{\alpha}} & : \quad x \neq x_*(t), \\ 0 & : \quad x = x_*(t), \end{cases} \tag{6}$$

and its limiting behavior at $x = x_*(t)$ for $\alpha \geq 1$. In sections 3 and 5, where we study superlinear convergence of Newton's method, we will assume continuity of $\psi_*(x,t;1)$ at $x_*(t)$ a.e. In section 4 we discuss local boundedness of $\psi_*(x,t;\alpha)$ for $\alpha > 1$ and rates of convergence. We may view these assumptions as *pointwise consistency conditions*.

As a general strategy we will carry over the properties of $\psi_*(x,t;\alpha)$ to its corresponding Nemyckii operator $\Psi_*(x;\alpha)$ in suitable spaces. Then, under a *smoothing* assumption on $G'(x)^{-1}$ we invoke Theorem 2.1 by estimating

$$\begin{aligned} \Theta(x) &:= \frac{\left\| G'(x)^{-1}\big(G'(x)(x - x_*) - (G(x) - G(x_*))\big)\right\|_{L_p}}{\|x - x_*\|_{L_p}} \\ &= \frac{\left\| G'(x)^{-1}\big(F'(x)(x - x_*) - (F(x) - F(x_*))\big)\right\|_{L_p}}{\|x - x_*\|_{L_p}} \end{aligned} \tag{7}$$

and showing consistency in $L_p(\Omega, \mathcal{X})$. Observe that T cancels out by linearity.

Relation to semismoothness. Theorem 2.1 is closely related to a semismooth Newton theorem. Its formulation purely in terms of the domain space makes it independent of the topology of the image space (an *affine covariant* formulation; cf. [4]). In view of (3) the use of operator norms for $G'(x)^{-1}$ recovers known results on semismooth Newton methods, such as [12, Theorem 3.12] or [6, Theorem 1.1].

Although semismooth Newton theorems are very simple, the characterization of semismoothness as an *intrinsic property* of G is involved. A straightforward definition would be via (3) and the requirement $\lim_{x\to x_*}\Theta(x)=0$. However, for fixed x_* there is always some $G'(x)$ such that $\Theta(x)=0$, which renders this definition of semismoothness meaningless. The choice of $G'(x)$ has to be restricted and localized to x. In finite dimensions Rademacher's theorem gives us an appropriate tool to define a set-valued generalized derivative $\partial G(x)$, from which $G'(x)$ has to be chosen. With this restriction a definition via (3) makes sense. However, this construction is not possible anymore for infinite dimensional spaces.

For this reason [12] first chooses a set-valued mapping $\partial G(x)$ from which $G'(x)$ has to be taken, and then defines $\partial G(x)$-semismoothness via this set. For nonlinear Nemyckii operators $F(x)$, [12] defines $\partial F(x)$ pointwise via $\partial f(x,t)$.

If f is semismooth and $f'(x,t)\in\partial f(x,t)$, then $\lim_{x\to x_*(t)}\psi_*(x,t;1)=0$ a.e. If f is γ-order semismooth, then $\psi_*(x,t;1+\gamma)$ is bounded a.e. in Ω. So semismoothness is a particular case in which our results can be applied.

3. Continuity of $\Psi_*(x;1)$ and superlinear convergence. Our first qualitative convergence result is proved in two simple steps. First, we prove a local version of a standard result on continuity of Nemyckii operators (cf., e.g., [16, Proposition 26.7(a)] or [5, Proposition IV.1.1]) for Baire–Carathéodory functions. Second, we apply this result to $\psi_*(x,t;1)$ and conclude superlinear convergence of Newton's method via the Hölder inequality and a smoothing assumption on $G'(x)^{-1}$.

LEMMA 3.1 (local continuity of Nemyckii operators). *Let $\mathcal{X}$, $\mathcal{Y}$ be separable Banach spaces, Ω a measurable subset of $\mathbb{R}^d$, and $\psi:\mathcal{X}\times\Omega\to\mathcal{Y}$ a Baire–Carathéodory function. For each measurable function $x:\Omega\to\mathcal{X}$ let $\Psi(x)$ be the measurable function $t\to\psi(x(t),t)$. Let $x_*\in L_p(\Omega,\mathcal{X})$ be given. Then the following assertion holds: If ψ is continuous with respect to x at $(x_*(t),t)$ for almost all $t\in\Omega$, and Ψ maps $L_p(\Omega,\mathcal{X})$ into $L_s(\Omega,\mathcal{Y})$ for $1\le p,s<\infty$, then Ψ is continuous at x_* in the norm topology.*

Proof. This is a slight modification of a well-known lemma of Krasnoselski, which states global continuity of Ψ for Carathéodory functions ψ. Our proof stays very close to the proof in [5, Proposition IV.1.1], and our weakened assumptions do not cause any additional difficulties compared to the standard case.

Because ψ is a Baire–Carathéodory function, Ψ maps measurable functions to measurable functions. To show its continuity at x_* for $s<\infty$ we consider an arbitrary sequence $\|x_n-x_*\|_{L_p}\to 0$. By picking a suitable subsequence, we may assume w.l.o.g. that $\|x_n-x_*\|_{L_p}^p\le 2^{-n}$ and $x_n(t)\to x_*(t)$ pointwise a.e. (cf., e.g., [9, Theorem 22.31]). Define the function

$$w(x,t):=|\psi(x,t)-\psi(x_*(t),t)|_{\mathcal{Y}}^s,$$

and denote by W the corresponding Nemyckii operator. Inserting the sequence x_n, we conclude that $W(x_n)\to 0$ pointwise a.e. because $\psi(x,t)$ is continuous in x at $(x_*(t),t)$ a.e.

Next we will show $W(x_n)\to 0$ in $L_1(\Omega,\mathbb{R})$ via the convergence theorem of Lebesgue. For this we construct a function $\overline{w}\in L_1(\Omega,\mathbb{R})$ that dominates the sequence $W(x_n)$. Since $W(x)\ge 0$ and $W(x_n)\to 0$ a.e., we can define the measurable sets

$$\Omega_n:=\{t\in\Omega: w(x_n(t),t)\ge w(x_m(t),t)\ \forall m\in\mathbb{N}\}\setminus\left(\bigcup_{1\le k<n}\Omega_k\right).$$

Hence, the function $\overline{x} := \sum_n \chi_{\Omega_n} x_n$ is measurable. Let $\overline{w} := W(\overline{x})$, which is measurable as well. By construction, $\overline{w}(t) = \sup_n w(x_n(t), t)$ a.e. Moreover,

$$\begin{aligned}\int_\Omega |\overline{x}(t) - x_*(t)|_{\mathcal{X}}^p \, dt &\le \int_\Omega \sup_n |x_n(t) - x_*(t)|_{\mathcal{X}}^p \, dt \\ &\le \sum_{n=1}^\infty \|x_n(t) - x_*(t)\|_{L_p}^p \le \sum_{n=1}^\infty 2^{-n} = 1.\end{aligned} \tag{8}$$

Consequently, since $x_* \in L_p(\Omega, \mathcal{X})$, $\overline{x} \in L_p(\Omega, \mathcal{X})$ by the triangle inequality and hence, because $\Psi : L_p(\Omega, \mathcal{X}) \to L_s(\Omega, \mathcal{Y})$, $\overline{w} = W(\overline{x}) \in L_1(\Omega, \mathbb{R})$. Since $\overline{w}$ dominates $W(x_n)$ we can apply the convergence theorem of Lebesgue to obtain $W(x_n) \to 0$ in $L_1(\Omega, \mathbb{R})$. Hence, $\Psi(x_n) \to \Psi(x_*)$ in $L_s(\Omega, \mathcal{Y})$. Because x_n was arbitrary, we conclude continuity of the operator $\Psi : L_p(\Omega, \mathcal{X}) \to L_s(\Omega, \mathcal{Y})$ at x_*. □

Remark 3.2. Let us add some remarks, concerning the indices s and p:

(i) In general one has to verify the growth condition (4) for ψ to assert that Ψ maps $L_p(\Omega, \mathcal{X})$ to $L_s(\Omega, \mathcal{Y})$.

(ii) If $p < \infty$ and $\|\Psi(x)\|_{L_\infty} \le M$ uniformly for all $x \in L_p(\Omega, \mathcal{X})$, then Lemma 3.1 holds for all $s < \infty$ but *not* for $s = \infty$ (except for the case of constant Ψ). This will turn out to be the main reason for the so-called *norm gap* that is observed in the analysis of semismooth Newton methods. The proof of Lemma 3.1 simplifies in this case because we can use the domination function $\overline{w} = (2M)^s$. In section 5 we consider a variant of Lemma 3.1 that shows a weak form of continuity for $s = \infty$.

(iii) The case $p = \infty$ is not covered by Lemma 3.1. However, for $s < \infty$ the proof carries over, replacing (8) by a suitable expression with essential suprema. Then, instead of a growth condition it is sufficient that $\|\Psi(x)\|_{L_\infty} \le M$ uniformly in a neighborhood of $x_* \in L_\infty(\Omega, \mathcal{X})$.
The case $p = s = \infty$ is different in character. Then Ψ is continuous at x_* if continuity of $\psi(x, t)$ at $x_*(t)$ is *uniform* in Ω. This is usually a too strong assumption in the context of semismoothness (cf. Example 3.4 below).

THEOREM 3.3. *Let $G : X \to Y$ be a semilinear operator as defined in* (5)*, and suppose that Basic Assumptions* 2.2 *hold. Assume additionally the following:*

(i) *$\psi_*(x, t; 1)$ as defined in* (6) *is a Baire–Carathéodory function and*

$$\lim_{x \to x_*(t)} \psi_*(x, t; 1) = 0 \quad \text{for almost all } t \in \Omega.$$

The corresponding Nemyckii operator $\Psi_*(x; 1)$ maps $L_p(\Omega, \mathcal{X})$ into $L_s(\Omega, \mathcal{Y})$ with $s < \infty$ and $s^{-1} + p^{-1} = q^{-1}$ (hence $p > q$).

(ii) *$\left\|G'(x)^{-1}\right\|_{L_q(\Omega, \mathcal{Y}) \to L_p(\Omega, \mathcal{X})} \le M$ holds uniformly on a neighborhood of x_*.*

Then Newton's method converges locally superlinearly to x_*.

Proof. By definition of $\psi_*(\cdot, \cdot; 1)$ we have for $x, x_* \in X$

$$F'(x)(x - x_*) - (F(x) - F(x_*)) = \Psi_*(x; 1)|x - x_*|_{\mathcal{X}}, \tag{9}$$

by interpretation of $|\cdot|_{\mathcal{X}} : L_p(\Omega, \mathcal{X}) \to L_p(\Omega, \mathbb{R})$ as a Nemyckii operator. By the Hölder inequality

$$\|\Psi_*(x; 1)|x - x_*|_{\mathcal{X}}\|_{L_q} \le \|\Psi_*(x; 1)\|_{L_s} \|x - x_*\|_{L_p} \quad \text{for } q^{-1} = s^{-1} + p^{-1}.$$

By (i) $\psi_*(x,t;1)$ is continuous at $x_*(t)$ a.e., and we can apply Lemma 3.1 to get $\lim_{x\to x_*}\|\Psi_*(x;1)\|_{L_s}=0$. Now we estimate $\Theta(x)$ as defined in (3), using (ii):

$$\Theta(x)\le\frac{\left\|G'(x)^{-1}\right\|_{L_q\to L_p}\|F'(x)(x-x_*)-(F(x)-F(x_*))\|_{L_q}}{\|x-x_*\|_{L_p}}\le M\|\Psi_*(x;1)\|_{L_s}.$$

Hence, $\lim_{x\to x_*}\Theta(x)=0$, which implies superlinear convergence by Theorem 2.1. ∎

Connection to known results. For semismooth functions [12] studies semismoothness of the corresponding Nemyckii operators under an additional *global Lipschitz condition*, which corresponds to global boundedness of ψ_*. In [6] semismoothness of the max-function is studied. In both works Ω is split into three subdomains whose sizes are carefully balanced. Then the remainder terms are estimated separately.

In our approach via Lemma 3.1 we can relax the global Lipschitz condition and replace it by weaker *growth conditions* (4) on $\psi_*(x,t;1)$. By Proposition A.1 these follow from growth conditions on f' and on *local* Lipschitz constants of f.

From a structural point of view, our approach via Lemma 3.1 parallels completely the study of *local Fréchet differentiability* of Nemyckii operators: replace $f'(x,t)$ by $f'(x_*(t),t)$ in (6) and assume again continuity of $\psi_*(x,t;1)$ at $x_*(t)$ a.e. Then Lemma 3.1 and the Hölder inequality yield local Fréchet differentiability of $F:L_p(\Omega,\mathcal{X})\to L_q(\Omega,\mathcal{Y})$ at x_*. Just as in Theorem 3.3, $p>q$ is necessary and depends on a growth condition for $\psi_*(x,t;1)$. Hence, one technique applied to different remainder terms yields semismoothness or Fréchet differentiability at x_*, respectively. This unifies the treatment of both concepts and clarifies their relation.

Example 3.4. As an illustration consider the following simple class of examples:

$$f(x):=\max(x,\tau)^\sigma,\qquad f'(x):=\begin{cases}0 & : \quad x\le\tau,\\ \sigma x^{\sigma-1} & : \quad \text{otherwise},\end{cases}\qquad \tau\ge 0,\ \sigma\ge 1. \tag{10}$$

Except for special cases f is not differentiable at $x=\tau$ and not globally Lipschitz continuous for $\sigma>1$. For f, f', and $x_*:\Omega\to\mathbb{R}$ we will study $\psi_*(x,t;\alpha)$. Let now $t_0\in\Omega$ and $x_0:=x_*(t_0)$ be fixed. Then by (6),

$$\psi_*(x,t_0;\alpha)=\begin{cases}0 & : \quad x=x_0,\\ \dfrac{-\tau^\sigma+\max(x_0,\tau)^\sigma}{|x-x_0|^\alpha} & : \quad x_0\ne x\le\tau,\\ \dfrac{\sigma x^{\sigma-1}(x-x_0)-x^\sigma+\max(x_0,\tau)^\sigma}{|x-x_0|^\alpha} & : \quad \text{otherwise}.\end{cases} \tag{11}$$

If $\tau>0$ and $x_0\ne\tau$, then $\psi_*(x,t_0;\alpha)$ has a *jump* at $x=\tau$ and is thus not a Carathéodory function, but it is easy to see that it is a Baire–Carathéodory function. We will show now that $\lim_{x\to x_0}\psi_*(x,t_0;1)=0$, regardless of the choice of x_0.

If $x_0\ne\tau$, then f is differentiable in a neighborhood of x_0 and f' is Lipschitz continuous there. If $x_0=\tau$, then the same holds separately for both one-sided neighborhoods of x_0, excluding x_0. In both cases there is $\varepsilon(x_0)>0$ such that we can apply the fundamental theorem of calculus for all $\tilde{x}$ with $|\tilde{x}-x_0|\le\varepsilon$ to obtain

$$|\psi_*(\tilde{x},t_0;1)|\le\int_0^1|f'(\tilde{x})-f'(s\tilde{x}+(1-s)x_0)|\,ds\le L|\tilde{x}-x_0|. \tag{12}$$

Observe that this integral is independent of $f'(x_0)$. Hence, $\lim_{x\to x_0}\psi_*(x,t_0;1)=0$ and $\psi_*(x,t_0;\alpha)$ is bounded for $\alpha\le 2$. If $\sigma=1$, f' is piecewise constant and (12) vanishes, which implies boundedness of $\psi_*(x,t_0;\alpha)$ for all $\alpha\ge 1$.

However, and this is a particular feature of semismoothness, these results *do not hold uniformly* with respect to x_0 (and hence t_0). Because f' has a jump at $x=\tau$, the estimate (12) holds only for $\varepsilon(x_0)<|x_0-\tau|$. Also boundedness of $\psi_*(x,t_0;\alpha)$ depends on the choice of x_0. If x_0 is very close to τ, then this bound is very large, and tends to ∞, if $x_0\to\tau$.

If we consider $\psi_*(x,t;\alpha)$ for a continuous function $x_*:\Omega\to\mathbb{R}$, then this nonuniformity occurs if $x_*(t)=\tau$ for some $t\in\Omega$. As a consequence, L_s bounds for $\Psi_*(x;\alpha)$ depend on the size of the sets for which $x_*(t)$ is close to τ. In section 4 we discuss assumptions (essentially on x_*) that allow us to quantify this effect.

Concerning growth conditions in the case $\sigma>1$, f is locally Lipschitz continuous, and the corresponding local Lipschitz constant satisfies $L(x)\le a+b|x|^{\sigma-1}$ and also $|f'(x)|\le a+b|x|^{\sigma-1}$. Hence, by Proposition A.1, $\Psi_*(\cdot;1)$ maps $L_p(\Omega,\mathbb{R})$ into $L_s(\Omega,\mathbb{R})$ for $p/s\ge\sigma-1$. Thus, for successful application of Theorem 3.3 to an equation $G(x)=Tx+F(x)$, we have to show (ii) for $p/q\ge\sigma$.

If $\sigma=1$, then f is globally Lipschitz continuous. Then ψ_* is uniformly bounded and Theorem 3.3 holds if there is some $p>q$ such that (ii) holds (cf. Remark 3.2). This is the only case covered by previous results [12, 13, 6].

4. Boundedness of $\Psi_*(x;\alpha)$ and rates of convergence. One way of showing boundedness of an integral is limiting the size of the sets where the integrand is large. In the following we will consider an assumption that asserts $\Psi_*(x;\alpha)\in L_s(\Omega,\mathcal{Y})$ based on this principle. Following the ideas of [12], we define the set

$$\Omega_\varepsilon(x):=\left\{t\in\Omega:|\psi_*(x(t),t;\alpha)|_{\mathcal{Y}}>\frac{1}{\varepsilon^{\alpha-1}}\right\}\qquad\text{for }\alpha>1$$

and assume that there are $\rho>0$ and $C<\infty$ such that for its measure $\mu(\Omega_\varepsilon(x))$ the following bounds hold for some $\gamma>0$:

$$\sup_{\|x-x_*\|_{L_p}\le\rho}\mu(\Omega_\varepsilon(x))\le C\varepsilon^\gamma\qquad\forall\varepsilon>0. \tag{13}$$

This assumption means qualitatively that the sets of high nonlinearity are small near x_* (cf. the discussion in Example 3.4). In section 6 this assumption is reformulated as a strict complementarity assumption as known in constrained optimization.

Relations like (13) can be described conveniently in terms of the *distribution function* $\mathcal{S}_v$ of a measurable nonnegative function $v:\Omega\to\mathbb{R}_+$, defined by

$$\mathcal{S}_v(e):=\mu\left(\{t\in\Omega:v(t)>e\}\right).$$

The distribution function measures the size of the sets where v is large. Obviously, $\mathcal{S}$ is positive, monotonically decreasing, and bounded on bounded domains. It has already been useful in the convergence analysis of interior point methods in function space (cf. [11, 10]). We will use it now to obtain sharpened estimates for convergence rates, compared to [12].

LEMMA 4.1 (the distribution function). *Let Ω be a σ-finite measure space and*

$v\colon \Omega \to \mathbb{R}_+$ measurable and nonnegative. Then

$$\int_\Omega v(t)dt = \int_{[0,\infty]} \mathcal{S}_v(e)\, de, \tag{14}$$

$$\mathcal{S}_v(e) = \mathcal{S}_{v^s}(e^s). \tag{15}$$

Let $\varphi : [0,\infty] \to [0,\infty]$ be locally absolutely continuous, strictly monotone (increasing or decreasing), and bijective. Then

$$\int_\Omega v(t)dt = \int_{[0,\infty]} \mathcal{S}_v(\varphi(e))|\varphi'(e)|\, de. \tag{16}$$

Proof. Equation (14) is a special case of [15, Theorem 8.8]. Equation (16) follows from the substitution rule (cf. [15, Theorem 8.1]), which shows that with $\tilde{e} = \varphi(e)$

$$\int_{[0,\infty]} \mathcal{S}_v(\tilde{e})\, d\tilde{e} = \int_{[0,\infty]} \mathcal{S}_v(\varphi(e))|\varphi'(e)|\, de.$$

Equation (15) follows from

$$\mathcal{S}_{v^s}(e^s) = \mu\left(\{t \in \Omega : v(t)^s > e^s\}\right) = \mu\left(\{t \in \Omega : v(t) > e\}\right) = \mathcal{S}_v(e). \qquad \square$$

In terms of the distribution function our assumption (13) reads

$$\mathcal{S}_{|\Psi_*(x;\alpha)|_{\mathcal{Y}}}\left(\varepsilon^{1-\alpha}\right) \le \min\left\{\mu(\Omega); C\varepsilon^\gamma\right\} \qquad \forall \varepsilon > 0. \tag{17}$$

LEMMA 4.2. *If (13) holds for some $1 < \alpha < 1+\gamma s^{-1}$, then there is a neighborhood of x_* in X in which $\Psi_*(x;\alpha)$ is uniformly bounded in $L_s(\Omega, \mathcal{Y})$.*

Proof. By (15) and (17) we have

$$\mathcal{S}_{|\Psi_*(x;\alpha)|_{\mathcal{Y}}^s}\left(\varepsilon^{(1-\alpha)s}\right) = \mathcal{S}_{|\Psi_*(x;\alpha)|_{\mathcal{Y}}}\left(\varepsilon^{1-\alpha}\right) \le \min\left\{\mu(\Omega); C\varepsilon^\gamma\right\}. \tag{18}$$

By (16) with $\varphi(\varepsilon) = \varepsilon^{(1-\alpha)s}$ and thus $\varphi'(\varepsilon) = (1-\alpha)s\varepsilon^{(1-\alpha)s-1}$ we deduce

$$\begin{aligned}
\|\Psi_*(x;\alpha)\|_{L_s}^s &= \int_\Omega |\Psi_*(x;\alpha)|_{\mathcal{Y}}^s\, dt \\
&= \int_{[0,\infty]} \mathcal{S}_{|\Psi_*(x;\alpha)|_{\mathcal{Y}}^s}\left(\varepsilon^{(1-\alpha)s}\right) \left|(1-\alpha)s\varepsilon^{(1-\alpha)s-1}\right| d\varepsilon.
\end{aligned}$$

Inserting (18), we finally obtain for $\gamma + (1-\alpha)s > 0$ and $\alpha > 1$ the boundedness of

$$\begin{aligned}
\|\Psi_*(x;\alpha)\|_{L_s}^s &\le C \int_{[0,\infty]} \min\left\{\mu(\Omega); \varepsilon^\gamma\right\} \varepsilon^{(1-\alpha)s-1} d\varepsilon \\
&= C \int_{[0,\mu(\Omega)]} \varepsilon^{\gamma+(1-\alpha)s-1} d\varepsilon + C \int_{[\mu(\Omega),\infty]} \mu(\Omega)\varepsilon^{(1-\alpha)s-1} d\varepsilon \\
&\le c \lim_{\varepsilon\to 0} \varepsilon^{\gamma+(1-\alpha)s} + C + c \lim_{\varepsilon\to\infty} \varepsilon^{(1-\alpha)s}. \qquad \square
\end{aligned} \tag{19}$$

THEOREM 4.3. *Let $G : X \to Y$ be a semilinear operator as defined in (5), and suppose that Basic Assumptions 2.2 hold. Assume additionally the following:*

(i) $\psi_*(x,t;\alpha)$ *as defined in* (6) *is a Baire–Carathéodory function that satisfies assumption* (13) *for some* α *in the range*

$$1 < \alpha < \alpha_0 := \frac{1+\gamma q^{-1}}{1+\gamma p^{-1}}. \tag{20}$$

(ii) $\left\|G'(x)^{-1}\right\|_{L_q(\Omega,\mathcal{Y})\to L_p(\Omega,\mathcal{X})} \le M$ *holds uniformly on a neighborhood of* x_*.

Then Newton's method converges locally superlinearly to x_* *with the rate* α.

Proof. By definition of $\psi_*(\cdot,\cdot;\alpha)$ we have for $x, x_* \in X$

$$F'(x)(x-x_*) - (F(x)-F(x_*)) = \Psi_*(x;\alpha)|x-x_*|_{\mathcal{X}}^{\alpha}.$$

By the Hölder inequality we obtain for $s^{-1} + \alpha p^{-1} = q^{-1}$

$$\left\|\Psi_*(x;\alpha)|x-x_*|_{\mathcal{X}}^{\alpha}\right\|_{L_q} \le \left\|\Psi_*(x;\alpha)\right\|_{L_s} \left\|x-x_*\right\|_{L_p}^{\alpha}.$$

Application of Lemma 4.2 shows that $\left\|\Psi_*(x;\alpha)\right\|_{L_s}$ is uniformly bounded for $1 < \alpha < 1+\gamma s^{-1}$, which holds due to (20). Now we can estimate $\Theta(x)$ as defined in (3) and use (ii):

$$\begin{aligned}\Theta(x) &\le \frac{\left\|G'(x)^{-1}\right\|_{L_q\to L_p} \left\|F'(x)(x-x_*) - (F(x)-F(x_*))\right\|_{L_q}}{\left\|x-x_*\right\|_{L_p}} \\ &\le M \left\|\Psi_*(x;\alpha)\right\|_{L_s} \left\|x-x_*\right\|_{L_p}^{\alpha-1}.\end{aligned}$$

By Theorem 2.1 this yields local superlinear convergence with the rate α. ∎

Remark 4.4. Under the given restrictions on α, Theorem 4.3 gives us a result as good as we can expect, but often (13) holds for a whole range $\alpha \in [1, \overline{\alpha}]$ with $\overline{\alpha} \ge \alpha_0$. Then Theorem 4.3 holds only for all $\alpha < \alpha_0$. We may interpret this as Newton's method approaching the rate α_0 asymptotically from below, a behavior often called *convergence of order* α_0.

If we *additionally* assume that $\left\|\Psi_*(x;1)\right\|_{L_\infty}$ is uniformly bounded near x_* and $\alpha > \alpha_0$, then we can refine our results slightly. For this we have to use a splitting of Ω similar to that used in [12] or [6]. However, despite a considerably increased technical effort, no substantial improvement on α is possible. We merely obtain the closure of the interval (20).

THEOREM 4.5. *Let* $G : X \to Y$ *be a semilinear operator as defined in* (5), *suppose that Basic Assumptions* 2.2 *hold, and let* $p > q$. *Assume additionally the following:*

(i) $\psi_*(x,t;\alpha)$ *as defined in* (6) *is a Baire–Carathéodory function that satisfies assumption* (13) *for some* α *in the range*

$$\frac{1+\gamma q^{-1}}{1+\gamma p^{-1}} =: \alpha_0 < \alpha \le \frac{p}{q}. \tag{21}$$

(ii) $\left\|G'(x)^{-1}\right\|_{L_q(\Omega,\mathcal{Y})\to L_p(\Omega,\mathcal{X})} \le M$ *holds uniformly on a neighborhood of* x_*.

(iii) $\left\|\Psi_*(x;1)\right\|_{L_\infty}$ *is uniformly bounded in a neighborhood of* x_*.

Then Newton's method converges locally superlinearly to x_* *with the rate* α_0.

Proof. For $x, x_* \in X$ we abbreviate $R(x) := F'(x)(x-x_*) - (F(x)-F(x_*))$ and $\delta x := x - x_*$. Depending on a parameter κ, we divide Ω into two parts. In view of (13) we call Ω_κ the set where $|\Psi_*(x;\alpha)|_{\mathcal{Y}} > \kappa^{1-\alpha}$ and obtain $\mu(\Omega_\kappa) \le C\kappa^{\gamma}$. On Ω_κ we use

the boundedness of $\|\Psi_*(x;1)\|_{L_\infty}$ and the relation $\|v\|_{L_q(S)} \le \mu(S)^{(q^{-1}-p^{-1})} \|v\|_{L_p(S)}$, which holds for $p \ge q$ (cf. [1, Theorem 2.8]). This yields

$$\|R(x)\|_{L_q(\Omega_\kappa)} \le \|\Psi_*(x;1)\|_{L_\infty} \mu(\Omega_\kappa)^{(q^{-1}-p^{-1})} \|\delta x\|_{L_p(\Omega_\kappa)} \le C\kappa^{\gamma(q^{-1}-p^{-1})} \|\delta x\|_{L_p(\Omega)}.$$

Setting $\kappa := \|\delta x\|^\nu_{L_p(\Omega)}$ (with $\nu > 0$ to be chosen later), we conclude

$$\|R(x)\|_{L_q(\Omega_\kappa)} \le C \|\delta x\|^{\nu\gamma(q^{-1}-p^{-1})+1}_{L_p(\Omega)}. \tag{22}$$

On the remaining set $\Omega \setminus \Omega_\kappa$ we apply, as before, the Hölder inequality

$$\|R(x)\|_{L_q(\Omega\setminus\Omega_\kappa)} \le \|\Psi_*(x;\alpha)\|_{L_s(\Omega\setminus\Omega_\kappa)} \|\delta x\|^\alpha_{L_p(\Omega)} \tag{23}$$

with $s^{-1} := q^{-1} - \alpha p^{-1}$. By construction, $|\Psi_*(x;\alpha)|^s \le \kappa^{(1-\alpha)s}$ on $\Omega \setminus \Omega_\kappa$ and hence $\mathcal{S}_{|\Psi_*(x;\alpha)|^s_{\mathcal{Y}}}\left(\varepsilon^{(1-\alpha)s}\right) = 0$ for $\varepsilon < \kappa$. Consequently, just as in the proof of Lemma 4.2, we have the following estimate for $\alpha \ne 1$ and $\gamma + (1-\alpha)s \ne 0$:

$$\begin{aligned}\|\Psi_*(x;\alpha)\|^s_{L_s(\Omega\setminus\Omega_\kappa)} &\le C \int_{[\kappa,\infty]} \min\left\{\mu(\Omega); \varepsilon^\gamma\right\} \varepsilon^{(1-\alpha)s-1} d\varepsilon \\ &\le c\kappa^{\gamma+(1-\alpha)s} + C + c \lim_{\varepsilon\to\infty} \varepsilon^{(1-\alpha)s}.\end{aligned}$$

Hence, for $(1-\alpha)s < 0$, $\gamma + (1-\alpha)s < 0$ and for $\kappa \in [0, \overline{\kappa}]$ with arbitrary $\overline{\kappa} > 0$

$$\|\Psi_*(x;\alpha)\|_{L_s(\Omega\setminus\Omega_\kappa)} \le C(\kappa^{\gamma s^{-1}+1-\alpha} + 1) = C(\overline{\kappa})\kappa^{\gamma s^{-1}+1-\alpha}. \tag{24}$$

Hypothesis (21) ensures that the above inequalities for s^{-1} are valid. By our definition $\kappa := \|\delta x\|^\nu$ with $\nu > 0$, κ remains bounded in the following, since δx is always chosen from a bounded set. Thus we can drop the argument of $C(\overline{\kappa})$. Hence, inserting (24) with $\kappa = \|\delta x\|^\nu_{L_p(\Omega)}$ and $s^{-1} = q^{-1} - \alpha p^{-1}$ into (23), we obtain

$$\|R(x)\|_{L_q(\Omega\setminus\Omega_\kappa)} \le C \|\delta x\|^{\nu(\gamma s^{-1}+1-\alpha)+\alpha}_{L_p(\Omega)} \le C \|\delta x\|^{\nu\gamma(q^{-1}-\alpha p^{-1})-\nu(\alpha-1)+\alpha}_{L_p(\Omega)}. \tag{25}$$

Finally we compute the norm of $R(x)$ on Ω by adding both components:

$$\|R(x)\|^q_{L_q(\Omega)} = \|R(x)\|^q_{L_q(\Omega_\kappa)} + \|R(x)\|^q_{L_q(\Omega\setminus\Omega_\kappa)}.$$

The summands can be estimated by (22) and (25), respectively, and a choice of ν that balances both estimates will provide the sharpest results. Thus, comparing the exponents in (22) and (25), we choose ν such that

$$\nu\gamma(q^{-1} - p^{-1}) + 1 = \nu\gamma(q^{-1} - \alpha p^{-1}) - \nu(\alpha - 1) + \alpha. \tag{26}$$

Solving this linear equation for ν yields $\nu = (\gamma p^{-1} + 1)^{-1}$. Note that α cancels out. Inserting ν into (22), we finally obtain for the contraction $\Theta(x)$

$$\Theta(x) \le \left\|G'(x)^{-1}\right\|_{L_q\to L_p} \frac{\|R(x)\|_{L_q(\Omega)}}{\|\delta x\|_{L_p(\Omega)}} \le C \|\delta x\|^{\alpha_0-1}_{L_p(\Omega)},$$

with α_0 as defined in (21). This shows superlinear convergence with rate α_0. ∎

Relation to known estimates. In [12, Theorem 3.45] the rate of convergence of Newton's method is estimated under assumption (13) for the case of uniformly bounded ψ_*. Translating the notation used there into our framework, the rate of convergence β of Newton's method was estimated in that work by

$$\beta = \min\left\{\alpha_0\,;\,\alpha\cdot\frac{(\alpha-1)/\alpha+\tau}{(\alpha-1)+\tau}\right\},\quad \tau=\gamma(q^{-1}-p^{-1}).$$

Theorems 4.3 and 4.5 show that the second bound can essentially be replaced by α, which is an improvement because $\alpha>1$. Example 3.53 in [12] shows that the rate in Theorem 4.5 can be considered sharp.

5. Newton's method and compactness. Consider again the equation $G(x)=0$, where G is a semilinear operator satisfying (5). In section 3 we used continuity of the Nemyckii operator $\Psi_*(x;1)$ to show local superlinear convergence of Newton's method. By lack of continuity from $L_p(\Omega,\mathcal{X})$ to $L_\infty(\Omega,\mathcal{Y})$ (cf. Remark 3.2(ii)) we encountered a norm gap and needed an $L_q\to L_p$ smoothing property of $G'(x)^{-1}$.

The classical qualitative notion of a smoothing operation is that of a compact operator, and we will now explore its connection to convergence of Newton's method. By the Sobolev embedding theorems, compact embeddings in a space L_q usually imply some continuous embedding into a stronger space L_p (cf. [1]). So our considerations are mainly of theoretical interest, but due to the fundamental role of compactness in analysis this connection is worth investigating.

LEMMA 5.1. *Let $\mathcal{X}$, $\mathcal{Y}$ be separable Banach spaces and Ω a measurable subset of $\mathbb{R}^d$. Let $\psi:\mathcal{X}\times\Omega\to\mathcal{Y}$ be a Baire–Carathéodory function. Assume that there is a constant $M<\infty$ independent of x such that $|\psi(x,t)|_{\mathcal{Y}}\le M$ a.e. in Ω.*

For some $1\le p\le\infty$ let $x_*\in L_p(\Omega,\mathcal{X})$ be given, and let x_n be a sequence of functions that converges to x_* in $L_p(\Omega,\mathcal{X})$. If $\psi(x,t)$ is continuous with respect to x at $(x_*(t),t)$ for almost all $t\in\Omega$, then

$$\lim_{n\to\infty}\int_\Omega |\psi(x_n(t),t)-\psi(x_*(t),t)|_{\mathcal{Y}}^q\,v(t)\,dt=0\quad\forall\,v\in L_1(\Omega,\mathbb{R}),\ \ \forall\,1\le q<\infty. \tag{27}$$

Proof. Since $x_n\to x_*$ in $L_p(\Omega,\mathcal{X})$, by picking a suitable subsequence we may assume that $x_n(t)\to x_*(t)$ a.e. Thus, by continuity of $\psi(x,t)$ in x at $x_*(t)$ a.e. and by continuity of the power function, also $|\psi(x_n(t),t)-\psi(x_*(t),t)|_{\mathcal{Y}}^q\to 0$ a.e. Consequently, for each $v\in L_1(\Omega,\mathbb{R})$ we have

$$|\psi(x_n(t),t)-\psi(x_*(t),t)|_{\mathcal{Y}}^q\,v(t)\to 0\quad\text{ a.e. in }\Omega.$$

Because $|\psi(x,t)|_{\mathcal{Y}}\le M$ for all $x\in\mathcal{X}$ a.e., this sequence is dominated by $(2M)^q v\in L_1(\Omega,\mathbb{R})$, and the convergence theorem of Lebesgue yields (27). ∎

We conclude with a semismooth Newton theorem based on compactness. Just as L_q-L_p continuity of each $G'(x)^{-1}$ was not sufficient in Theorem 3.3 (we needed a uniform bound), compactness of each $G'(x)^{-1}$ is not sufficient in Theorem 5.2.

THEOREM 5.2. *Let $G:X\to Y$ be a semilinear operator as defined in* (5)*, suppose that Basic Assumptions* 2.2 *hold, and let $1<p<\infty$. Assume additionally the following:*

(i) *$\psi_*(x,t;1)$ as defined in* (6) *is a Baire–Carathéodory function, and*

$$\lim_{x\to x_*(t)}\psi_*(x,t;1)=0\quad\text{ for almost all } t\in\Omega.$$

The corresponding Nemyckii operator $\Psi_*(x;1)$ maps $L_p(\Omega,\mathcal{X})$ into $L_\infty(\Omega,\mathcal{Y})$.

(ii) *The linear space Y can be equipped with a norm $\|\cdot\|_Y$ such that $L_p(\Omega,\mathcal{Y})$ is compactly embedded into $(Y,\|\cdot\|_Y)$.*

(iii) $\left\|G'(x)^{-1}\right\|_{Y\to L_p(\Omega,\mathcal{X})} \le M$ *holds uniformly on a neighborhood of x_*.*

Then Newton's method converges locally superlinearly to x_*.

Proof. Let x_n be an arbitrary sequence in X with $\|x_n - x_*\|_{L_p} \to 0$, and define $R(x) := F'(x)(x-x_*) - (F(x) - F(x_*))$. Let p' be defined by $p'^{-1} + p^{-1} = 1$. Then $1 < p' < \infty$ and $L_{p'}(\Omega,\mathcal{Y}^*) = (L_p(\Omega,\mathcal{Y}))^*$. Consider $v \in L_{p'}(\Omega,\mathcal{Y}^*)$. By the Hölder inequality we estimate

$$|\langle R(x_n), v\rangle| \le \| |R(x_n)|_{\mathcal{Y}}\, |v|_{\mathcal{Y}^*} \|_{L_1} \le \|x_n - x_*\|_{L_p}\, \| |\Psi_*(x_n;1)|_{\mathcal{Y}}\, |v|_{\mathcal{Y}^*} \|_{L_{p'}}$$

$$= \|x_n - x_*\|_{L_p} \left(\int_\Omega |\psi_*(x_n(t),t;1)|_{\mathcal{Y}}^{p'}\, |v(t)|_{\mathcal{Y}^*}^{p'}\, dt \right)^{p'^{-1}}.$$

Because $\psi_*(x_*(t),t;1) = 0$ a.e., by Lemma 5.1 the last integral expression converges to 0 for each $|v|_{\mathcal{Y}^*}^{p'} \in L_1(\Omega)$. Thus, division by $\|x_n - x_*\|_{L_p}$ yields

$$r_n := \frac{R(x_n)}{\|x_n - x_*\|_{L_p}} \rightharpoonup 0 \quad \text{in } L_p(\Omega,\mathcal{Y}).$$

By (ii), $r_n \to 0$ in Y, and thus by (iii), $e_n := G'(x_n)^{-1} r_n \to 0$ in $L_p(\Omega,\mathcal{X})$. Hence,

$$\lim_{n\to\infty} \Theta(x_n) = \lim_{n\to\infty} \frac{\left\|G'(x_n)^{-1} R(x_n)\right\|_{L_p}}{\|x_n - x_*\|_{L_p}} = \lim_{n\to\infty} \|e_n\|_{L_p} = 0.$$

Because x_n was arbitrary, this implies superlinear convergence by Theorem 2.1. ∎

Often Y is the dual of some Sobolev space W. If there is a compact embedding $E : W \to L_{p'}$, then its adjoint $E^* : L_p \to Y$ with $p^{-1} + p'^{-1} = 1$ is also a compact embedding, suitable for Theorem 5.2.(ii). For a characterization of compactness in L_p spaces we refer to [1, Theorem 2.21].

6. Application to an optimal control problem. As an illustration and in view of Example 3.4 consider the following optimal control problem:

$$\min \frac{1}{2}\|y - y_d\|_{L_2(\Omega)}^2 + \frac{1}{1/\sigma+1}\int_\Omega |u|^{1/\sigma+1}\, dt \qquad \text{for } \sigma \ge 1$$

$$\text{s.t. } -\Delta y - u = 0, \quad y\big|_{\partial\Omega} = 0, \qquad u \ge \tau^\sigma \qquad \text{for } \tau \ge 0.$$

To compute (u,y) we introduce an adjoint state λ and consider the first-order optimality conditions:

$$\begin{aligned} y - y_d - \Delta\lambda = 0, \quad \lambda\big|_{\partial\Omega} = 0, \\ -\Delta y - \max(\lambda,\tau)^\sigma = 0, \quad y\big|_{\partial\Omega} = 0 \end{aligned} \tag{28}$$

with $u = \max(\lambda,\tau)^\sigma$. This is a semilinear system of equations of the form (5) with $x = (y,\lambda)$. For $\sigma = 1$ its solution in function space by Newton's method has been considered in [6, 12] by a formulation in terms of the control u. The consideration of the system (28) in terms of y and λ is an interesting alternative, which naturally describes algorithms based on the discretization idea of [7].

Constructing a linearization. The nonlinearity in this system is given by the function $f(\lambda) := \max(\lambda, \tau)^\sigma$, which has been considered in Example 3.4. Using f' as defined in (10), we can construct a Jacobian matrix $G'(x)$ to (28), given by

$$G'(x) = \begin{pmatrix} I & -\Delta \\ -\Delta & -f'(\lambda) \end{pmatrix}. \tag{29}$$

Let $x_* = (y_*, \lambda_*)$ be the solution of (28). Computation of the remainder term yields

$$G'(x_* + \delta x)\delta x - (G(x_* + \delta x) - G(x_*)) = \begin{pmatrix} 0 \\ R(\lambda) \end{pmatrix}.$$

Here $R(\lambda)(t) = \psi_*(\lambda(t), t; \alpha)|\lambda(t) - \lambda_*(t)|^\alpha$, with ψ_* defined in (11). The properties of ψ_* at x_* have already been established in Example 3.4. In particular, $\lim_{x\to x_*} \psi_*(x, t; 1) = 0$ and $|\psi_*(x, t; 1)| \le a(t) + b|x|^{\sigma-1}$. Hence, by Theorem 3.3, local superlinear convergence of Newton's method follows if an $L_q - L_p$ *smoothing property* holds for some $p/q \ge \sigma$. If $\sigma = 1$, then we can use either Theorem 3.3 with any $p > q$ or Theorem 5.2.

Solvability and smoothing property of $\mathrm{G'(x)^{-1}}$. Let us not go into the details of invertibility of G' in the case of a block structure as in (29). It is, however, not hard to see that G' inherits a smoothing property of the solution operator of the state equation. One approach is via a general theory for linear saddle point problems as can be found, for example, in [3, Chapter III.4].

Since the nonlinear terms in our system depend on the adjoint state λ only, it is sufficient to consider this component only, setting $x = \lambda$. By duality techniques it can be shown (cf., e.g., [10, Chapter 4]) that a smoothing property holds for $q^{-1} + p^{-1} = 1$ as long as for $-\Delta y = v$ we have $\|y\|_{L_p} \le c\,\|v\|_{L_2}$. We then obtain

$$\begin{pmatrix} \delta y \\ \delta\lambda \end{pmatrix} = G'(x)^{-1} \begin{pmatrix} 0 \\ R(\lambda) \end{pmatrix} \Longrightarrow \|\delta\lambda\|_{L_p} \le C\,\|R\|_{L_q}.$$

In particular, on regular domains $p = \infty$ and $q = 1$ are often obtained by H^2-regularity results for solutions of the state equation. Appropriate spaces for convergence analysis are then $X = (H^2(\Omega), \|\cdot\|_{L_\infty})$ and $Y = X^* \supset L_1(\Omega)$.

Rates of convergence. To study convergence rates we restrict our analysis to the case $\sigma = 1$ for simplicity. Following [12], we assume

$$\mu\left(\{t \in \Omega : 0 < |\lambda_*(t)| < \varepsilon\}\right) \le C\varepsilon^\gamma, \tag{30}$$

a condition that resembles a strengthened strict complementarity condition. If $|\lambda - \lambda_*(t)| < |\lambda_*(t)|$ or $\lambda_*(t) = 0$, then $r = 0$. Otherwise, we have $r < |\lambda_*(t)|$. Hence, for any $1 < \alpha < \infty$

$$\psi_*(\lambda, t; \alpha) \le \frac{|\lambda_*(t)|}{|\lambda - \lambda_*(t)|^\alpha} \le |\lambda_*(t)|^{1-\alpha},$$

and thus finally (30) becomes (13):

$$\mu(\Omega_\varepsilon(\lambda)) := \mu\left(\left\{t \in \Omega : \psi_*(\lambda(t), t; \alpha) < \varepsilon^{1-\alpha}\right\}\right) \le C\varepsilon^\gamma.$$

By Theorem 4.5 we obtain the rate of convergence α as in (21). In particular, for the very common case $\gamma = 1$, $q = 1$, $p = \infty$ we obtain $\alpha = 2$. Thus, Newton's method converges locally quadratically in function space.

Appendix.

PROPOSITION A.1 (a growth condition for $\psi_*(x,t;1)$). *For given $x_* \in L_p(\Omega,\mathcal{X})$, $f(x,t)$, and $f'(x,t)$ let $\psi_*(x,t;1)$ be defined as in* (6). *Define the local Lipschitz constant $L(x,t)$ of f with respect to x at (x,t) by*

$$L(x,t) := \sup_{|x-\tilde{x}|_{\mathcal{X}} \le 1} \frac{|f(x,t) - f(\tilde{x},t)|_{\mathcal{Y}}}{|x-\tilde{x}|_{\mathcal{X}}}.$$

Assume that

$$L(x,t) \le a(t) + b\,|x|_{\mathcal{X}}^{p/s} \qquad \text{for some } a \in L_s(\Omega,\mathbb{R}), b \in \mathbb{R}, \tag{31}$$

$$|f'(x,t)|_{\mathcal{X}\to\mathcal{Y}} \le a(t) + b\,|x|_{\mathcal{X}}^{p/s} \qquad \text{for some } a \in L_s(\Omega,\mathbb{R}), b \in \mathbb{R}. \tag{32}$$

Then $\psi_*(x,t;1)$ satisfies the growth condition (4).

Proof. By the triangle inequality,

$$|\psi_*(x,t;1)|_{\mathcal{Y}} \le \frac{|f'(x,t)(x-x_*(t))|_{\mathcal{Y}}}{|x-x_*(t)|_{\mathcal{X}}} + \frac{|f(x,t)-f(x_*(t),t)|_{\mathcal{Y}}}{|x-x_*(t)|_{\mathcal{X}}}.$$

Taking the operator norm of $f'(x,t)$, by (32) the first part of this sum satisfies (4).

To show the same for the second part we define for $n \in \mathbb{N}$ and $i = 0,\dots,n$ the collinear points $x_i := x \cdot i/n + x_*(t) \cdot (n-i)/n$. For sufficiently large n we can apply (31) to obtain

$$\frac{|f(x,t)-f(x_*(t),t)|_{\mathcal{Y}}}{|x-x_*(t)|_{\mathcal{X}}} \le \frac{\sum_{i=1}^n L(x_i,t)|x_i - x_{i-1}|_{\mathcal{X}}}{|x_0 - x_n|_{\mathcal{X}}} \le a(t) + \frac{b}{n}\sum_{i=1}^n |x_i|_{\mathcal{X}}^{p/s}. \tag{33}$$

Application of the triangle inequality yields

$$|x_i|_{\mathcal{X}}^{p/s} \le \left(|x|_{\mathcal{X}} \cdot \frac{i}{n} + |x_*(t)|_{\mathcal{X}} \cdot \frac{(n-i)}{n}\right)^{p/s} \le \left(\max\{|x|_{\mathcal{X}}, |x_*(t)|_{\mathcal{X}}\}\right)^{p/s}.$$

Inserting this into (33), using $x_* \in L_p(\Omega,\mathcal{X})$, yields the desired estimate. ∎

Often there is a relation $L(x,t) \approx \sup_{|\tilde{x}-x|_{\mathcal{X}}\le 1} |f'(\tilde{x},t)|$. Then one of (31) and (32) is redundant.

Acknowledgments. The author would like to thank Prof. Peter Deuflhard and Dr. Martin Weiser for support and encouragement and for valuable comments concerning this work. Further, the author wants to thank the referees, who contributed to the clarity and accuracy of this paper by a large number of valuable suggestions.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] J. APPELL AND P. P. ZABREJKO, *Nonlinear Superposition Operators*, Cambridge University Press, Cambridge, UK, 1990.
[3] D. BRAESS, *Finite Elements*, 2nd ed., Cambridge University Press, Cambridge, UK, 2001.
[4] P. DEUFLHARD, *Newton Methods for Nonlinear Problems. Affine Invariance and Adaptive Algorithms*, Ser. Comput. Math. 35, Springer, New York, 2004.

[5] I. Ekeland and R. Témam, *Convex Analysis and Variational Problems*, Classics Appl. Math. 28, SIAM, Philadelphia, 1999.
[6] M. Hintermüller, K. Ito, and K. Kunisch, *The primal-dual active set strategy as a semismooth Newton method*, SIAM J. Optim., 13 (2003), pp. 865–888.
[7] M. Hinze, *A variational discretization concept in control constrained optimization: The linear-quadratic case*, Comput. Optim. Appl., 30 (2005), pp. 45–63.
[8] B. Kummer, *Generalized Newton and NCP-methods: Convergence, regularity and actions*, Discuss. Math. Differential Incl., 20 (2000), pp. 209–244.
[9] E. Schechter, *Handbook of Analysis and Its Foundations*, Academic Press, New York, 1997.
[10] A. Schiela, *The Control Reduced Interior Point Method—A Function Space Oriented Algorithmic Approach*, Ph.D. thesis, Department of Mathematics and Computer Science, Free University of Berlin, Berlin, 2006.
[11] A. Schiela and M. Weiser, *Superlinear convergence of the control reduced interior point method for PDE constrained optimization*, Comput. Optim. Appl., 39 (2008), pp. 369–393.
[12] M. Ulbrich, *Nonsmooth Newton-like Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*, Habilitation thesis, Fakultät für Mathematik, Technische Universität München, Munich, 2002.
[13] M. Ulbrich, *Semismooth Newton methods for operator equations in function spaces*, SIAM J. Optim., 13 (2003), pp. 805–841.
[14] M. Ulbrich and S. Ulbrich, *Superlinear convergence of affine-scaling interior-point Newton methods for infinite-dimensional nonlinear problems with pointwise bounds*, SIAM J. Control Optim., 38 (2000), pp. 1938–1984.
[15] M. Väth, *Integration Theory. A Second Course*, World Scientific Publishing, River Edge, NJ, 2002.
[16] E. Zeidler, *Nonlinear Functional Analysis and Its Applications*, Vol. II/B, Springer, New York, 1990.

SIAM J. OPTIM.
Vol. 19, No. 3, pp. 1433–1450

NEW FORMULATIONS FOR OPTIMIZATION UNDER STOCHASTIC DOMINANCE CONSTRAINTS*

JAMES LUEDTKE†

Abstract. Stochastic dominance constraints allow a decision maker to manage risk in an optimization setting by requiring his or her decision to yield a random outcome which stochastically dominates a reference random outcome. We present new integer and linear programming formulations for optimization under first- and second-order stochastic dominance constraints, respectively. These formulations are more compact than existing formulations, and relaxing integrality in the first-order formulation yields a second-order formulation, demonstrating the tightness of this formulation. We also present a specialized branching strategy and heuristics which can be used with the new first-order formulation. Computational tests illustrate the potential benefits of the new formulations.

Key words. stochastic programming, stochastic dominance constraints, risk, probabilistic constraints, integer programming

AMS subject classifications. 90C15, 90C11

DOI. 10.1137/070707956

1. Introduction. Optimization under stochastic dominance constraints is an attractive approach to managing risk in an optimization setting. The idea is to optimize an objective, such as the expected profit, subject to a constraint that a random outcome of interest, such as the actual profit, is preferable in a strong sense than a given reference random outcome. Here, "preferable" is taken to mean that the random outcome we achieve stochastically dominates the reference outcome. A simple example application is to choose investments to maximize the expected return, subject to the constraint that the actual return should stochastically dominate the return from a given index, such as the S&P 500; see, e.g., [7]. Stochastic dominance constraints have also been used in risk modeling in power systems with dispersed generation [10]. In addition, dose-volume restrictions appearing in radiation treatment planning problems [18] can be formulated as a first-order stochastic dominance constraint. Stochastic programming under stochastic dominance constraints has recently been studied in [4, 5, 6, 8, 11, 12, 24, 25, 26].

Let W and Y be random variables with distribution functions F and G. The random variable W dominates Y in the first order, written $W \succeq_{(1)} Y$, if

$$F(\eta) \leq G(\eta) \qquad \forall\, \eta \in \mathbb{R}. \tag{1.1}$$

The random variable W dominates Y in the second order, written $W \succeq_{(2)} Y$, if

$$\mathbb{E}\big[\max\{\eta - W, 0\}\big] \leq \mathbb{E}\big[\max\{\eta - Y, 0\}\big] \qquad \forall\, \eta \in \mathbb{R}. \tag{1.2}$$

If W and Y represent random outcomes for which we prefer larger values, then stochastic dominance of W over Y implies a very strong preference for W. In particular, it is known that (see, e.g., [29]) $W \succeq_{(1)} Y$ if and only if

$$\mathbb{E}[h(W)] \geq \mathbb{E}[h(Y)]$$

*Received by the editors November 12, 2007; accepted for publication (in revised form) July 11, 2008; published electronically December 5, 2008.
http://www.siam.org/journals/siopt/19-3/70795.html
†Department of Industrial and Systems Engineering, University of Wisconsin, Madison, Wisconsin 53706 (jrluedt1@wisc.edu).

for all nondecreasing functions $h : \mathbb{R} \to \mathbb{R}$ for which the above expectations exist and are finite. Thus, if $W \succeq_{(1)} Y$, any rational decision maker would prefer W to Y. In addition, $W \succeq_{(2)} Y$ if and only if

$$\mathbb{E}[h(W)] \geq \mathbb{E}[h(Y)]$$

for all nondecreasing and concave functions $h : \mathbb{R} \to \mathbb{R}$ for which the above expectations exist and are finite. Thus, if $W \succeq_{(2)} Y$, any rational and risk-averse decision maker will prefer W to Y.

In this paper, we present new, computationally attractive formulations for *optimization under stochastic dominance constraints*. Let $X \subseteq \mathbb{R}^n$, and let $f : \mathbb{R}^n \to \mathbb{R}$ represent an objective we want to maximize. Let Y be a given random variable, which we refer to as the *reference random variable*, and let ξ be a random vector taking values in $\mathbb{R}^m$. Finally, let $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ be a given mapping which represents a random outcome depending on the decision x and the random vector ξ. We consider the two optimization problems

$$\max_x \left\{ f(x) : x \in X, g(x, \xi) \succeq_{(1)} Y \right\} \tag{FSDP}$$

and

$$\max_x \left\{ f(x) : x \in X, g(x, \xi) \succeq_{(2)} Y \right\}. \tag{SSDP}$$

We will present formulations for these problems for instances when the random vector ξ and reference random variable Y have finite distributions. That is, we assume ξ can take at most N values, and Y can take at most D values. In particular, we have the following.

1. We introduce two new linear formulations for SSDP which have $O(N + D)$ constraints, as opposed to $O(ND)$ constraints in an existing linear formulation. Computational results indicate that this yields significant improvement in solution time for instances in which $N = D$.

2. We introduce a new mixed-integer programming (MIP) formulation for FSDP which also has $O(N + D)$ constraints. In addition, the linear relaxation of this formulation is also a formulation of SSDP. As a result, the linear programming relaxation of this formulation is equivalent to the SSDP relaxation proposed in [24], and is shown to be a tight relaxation of FSDP in [25].

3. We present a specialized branching rule and heuristics for the new FSDP formulation and conduct computational tests which indicate that provably good, and in some cases provably optimal, solutions can be obtained for relatively large instances using this approach.

We do not make any assumptions on the set X or the mapping g in the development of the formulations, but computationally we are interested in the case when X is a polyhedron and $g(x, \xi)$ is affine in x for all possible values of ξ, so that the formulations become linear and linear integer programs, for SSDP and FSDP, respectively.

In [6] it is shown that in some special cases the convex second-order dominance constraint yields the convexification of the nonconvex first-order dominance constraint, and that in all cases, the second-order constraint is a relaxation of the first-order constraint. Our new formulations further illustrate this close connection by showing that relaxing integrality in the new formulation for FSDP yields a formulation for SSDP.

In section 2 we review some basic results about stochastic dominance and present existing formulations for FSDP and SSDP. In section 3 we present the new formula-

tions for SSDP, and in section 4 we present the new formulation for FSDP. In section 5 we present a specialized branching scheme and some heuristics for solving the new formulation of FSDP. In section 6 we present some illustrative computational results, and we close with some concluding remarks in section 7.

2. Review of existing results. For the purpose of developing formulations for FSDP and SSDP, it will be sufficient to present conditions which characterize instances when a random variable W stochastically dominates the reference random variable Y. We will assume the distributions of W and Y are finite and described by

$$\mu\{W = w_i\} = p_i, \qquad i \in \mathcal{N} := \{1, \dots, N\}, \tag{2.1}$$

$$\nu\{Y = y_k\} = q_k, \qquad k \in \mathcal{D} := \{1, \dots, D\}, \tag{2.2}$$

where μ and ν are the probability distributions induced by W and Y, respectively. Furthermore, we assume without loss of generality that $y_1 < y_2 < \cdots < y_D$.

Given a formulation which guarantees W stochastically dominates Y, a formulation for FSDP or SSDP can be obtained by simply enforcing that $g(x, \xi) \geq W$. Then, if ξ has distribution given by $\mathbb{P}\{\xi = \xi^i\} = p_i$ for $i \in \mathcal{N}$ and we add the constraints

$$w_i \leq g(x, \xi^i), \qquad i \in \mathcal{N} \tag{2.3}$$

to the formulation, then we will have $g(x, \xi) \succeq_{(1)} Y$ if and only if $W \succeq_{(1)} Y$, and $g(x, \xi) \succeq_{(2)} Y$ if and only if $W \succeq_{(2)} Y$. Henceforth, we will consider only formulations which guarantee stochastic dominance of W over Y, but based on the relation (2.3), the reader should think of the values w_i as decision variables, whereas the values y_k are fixed.

When the reference random variable Y has finite distribution, the conditions for stochastic dominance can be simplified, as has been observed, for example, in [4, 5]. We let $y_0 \in \mathbb{R}$ be such that $y_0 < y_1$ and introduce the notation $(\cdot)^+ = \max\{0, \cdot\}$.

LEMMA 2.1. *Let W, Y be random variables with distributions given by* (2.1) *and* (2.2). *Then, $W \succeq_{(2)} Y$ if and only if*

$$\mathbb{E}[(y_k - W)^+] \leq \mathbb{E}[(y_k - Y)^+], \qquad k \in \mathcal{D}, \tag{2.4}$$

and $W \succeq_{(1)} Y$ if and only if

$$\mu\{W < y_k\} \leq \nu\{Y \leq y_{k-1}\}, \qquad k \in \mathcal{D}. \tag{2.5}$$

The key simplification is that the infinite sets of inequalities in the definitions (1.1) and (1.2) can be reduced to a finite set when Y has a finite distribution.

Second-order stochastic dominance (SSD) constraints are known to define a convex feasible region [4]. In fact, condition (2.4) can be used to derive a linear formulation (in an extended variable space) for SSD by introducing variables s_{ik} representing the terms $(y_k - w_i)^+$; see, e.g., [4]. Thus, $W \succeq_{(2)} Y$ if and only if there exists $s \in \mathbb{R}_+^{ND}$ such that

$$\begin{aligned} &\sum_{i=1}^{N} p_i s_{ik} \leq \sum_{j=1}^{D} q_j (y_k - y_j)^+, \quad k \in \mathcal{D}, \\ &s_{ik} + w_i \geq y_k, \qquad\qquad\qquad\qquad i \in \mathcal{N}, k \in \mathcal{D}. \end{aligned}$$

We refer to this formulation as SDLP. Note that this formulation introduces ND variables and $(N+1)D$ constraints.

It is possible to use the nonsmooth convex constraints (2.4) directly, yielding a formulation for SSDP that does not introduce auxiliary variables and has $O(D)$ constraints; and specialized methods can be used to solve this formulation; see [5]. The advantage of using a linear formulation is that it can be solved directly by readily available linear programming solvers such as the open source solver CLP [9] or the commercial solver Ilog CPLEX [14]. In addition, if the base problem contains integer restrictions on some of the variables x, then a linear formulation is advantageous because it can be solved as a mixed-integer *linear* program, as opposed to a mixed-integer nonlinear program.

The condition for second-order dominance given in (2.4) can also be interpreted as a collection of D *integrated chance constraints*, as introduced by Klein Haneveld [15]. In [16], Klein Haneveld and van der Vlerk proposed a cutting plane algorithm for solving problems with integrated chance constraints and demonstrated its computational efficiency. Due to (2.4), this approach can also be used for problems with second-order stochastic dominance constraints, as has been observed in [8]. Independently, Rudolf and Ruszczyński [26] proposed a primal cutting plane method and a dual column generation method for optimization problems with SSD constraints, and the primal method is shown to be computationally efficient. In the case of finite distributions, the primal cutting plane method is equivalent to the cutting plane method used for integrated chance constraints in [16].

Condition (2.5) can be used to derive an MIP formulation for a first-order stochastic dominance (FSD) constraint [24, 25]. $W \succeq_{(1)} Y$ if and only if there exists β such that

$$
\begin{aligned}
\sum_{i=1}^{N} p_i \beta_{ik} &\le \sum_{j=1}^{k-1} q_j, \quad && k \in \mathcal{D}, \\
w_i + M_{ik}\beta_{ik} &\ge y_k, && i \in \mathcal{N}, k \in \mathcal{D}, \\
\beta_{ik} &\in \{0,1\}, && i \in \mathcal{N}, k \in \mathcal{D}.
\end{aligned} \tag{2.6}
$$

We refer to this formulation as FDMIP. Here, M_{ik} is sufficiently large to guarantee that if $\beta_{ik} = 1$, then the corresponding constraint (2.6) will be redundant. For example, if other constraints in the model imply $w_i \ge l_i$, then we can take $M_{ik} = y_k - l_i$. Although this formulation was presented in [24, 25], the authors do not recommend using this formulation for computation, since the linear programming relaxation bounds are too weak. Instead, because first-order stochastic dominance implies second-order dominance, any formulation for second-order dominance is a relaxation of first-order dominance, and the authors therefore propose using the problem SSDP as a relaxation for FSDP. Thus, they use the cutting plane algorithm proposed in [26] for solving problem SSDP, which yields bounds for FSDP, and then they improve these bounds using disjunctive cuts [1]. In addition, problem SSDP is used as a basis for heuristics to find feasible solutions for FSDP. It is demonstrated in [25] that the bounds from using SSDP as a relaxation of FSDP are usually good, and that the heuristics are able to obtain good feasible solutions. However, these results do not yield a convergent algorithm for finding an optimal solution to FSDP. As these results are based on solving problem SSDP, an easily implementable and computationally efficient formulation for solving SSDP will also enhance this approach.

3. New formulations for second-order stochastic dominance. When all outcomes are equally likely and $N = D$, a formulation for SSDP based on majorization theory [13, 21] can be derived which introduces $O(N^2)$ variables but only $O(N)$

rows. This has been done implicitly in [6] when proving that in this case the SSD constraint yields the convexification of the FSD constraint, and explicitly in [17] to derive a test for SSD. In this section we present two formulations for second-order dominance between finitely distributed random variables which do not require all outcomes to be equally likely and allow $N \neq D$. The formulations will not be based on the majorization theory, and instead will follow from the following theorem due to Strassen, which we state here in a form that is convenient for our use.

THEOREM 3.1 (see Corollary 1.5.21 in [22]). *Let W and Y be random variables with finite means. Then $W \succeq_{(2)} Y$ if and only if there exists random variables W' and Y', with the same distributions as W and Y, such that almost surely*

$$\mathbb{E}[Y'|W'] \leq W'.$$

THEOREM 3.2. *Let W, Y be random variables with distributions given by* (2.1) *and* (2.2). *Then $W \succeq_{(2)} Y$ if and only if there exists $\pi \in \mathbb{R}_+^{ND}$ which satisfies*

$$\sum_{j=1}^{D} y_j \pi_{ij} \leq w_i, \qquad i \in \mathcal{N}, \tag{3.1}$$

$$\sum_{j=1}^{D} \pi_{ij} = 1, \qquad i \in \mathcal{N}, \tag{3.2}$$

$$\sum_{i=1}^{N} p_i \pi_{ik} = q_k, \qquad k \in \mathcal{D}. \tag{3.3}$$

Proof. First suppose $W \succeq_{(2)} Y$. By Theorem 3.1, there exists random variables W' and Y' (defined, say, on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$) such that $\mathbb{E}[Y'|W'] \leq W'$ and $\mathbb{P}\{W' = w_i\} = p_i$ for $i \in \mathcal{N}$ and $\mathbb{P}\{Y' = y_k\} = q_k$ for $k \in \mathcal{D}$. Define a vector $\pi \in \mathbb{R}^{ND}$ by $\pi_{ik} = \mathbb{P}\{Y' = y_k | W' = w_i\}$ for $i \in \mathcal{N},\ k \in \mathcal{D}$. By definition, $\pi \geq 0$ and $\sum_{k \in \mathcal{D}} \pi_{ik} = 1$ for each $i \in \mathcal{N}$. Also, for each $k \in \mathcal{D}$

$$q_k = \mathbb{P}\{Y' = y_k\} = \sum_{i=1}^{N} \mathbb{P}\{Y' = y_k | W' = w_i\} \mathbb{P}\{W' = w_i\} = \sum_{i=1}^{N} p_i \pi_{ik}.$$

Finally, for each $i \in \mathcal{N}$

$$w_i \geq \mathbb{E}[Y'|W' = w_i] = \sum_{k=1}^{D} y_k \pi_{ik}$$

and hence π satisfies (3.1)–(3.3).

Now suppose there exists $\pi \in \mathbb{R}_+^{ND}$ which satisfies (3.1)–(3.3). Let $\Omega = \{(i,k) : i \in \mathcal{N}, k \in \mathcal{D}\}$, and define the probability measure $\mathbb{P}$ on Ω by $\mathbb{P}\{(i,k)\} = p_i \pi_{ik}$. Note that $\mathbb{P}$ is well defined since by (3.2) $\sum_{k \in \mathcal{D}} \sum_{i \in \mathcal{N}} p_i \pi_{ik} = 1$. Now define W' by $W'\big((i,k)\big) = w_i$ for $i \in \mathcal{N}$, $k \in \mathcal{D}$, and Y' by $Y'\big((i,k)\big) = y_k$ for $i \in \mathcal{N}$, $k \in \mathcal{D}$. Then, $\mathbb{P}\{W' = w_i\} = p_i \sum_{k \in \mathcal{D}} \pi_{ik} = p_i$ by (3.2), and so W' has the same distribution as W. Also, $\mathbb{P}\{Y' = y_k\} = \sum_{i \in \mathcal{N}} p_i \pi_{ik} = q_k$ by (3.3), and so Y' has the same distribution as Y. Finally, for each $i \in \mathcal{N}$,

$$\mathbb{E}[Y'|W' = w_i] = \sum_{k=1}^{D} y_k \pi_{ik} \leq w_i$$

by (3.1). It follows from Theorem 3.1 that $W \succeq_{(2)} Y$. □

To use Theorem 3.2 to obtain a formulation for SSDP, we replace w_i with $g(x, \xi^i)$ so that (3.1) becomes

$$g(x, \xi^i) \geq \sum_{j=1}^{D} y_j \pi_{ij}, \qquad i \in \mathcal{N}, \tag{3.4}$$

and thus obtain our first new formulation for SSDP given by

$$f^*_{\text{SSDP}} = \max_{x,\pi} \left\{ f(x) : \ (3.2),\ (3.3),\ (3.4), x \in X, \pi \in \mathbb{R}^{ND}_+ \right\}. \tag{cSSD1}$$

This formulation, which we refer to as cSSD1, introduces ND variables and $O(N+D)$ linear constraints.

THEOREM 3.3. *Let W, Y be random variables with distributions given by* (2.1) *and* (2.2). *Then $W \succeq_{(2)} Y$ if and only if there exists $\pi \in \mathbb{R}^{ND}_+$ which satisfies* (3.1), (3.2), *and*

$$\sum_{i=1}^{N} p_i \sum_{j=1}^{k-1} (y_k - y_j)\pi_{ij} \leq \sum_{j=1}^{k-1} (y_k - y_j), q_j \qquad k = 2, \ldots, D. \tag{3.5}$$

Proof. First suppose $W \succeq_{(2)} Y$. Then by Theorem 3.2 there exists $\pi \in \mathbb{R}^{ND}_+$ which satisfies (3.1)–(3.3). Then,

$$\sum_{i=1}^{N} p_i \sum_{j=1}^{k-1} (y_k - y_j)\pi_{ij} = \sum_{j=1}^{k-1} (y_k - y_j) \sum_{i=1}^{N} p_i \pi_{ij} = \sum_{j=1}^{k-1} (y_k - y_j) q_j$$

for $k = 2, \ldots, D$ by (3.3) and hence π satisfies (3.5).

Now suppose there exists $\pi \in \mathbb{R}^{ND}_+$ which satisfies (3.1), (3.2), and (3.5). For any $i \in \mathcal{N}$, $k \in \mathcal{D}$, we have

$$\begin{aligned}
(y_k - w_i)^+ &\leq \left(y_k - \sum_{j=1}^{D} y_j \pi_{ij} \right)^+ && \text{by (3.1)} \\
&= \left(\sum_{j=1}^{D} (y_k - y_j)\pi_{ij} \right)^+ && \text{by (3.2)} \\
&\leq \sum_{j=1}^{D} (y_k - y_j)^+ \pi_{ij} = \sum_{j=1}^{k-1} (y_k - y_j)\pi_{ij} && \text{since } \pi \geq 0.
\end{aligned}$$

Thus, for each $k \in \mathcal{D}$,

$$\begin{aligned}
\mathbb{E}[(y_k - W)^+] = \sum_{i=1}^{N} p_i (y_k - w_i)^+ &\leq \sum_{i=1}^{N} p_i \sum_{j=1}^{k-1} (y_k - y_j)\pi_{ij} \\
&\leq \sum_{j=1}^{k-1} (y_k - y_j) q_j = \mathbb{E}[(y_k - Y)^+],
\end{aligned}$$

where the second inequality follows from (3.5). Thus, condition (2.4) in Lemma 2.1 implies that $W \succeq_{(2)} Y$. □

When using the formulation arising from Theorem 3.3, it is beneficial for computational purposes to use an equivalent formulation in which we introduce variables $v \in \mathbb{R}^D$ and replace the constraints in (3.5) with the $2D$ constraints

$$v_j - \sum_{i=1}^{N} p_i \pi_{ij} = 0, \qquad j \in \mathcal{D}, \tag{3.6}$$

$$\sum_{j=1}^{k-1} (y_k - y_j) v_j \leq \sum_{j=1}^{k-1} (y_k - y_j) q_j, \qquad k \in \mathcal{D}. \tag{3.7}$$

Thus, our second formulation for SSDP is given by

$$\text{(cSSD2)} \quad f^*_{\text{SSDP}} = \max_{x,\pi,v} \left\{ f(x) : (3.2), (3.4), (3.6), (3.7), x \in X, \pi \in \mathbb{R}^{ND}_+, v \in \mathbb{R}^D \right\}.$$

The advantage of using (3.6) and (3.7) instead of (3.5) is that this yields a formulation with $O(ND)$ nonzeros, as compared to $O(ND^2)$ nonzeros if we used (3.5). This formulation, which we refer to as cSSD2, introduces $(N+1)D$ new variables and $O(N+D)$ linear constraints.

One motivation for introducing formulation cSSD2 is that we have empirical evidence (section 6) that it performs better than cSSD1, at least when solved with the dual simplex algorithm (as implemented in Ilog CPLEX [14]). cSSD2 is also interesting because a slight generalization of this formulation can be used to compactly model a collection of expected shortfall constraints of the form

$$\mathbb{E}[(y_k - g(x,\xi))^+] \leq L_k, \qquad k \in \mathcal{D}, \tag{3.8}$$

where $y_1 < y_2 < \cdots < y_D$ are given targets and $0 \leq L_1 \leq L_2 \leq \cdots \leq L_D$ are given limits on the expected shortfalls of these targets. Note that if

$$L_k = \mathbb{E}[(y_k - Y)^+], \qquad k \in \mathcal{D}, \tag{3.9}$$

where Y is a random variable with distribution given by (2.2), then the inequalities (3.8) are equivalent to (2.4), and hence (3.8) is satisfied exactly when $W \succeq_{(2)} Y$. If L_k is not defined by a random variable Y as in (3.9), formulation cSSD2 can still be extended to directly model (3.8), provided that $L_1 = 0$, which implies that we require $g(x,\xi) \geq y_1$ with probability 1. All that is required is to replace the term $\sum_{j=1}^{k-1}(y_k - y_j)q_j$ in the right-hand side of (3.5) with L_k.

4. A new formulation for first-order stochastic dominance. As in the case for second-order stochastic dominance, if $N = D$ and all outcomes are equally likely, a formulation for first-order stochastic dominance which introduces N^2 (binary) variables and $O(N)$ constraints has been presented in [17]. Once again, we are able to generalize this to the case in which the probabilities are not necessarily equal and $N \neq D$.

THEOREM 4.1. *Let W, Y be random variables with distributions given by* (2.1) *and* (2.2). *Then $W \succeq_{(1)} Y$ if and only if there exists $\pi \in \{0,1\}^{ND}$ such that (w, π) satisfy* (3.1), (3.2) *and*

$$\sum_{i=1}^{N} p_i \sum_{j=1}^{k-1} \pi_{ij} \leq \sum_{j=1}^{k-1} q_j, \qquad k = 2, \ldots, D. \tag{4.1}$$

Proof. First suppose $W \succeq_{(1)} Y$. Then, by condition (2.5) in Lemma 2.1 we have

$$\mu\{W < y_k\} \leq \nu\{Y \leq y_{k-1}\} = \sum_{i=1}^{k-1} q_i \tag{4.2}$$

for each $k \in \mathcal{D}$. In particular, (4.2) for $k = 1$ implies $\mu\{W \geq y_1\} = 1$, and hence $w_i \geq y_1$ for all $i \in \mathcal{N}$. Now, for each $i \in \mathcal{N}, k \in \mathcal{D}$, let $\pi_{ik} = 1$ if $y_k \leq w_i < y_{k+1}$ and $\pi_{ik} = 0$ otherwise, where we take $y_{D+1} \equiv +\infty$. Then, $\sum_{k=1}^{D} \pi_{ik} = 1$ because $w_i \geq y_1$ for all i, and so π satisfies (3.2). It is also immediate by the definition of π_{ik} that $w_i \geq \sum_{k=1}^{D} y_k \pi_{ik}$ and so π satisfies (3.1). Finally, note that $w_i < y_k$ if and only if $\sum_{j=1}^{k-1} \pi_{ij} = 1$. Thus,

$$\mu\{W < y_k\} = \sum_{i \in \mathcal{N}: w_i < y_k} p_i = \sum_{i=1}^{N} p_i \sum_{j=1}^{k-1} \pi_{ij}.$$

This combined with (4.2) proves that π satisfies (4.1).

Now suppose $\pi \in \{0,1\}^{ND}$ satisfies (3.1), (3.2), and (4.1). Note that by (3.1) and (3.2) if $w_i < y_k$, then $\sum_{j=1}^{k-1} \pi_{ij} = 1$. Thus,

$$\mu\{W < y_k\} = \sum_{i \in \mathcal{N}: w_i < y_k} p_i \leq \sum_{i=1}^{N} p_i \sum_{j=1}^{k-1} \pi_{ij} \leq \sum_{j=1}^{k-1} q_j = \nu\{Y \leq y_{k-1}\},$$

where the second inequality follows from (4.1). It follows that $W \succeq_{(1)} Y$ by condition (2.5) in Lemma 2.1. □

As in the new formulation for second-order stochastic dominance cSSD2, for computational purposes it is beneficial to use the equivalent formulation obtained by introducing variables $v \in \mathbb{R}^D$ and replacing the constraints (4.1) with the constraints

$$v_j - \sum_{i=1}^{N} p_i \pi_{ij} = 0, \qquad j \in \mathcal{D}, \tag{4.3}$$

$$\sum_{j=1}^{k-1} v_j \leq \sum_{j=1}^{k-1} q_j, \qquad k \in \mathcal{D}. \tag{4.4}$$

Thus, taking $w_i = g(x, \xi^i)$, and using (4.3) and (4.4) in place of (4.1), Theorem 4.1 yields the formulation for FSDP given by

$$\text{(cFSD)} \qquad f^*_{\text{FSDP}} = \max_{x,\pi} \left\{ f(x) : \ (3.2), (3.4), (4.3), (4.4), x \in X, \pi \in \{0,1\}^{ND} \right\}.$$

One advantage of formulation cFSD over FDMIP is that the number of constraints is reduced from $O(ND)$ to $O(N + D)$, which means it should be more efficient to solve the linear programming relaxation of cFSD than to solve that of FDMIP. We now consider the relationship between the relaxation of this formulation and *second*-order stochastic dominance.

THEOREM 4.2. *Let W, Y be random variables with distributions given by* (2.1) *and* (2.2). *Then the linear programming relaxation of* cFSD *yields a valid formulation for second-order stochastic dominance. That is, $W \succeq_{(2)} Y$ if and only if there exists $\pi \in \mathbb{R}^{ND}_+$ such that (w, π) satisfy* (3.1), (3.2), *and* (4.1).

Proof. Let $\pi \in \mathbb{R}_+^{ND}$ and (w, π) satisfy (3.1), (3.2), and (4.1). Then,

$$\begin{aligned}\sum_{i=1}^{N} p_i \sum_{j=1}^{k-1} \pi_{ij}(y_k - y_j) &= \sum_{i=1}^{N} p_i \sum_{j=1}^{k-1} \pi_{ij} \sum_{l=j+1}^{k} (y_l - y_{l-1}) \\ &= \sum_{l=2}^{k} (y_l - y_{l-1}) \sum_{i=1}^{N} p_i \sum_{j=1}^{l-1} \pi_{ij} \\ &\le \sum_{l=1}^{k} (y_l - y_{l-1}) \sum_{j=1}^{l-1} q_j \qquad \text{by (4.1)} \\ &= \sum_{j=1}^{k-1} q_j (y_k - y_j),\end{aligned}$$

and hence π also satisfies (3.5), which implies $W \succeq_{(2)} Y$ by Theorem 3.3.

Now suppose $W \succeq_{(2)} Y$. Then by Theorem 3.2 there exists $\pi \in \mathbb{R}_+^{ND}$ which satisfies (3.1)–(3.3). Then, (3.3) implies

$$\sum_{i=1}^{N} p_i \sum_{j=1}^{k-1} \pi_{ik} = \sum_{j=1}^{k-1} \sum_{i=1}^{N} p_i \pi_{ik} = \sum_{j=1}^{k-1} q_j$$

for $k = 2, \ldots, D$ and hence (4.1) holds. ∎

As a result, we obtain another formulation for SSDP, but more importantly we know that the linear programming relaxation of cFSD yields a bound at least as strong as the bound obtained from the SSDP relaxation.

Next, we illustrate the relationship between the formulation cFSD and FDMIP by presenting a derivation of cFSD based on strengthening FDMIP. In FDMIP, if $\beta_{ik} = 0$, then $w_i \ge y_k$. But, because $y_k > y_{k-1} > \cdots > y_1$, then we also know $w_i \ge y_{k-1} > \cdots > y_1$. Thus, we lose nothing by setting $\beta_{i,k-1} = \cdots = \beta_{i1} = 0$. Hence, we can add the inequalities

$$\beta_{ik} \le \beta_{i,k+1}, \qquad i \in \mathcal{N}, k \in \mathcal{D}, \tag{4.5}$$

and maintain a valid formulation. The inequalities (2.6) can then be replaced by

$$w_i - \sum_{k=1}^{D} (\beta_{i,k+1} - \beta_{ik}) y_k \ge 0, \qquad i \in \mathcal{N},$$

which together with inequalities (4.5) ensure that when $\beta_{ik} = 0$, we have $w_i \ge y_k$. We finally obtain the new formulation cFSD by substituting $\pi_{ik} = \beta_{i,k+1} - \beta_{ik}$ for $k \in \mathcal{D}$ and $i \in \mathcal{N}$, where $\beta_{i,D+1} = 1$.

5. Branching and heuristics for FSDP. cFSD yields a mixed-integer programming formulation for FSDP. Moreover, if X is a polyhedron and $g(x, \xi^i)$ are affine in x for each i, then cFSD is a mixed-integer *linear* programming formulation. As has been shown in [25], the optimal value of SSDP yields a good bound on the optimal value of FSDP, and hence the bound obtained from relaxing integrality in cFSD should be good. In addition, because of the compactness of cFSD, this bound can be calculated efficiently. However, we have found that the default settings in the MIP solver we use (Ilog CPLEX 9.0 [14]) do not effectively generate good feasible solutions

for cFSD. In addition, the default branching setting does not help to find feasible solutions or effectively improve the relaxation bounds. In this section we present a specialized branching approach and two heuristics which exploit the structure of this formulation. The computational benefits of these techniques will be demonstrated in section 6.

5.1. Branching for FSDP. Standard variable branching for mixed-integer programming would select a variable π_{ij} which is fractional in the current node relaxation solution, and then branch to create two new nodes, one with π_{ij} fixed to one and one with π_{ij} fixed to zero. However, the constraints (3.1) and (3.2) imply that for a fixed i, the set of variables π_{ij} for $j \in \mathcal{D}$ are essentially selecting which value level y_j the variable w_i should be greater than. In particular, the set of variables $\{\pi_{ij} : j \in \mathcal{D}\}$ is a *special order set of type* 1 (SOS1), that is, at most one of the variables in this set can be positive. As a result, it is natural to consider using an SOS1 branching rule (see, e.g., [2]). In this branching scheme, we select a *set index* $i \in \mathcal{N}$, specifying which special ordered set to branch on, and also choose a *level index* $k \in \{2, \dots, D\}$. Then in the first branch the constraint $\sum_{j<k} \pi_{ij} = 0$ is enforced and in the second branch $\sum_{j<k} \pi_{ij} = 1$ is enforced. In an implementation, the first condition is enforced by changing the upper bound on the variables π_{ij} to zero for $j < k$, and the second condition is enforced by changing the upper bound on the variables π_{ij} to zero for $j \geq k$.

To specify an SOS1 branching rule, we must state how the set and level indices are chosen. Our branching scheme is based on attempting to enforce the feasibility condition (2.5),

$$\mu\{W < y_k\} \leq \nu\{Y \leq y_{k-1}\}, \qquad k \in \mathcal{D}.$$

At each node in which we must branch, we find $k^* = \min\{k \in \mathcal{D} : \mu\{W < y_k\} > \nu\{Y \leq y_{k-1}\}\}$ based on the values of w in the current relaxation solution. Note that if such a k^* does not exist, then we have $W \succeq_{(1)} Y$, so the current solution is feasible. In this case, if π is not integer feasible (which may happen), then we construct an integer feasible solution of the same cost as in the proof of Theorem 4.1, and as a result, branching is not required at this node. We will take k^* to be the level index on which we will branch. Note that (3.1) and (3.2) imply that $w_i \geq y_1$ for all i in any relaxation solution, so that $k^* \geq 2$, making it an eligible branching level index.

We next choose a set index $i \in \mathcal{N}$ such that

$$w_i < y_{k^*}, \tag{5.1}$$

$$\sum_{j<k^*} \pi_{ij} < 1. \tag{5.2}$$

We claim that such an index must exist. Indeed, let $\Omega_{k^*} = \{i \in \mathcal{N} : w_i < y_{k^*}\}$. By the definition of k^* we have $\sum_{i\in\Omega_{k^*}} p_i > \sum_{j=1}^{k-1} q_j$, and so, in particular, $\Omega_{k^*} \neq \emptyset$. If there were no $i \in \Omega_{k^*}$ which also satisfies (5.2), then we would have

$$\sum_{i=1}^{N} p_i \sum_{j=1}^{k^*-1} \pi_{ij} \geq \sum_{i\in\Omega_{k^*}} p_i > \sum_{j=1}^{k^*-1} q_j,$$

violating (4.1). If there are multiple set indices which satisfy (5.1) and (5.2), we choose an index which maximizes the product $(y_{k^*} - w_i)(1 - \sum_{j<k^*} \pi_{ij})$. In the

first branch, we enforce $\sum_{j<k^*} \pi_{ij} = 0$, which by (3.1) forces $w_i \geq y_{k^*}$. Because of (5.1), this will make the current relaxation solution infeasible to this branch, and will promote feasibility of (2.5) at the currently infeasible level k^*. In the second branch, we enforce $\sum_{j<k^*} \pi_{ij} = 1$, which because of (5.2) will make the current relaxation solution infeasible for this branch. The motivation for this choice of set index i is to make progress in both of the branches. The motivation for the choice of level index k^* is that in the first branch progress toward feasibility of (2.5) is made, whereas by selecting k^* as small as possible, reasonable progress is also made in the second branch since this enforces $\pi_{ij} = 0$ for all $j \geq k^*$.

5.2. Heuristics for FSDP. We now present some heuristics we have developed that can be used with formulation cFSD. We first present a simple and efficient heuristic, called the *order-preserving heuristic,* and then present a variant of a diving heuristic which can be integrated with the order-preserving heuristic.

Order-preserving heuristic. Given a solution x^* to a relaxation of cFSD, let $w^* \in \mathbb{R}^N$ be the vector given by $w_i^* = g(x^*, \xi^i)$ for $i \in \mathcal{N}$. The idea behind the order-preserving heuristic is to use w^* as a guide to build a solution $\hat{\pi} \in \{0,1\}^{ND}$ which satisfies (3.2) and (4.1), and then solve the problem with π fixed to $\hat{\pi}$. If this problem is feasible, it yields a feasible solution to cFSD. The heuristic is *order-preserving* because it chooses $\hat{\pi}$ in such a way that if $w_i^* < w_{i'}^*$, then $\sum_{j\in\mathcal{D}} y_j \hat{\pi}_{ij} \leq \sum_{j\in\mathcal{D}} y_j \hat{\pi}_{i'j}$, so that the constraints (3.4) obtained with this $\hat{\pi}$ enforce lower bounds on $g(x, \xi^i)$ which are consistent with the ordering of $w_i^* = g(x^*, \xi^i)$ obtained from the current relaxation solution. The order-preserving heuristic is given in Algorithm 1. The algorithm begins by sorting the values of w^*. Then, in lines 2 to 8 a solution $\hat{\pi}$ is constructed which is feasible to (3.2) and (4.1) by working in this order. To see that $\hat{\pi}$ satisfies (3.2), observe that the algorithm will terminate with $t = N+1$, since when $k = D$, $\sum_{j=1}^{t} p_{i_j} \leq \sum_{j=1}^{D} q_j$ for all $t \leq N$, so the loop on line 4 will terminate only when $t > N$. Since $\{i_1, \ldots, i_N\} = \mathcal{N}$, this implies that for each $i \in \mathcal{N}$, there is some k such that the algorithm sets $\hat{\pi}_{ik} = 1$. The condition $\sum_{j=1}^{t} p_{i_j} \leq \sum_{j=1}^{k} q_j$ in line 4

Algorithm 1. Order-preserving heuristic

Data: $w^* \in \mathbb{R}^N$

1 Sort w^* to obtain $\{i_1, \ldots, i_N\} = \mathcal{N}$ with $w_{i_1}^* \leq w_{i_2}^* \leq \cdots \leq w_{i_N}^*$;
2 Set $t := 1$ and $\hat{\pi}_{ij} := 0$ for all $i \in \mathcal{N}, j \in \mathcal{D}$;
3 **for** $k := 1$ **to** D **do**
4 **while** $t \leq N$ **and** $\sum_{j=1}^{t} p_{i_j} \leq \sum_{j=1}^{k} q_j$ **do**
5 $\hat{\pi}_{i_t k} := 1$;
6 $t := t + 1$;
7 **end**
8 **end**
9 Solve cSSD($\hat{\pi}$) $= \max_x \{f(x) : x \in X, g(x, \xi_i) \geq \sum_{j=1}^{D} \hat{\pi}_{ij} y_j \ i \in \mathcal{N}\}$;
10 **if** cSSD($\hat{\pi}$) *is feasible* **then**
11 Let $\hat{x}$ be the optimal solution to cSSD($\hat{\pi}$);
12 **return** $(\hat{x}, \hat{\pi})$;
13 **end**

ensures that (4.1) holds for $\hat{\pi}$, since it ensures that for each $k \in \mathcal{D}$,

$$\sum_{j=1}^{k} \sum_{i=1}^{N} p_i \hat{\pi}_{ij} = \sum_{j=1}^{t(k)} p_{i_j},$$

where $t(k) = \max\{t : \sum_{j=1}^{t} p_{i_j} \leq \sum_{j=1}^{k} q_j\}$.

The main work done in Algorithm 1 is the sorting of w^* and the solving of cSSD($\hat{\pi}$). Note that this problem is small relative to the original problem cFSD, since the $O(ND)$ variables π are fixed, the constraints (3.2) and (4.1) no longer need to be considered, and the constraints (3.4) reduce to lower bounds on the functions $g(x, \xi^i)$ for $i \in \mathcal{N}$.

Integrated order-preserving and diving heuristic. Diving is a classic heuristic strategy for integer programs which alternates between fixing one or more integer variables based on the current linear programming (LP) relaxation solution and resolving the relaxation. We have developed a variant of the diving heuristic for solving cSSD which we call the *aggressive diving heuristic*. For brevity, we only outline the idea of the heuristic here; for details, we refer the reader to [19]. Within each iteration of the aggressive diving heuristic, the heuristic repeatedly selects the index $i \in \mathcal{N}$ which has minimum value of $w_i^* = g(x^*, \xi^i)$ and has not yet had π_{ij} fixed to one for any $j \in \mathcal{D}$. A variable π_{ik} is then fixed to one, where k is the minimum index such that π_{ik} could feasibly be fixed to one and still satisfy (4.1). This is done until one of the fixings causes inequality (3.4) to be violated by the current solution, that is, until a π_{ik} is fixed to one with $w_i^* < y_k$. Also within an iteration, a similar sequence of fixings is done for indices $i \in \mathcal{N}$ which have *maximum* value of w_i^* until one of the fixings implies (3.4) is violated by the current solution. After these fixings have been done, the LP relaxation is resolved, and the next iteration begins. The heuristic terminates when the current LP relaxation yields a feasible integer solution or is infeasible (where infeasibility would be caused by the lower bounds implied by (3.4) due to the fixed variables). The key advantages of the aggressive diving heuristic are that it fixes multiple variables in each iteration, leading to faster convergence, and that the variables are fixed in such a way that constraints (3.2) and (4.1) will not become violated.

Integration of the order-preserving heuristic with the aggressive diving heuristic is accomplished by calling the order-preserving heuristic during each iteration of the diving heuristic, using the current relaxation solution. If this yields an improved feasible solution, it is saved, but the heuristic still continues the dive until it terminates. At the end, the best feasible solution found over all iterations in the dive is reported.

6. Computational results. We conducted computational experiments to test the new formulations for stochastic dominance. Following [17] and [25], we conducted tests on a portfolio optimization problem with stochastic dominance constraints. In this problem, we wish to choose the fraction of our investment to invest in n different assets. The return of asset j is a random variable given by R_j with $\mathbb{E}[R_j] = r_j$. We are also given a reference random variable Y, and the objective is to maximize the expected return subject to the constraint that the random return we achieve stochastically dominates Y. Thus, the portfolio optimization problems we consider are

$$\max\left\{\sum_{j=1}^{n} r_j x_j : \; x \in X, \; \sum_{j=1}^{n} R_j x_j \succeq_{(k)} Y\right\}, \qquad k = 1, 2, \tag{6.1}$$

where $X = \{x \in \mathbb{R}^n_+ : \sum_{j=1}^n x_j = 1\}$.

We constructed test instances using the daily returns of 435 stocks ($n = 435$) in the S&P 500, for which daily return data was available from January 2002 through March 2007. We take each daily return as an outcome that occurs with equal probability. For each desired number of outcomes N, we constructed three instances by taking the N daily returns immediately preceding March 14 of the years 2005, 2006, and 2007. For example, the instance for the year 2007 with $N = 100$ is obtained by taking the daily returns in the days from November 16, 2006 through March 14, 2007.

For the reference random variable Y, we use the returns that would be obtained by investing an equal fraction in each of the available assets. That is, we take $Y = \sum_{j=1}^n R_j/n$. Hence, if R^i_j is the return that is achieved under outcome i for asset j, then the distribution of Y is given by $\nu\{Y = \sum_{j=1}^n R^i_j/n\} = 1/N$ for $i \in \mathcal{N}$. Note that in this case, the number of outcomes of Y is the same as the number of outcomes of R, i.e., $D = N$. This is an extreme case: in many settings we would expect D to be significantly less than N. However, this extreme case will yield challenging instances for comparing the formulations.

We used CPLEX 9.0 [14] to solve the LP and MIP formulations, and all experiments were done on a computer with two 2.4 GHz processors (although no parallelism is used) and 2.0 GB of memory. The specialized heuristics and branching for FSDP were implemented using callback routines provided by the CPLEX callable library.

6.1. Second-order dominance. We first compared the solution times using the formulations SDLP, cSSD1, and cSSD2 to solve the portfolio optimization problem (6.1) with SSD constraint ($k = 2$ in (6.1)). We tested seven different sizes N and three instances for each size. These linear programs were solved using the dual simplex method (the default CPLEX setting), and a time limit of 100,000 seconds was used. Table 6.1 gives the solution time and number of simplex iterations for each formulation

TABLE 6.1
Computational results for SSDP *formulations. * indicates not solved within time limit.*

		Solution time(s)			Iterations		
Year	N	SDLP	cSSD1	cSSD2	SDLP	cSSD1	cSSD2
2005	200	103	18	3	30851	10336	1921
	300	1063	61	19	76438	16879	6019
	400	4859	127	23	118328	17692	5698
	500	10345	509	17	121067	34380	4770
	600	27734	528	39	202490	40430	5854
	700	69486	3366	434	318030	112848	20788
	800	*100122	8272	1476	*361600	222967	42773
2006	200	83	13	3	20009	7449	2134
	300	883	44	8	52457	12004	3244
	400	4253	190	25	109493	24398	5549
	500	11365	332	63	117559	37086	7904
	600	43927	670	198	307680	41360	16443
	700	58947	6067	94	346077	173483	13026
	800	*100100	10406	50	*433400	245401	6307
2007	200	122	25	9	19359	13771	4597
	300	757	61	30	64795	15585	8253
	400	4292	214	59	89731	28024	8265
	500	12551	609	178	154287	46973	14914
	600	27492	1213	271	172905	66164	18611
	700	59144	1888	338	308064	92365	19009
	800	*100095	23171	74	*385700	544342	8174

on each instance. From this table it is clear that when using a commercial LP solver, the new formulations cSSD1 and cSSD2 allow for a much more efficient solution of SSDP. Formulation cSSD1 yields a solution an order of magnitude faster than SDLP, whereas cSSD2 yields a solution roughly two orders of magnitude faster. Both formulations cSSD1 and cSSD2 have $O(N)$ rows as opposed to $O(ND) = O(N^2)$ rows in SDLP, leading to a significantly reduced basis size, so that the time per iteration using these formulations is significantly less. The additional reduction in computation time obtained from formulation cSSD2 can be explained by the large reduction in the number of simplex iterations.

We should stress that because $N = D$ in this test, the relative improvement of cSSD1 and cSSD2 over SDLP is likely the best case. For instances in which D is of much more modest size, such as $D = 10$, we would not expect such an extreme difference.

6.2. First-order dominance. We next present results of the tests on the portfolio optimization problem (6.1) in which a first-order stochastic constraint is enforced ($k = 1$ in (6.1)).

We tested four solution methods for solving FSDP:

(i) FDMIP: Solve FDMIP with default CPLEX settings.

(ii) cFSD: Solve cFSD with default CPLEX settings and CPLEX SOS1 branching.

(iii) cFSD+H: Solve cFSD with CPLEX SOS1 branching and specialized heuristic.

(iv) cFSD+H+B: Solve cFSD with CPLEX, specialized heuristic, and specialized branching.

When solving cFSD with or without the heuristic (but not with the specialized branching), we declare the sets of variables $\{\pi_{ij} : j \in \mathcal{D}\}$ for $i \in \mathcal{N}$ as SOS1, allowing CPLEX to perform its general purpose SOS1 branching, as discussed in section 5.1. We found that this yields better results than having CPLEX perform its default single variable branching. Note that the specialized branching scheme also uses SOS1 branching, but crucially differs from the CPLEX implementation in the selection of the SOS1 set and level on which to branch.

The heuristic used in the last two methods is the aggressive diving heuristic integrated with the order-preserving heuristic. In our implementation, we call the heuristic at every node of depth less than five, at every fifth node for the first 100 nodes, at every 20th node between 100 and 1000 nodes, and at every 100th node thereafter. When the heuristic is used we turn off the CPLEX heuristics and preprocessing. The preprocessing was turned off for implementation convenience, but we found it had little effect for formulation cFSD anyway.

The specialized branching used in the last method is the branching strategy given in section 5.1. For this case, we set the CPLEX branching variable selection to select the most fractional variable since this takes the least time and we do not use CPLEX's choice of branching variable anyway.

We first compare the time required to solve the root linear program relaxations and the resulting lower bound from formulations FDMIP and cFSD. These results are given in Table 6.2. For formulation FDMIP we report the results before and after the addition of CPLEX cuts. The results obtained after the addition of CPLEX cuts are under the FDMIP.C column. For cFSD, we report only the results after the initial relaxation solution, because CPLEX cuts had little effect in this formulation. The columns under the heading "Percent above cFSD UB report the percent by which the

TABLE 6.2
Comparison of root LP relaxations for FSDP *formulations.*

		Time(s)			Percent above cFSD UB	
Year	N	cFSD	FDMIP	FDMIP.C	FDMIP	FDMIP.C
2005	100	1.0	6.6	41.4	5.36%	3.46%
	150	1.8	19.7	89.5	7.64%	6.18%
	200	4.7	36.3	196.2	8.42%	5.87%
	250	15.1	49.9	365.0	9.34%	6.78%
	300	31.0	232.6	681.5	9.78%	7.50%
	350	88.0	509.7	1201.0	4.36%	3.05%
	400	97.6	427.7	1566.2	5.14%	3.19%
2006	100	0.4	3.9	4.3	0.21%	0.00%
	150	3.8	16.2	82.0	1.54%	1.03%
	200	4.8	26.3	140.9	1.38%	1.08%
	250	17.5	91.1	325.8	3.99%	2.45%
	300	16.4	191.3	575.6	4.60%	3.53%
	350	52.3	227.7	1157.8	8.49%	6.52%
	400	69.1	1254.7	2188.6	6.92%	5.77%
2007	100	2.0	4.5	33.5	7.55%	3.70%
	150	8.1	17.0	148.4	7.69%	6.06%
	200	17.8	33.3	300.8	9.75%	8.26%
	250	36.1	121.4	413.1	14.13%	10.71%
	300	43.5	298.6	732.6	11.12%	8.26%
	350	114.0	320.9	1060.7	10.80%	10.60%
	400	245.7	2010.8	3664.2	11.53%	11.02%

upper bound obtained from the relaxation of FDMIP with or without cuts, exceeds the upper bound obtained from the relaxation of cFSD. It is clear from Table 6.2 that the relaxation of formulation cFSD provides significantly better upper bounds in significantly less time.

We next tested how the different methods performed when run for a time limit of 10,000 seconds. Table 6.3 reports the optimality gap remaining after this time limit. All formulations were able to solve the 2006 instance with $N = 100$ in less than a minute, so this instance is excluded. Using formulation cFSD with the heuristic and specialized branching, 8 of the remaining 20 instances were solved to optimality within the time limit, and for these instances the solution time is reported (these are the instances with "-" in the "Gap" column). From Table 6.3 we observe that even without the use of specialized heuristic or branching formulation cFSD outperforms formulation FDMIP. However, in several instances cFSD fails to find a feasible solution, and in several others the optimality gaps for the feasible solutions found are quite bad. This is remedied to a significant extent by using the specialized heuristic, in which case a feasible solution is found for every instance, and in most cases it is within 2% of the upper bound. If, in addition, we use the specialized branching scheme, the final optimality gaps are reduced even further, with many of the instances being solved to optimality.

Table 6.4 gives more detailed results for the methods based on formulation cFSD for the 2005 instances (results for the other instances yield similar insights and are excluded for brevity). First, for each of these methods, the table indicates the percent by which the final upper bound ("UB" in the table) was below the initial upper bound obtained simply from solving the LP relaxation ("Root UB" in the table). These results indicate that by using CPLEX branching, with or without the specialized heuristic, very little progress is made in improving the upper bound through branching. In contrast, the specialized branching scheme improves the upper bound

TABLE 6.3
Comparison of optimality gaps for FSDP *after time limit. ** indicates no feasible solution found.*

		Optimality gap			cFSD+H+B	
Year	N	FDMIP	cFSD	cFSD+H	Gap	Time(s)
2005	100	1.69%	0.68%	0.68%	-	864.0
	150	2.84%	0.99%	0.73%	-	223.1
	200	4.46%	1.09%	0.87%	-	1987.3
	250	8.82%	0.31%	0.24%	-	2106.6
	300	**	3.41%	1.21%	1.15%	
	350	**	**	2.15%	1.39%	
	400	**	10.67%	0.73%	0.31%	
2006	150	1.71%	0.77%	0.55%	0.18%	
	200	1.25%	0.57%	0.55%	-	1752.1
	250	4.82%	0.97%	0.44%	-	274.9
	300	4.56%	4.24%	0.85%	-	9386.8
	350	**	1.96%	0.65%	0.53%	
	400	**	4.77%	1.21%	0.87%	
2007	100	0.13%	0.14%	0.15%	-	41.6
	150	13.90%	4.11%	2.37%	1.85%	
	200	**	3.80%	1.64%	0.67%	
	250	**	9.13%	2.12%	0.67%	
	300	**	**	2.43%	2.01%	
	350	**	**	6.74%	6.37%	
	400	**	**	5.82%	5.79%	

TABLE 6.4
Lower and upper bounds results using cFSD. *** indicates no feasible solution found.*

		% UB below Root UB			% LB below Best UB		
Year	N	cFSD	+H	+H+B	cFSD	+H	+H+B
2005	100	0.02%	0.02%	0.69%	0.01%	0.01%	0.01%
	150	0.00%	0.00%	0.66%	0.33%	0.07%	0.01%
	200	0.00%	0.00%	0.74%	0.36%	0.14%	0.01%
	250	0.00%	0.00%	0.20%	0.11%	0.04%	0.01%
	300	0.00%	0.00%	0.06%	3.47%	1.17%	1.16%
	350	0.00%	0.00%	0.26%	**	1.93%	1.41%
	400	0.00%	0.00%	0.23%	11.68%	0.50%	0.31%

considerably. Table 6.4 also reports the percent by which the value of the best feasible solution found ("LB" in the table) is below the best upper bound found over all methods ("Best UB" in the table). These results indicate that the specialized heuristic significantly improves the value of the feasible solutions found, and that integrating the specialized branching with the heuristic often yields even further improvement in solution quality.

7. Concluding remarks. More computational experiments need to be performed to test the effectiveness of the new formulations in different settings. For example, we tested the case in which the number of possible realizations of the reference random variable, D, is large. The case in which D is small should also be tested since this is likely the case when a stochastic dominance constraint is used to model a collection of risk constraints. It would be particularly interesting to test these formulations for radiation treatment planning models with dose-volume constraints. We expect that when D is small it will be possible to significantly increase the number of possible realizations, N, of the random vector appearing in the constraints. Another setting in which to test the new formulations is in two-stage stochastic programming with stochastic dominance constraints, as has been recently studied in [11, 12],

where they use the previous, less compact formulations for the stochastic dominance constraints.

Finally, it will be interesting to study a Monte Carlo sampling based approximation scheme for problems with stochastic dominance constraints that have more general distributions. Results on sample approximations for probabilistic constraints (e.g., [3, 20, 23]) may be applied to yield approximations for FSD constraints in which the random vector ξ appearing in the constraint may have general distribution. It will be interesting to explore whether the specific structure of the FSD constraint can yield results beyond direct application of the results for probabilistic constraints. Similarly, results on sample approximations for optimization problems with expected value constraints (e.g., [27, 28]) may be applied to yield approximations for second-order dominance constraints.

Acknowledgments. The author is grateful to Shabbir Ahmed for his helpful comments on a draft of this paper. The author also thanks Darinka Dentcheva, Andrzej Ruszczyński, and an anonymous referee for pointing out the connection between the SSD formulations presented and Strassen's theorem.

REFERENCES

[1] E. Balas, *Disjunctive programming: Properties of the convex hull of feasible points*, Discrete Appl. Math., 89 (1998), pp. 3–44.

[2] E. L. M. Beale and J. A. Tomlin, *Special facilities in a general mathematical programming system for non-convex problems using ordered sets of variables*, in Proceedings of the Fifth International Conference on Operational Research, J. Lawrence, ed., London, UK, 1969, Tavistock Publications, pp. 447–454.

[3] G. C. Calafiore and M. C. Campi, *The scenario approach to robust control design*, IEEE Trans. Automat. Control, 51 (2006), pp. 742–753.

[4] D. Dentcheva and A. Ruszczyński, *Optimization with stochastic dominance constraints*, SIAM J. Optim., 14 (2003), pp. 548–566.

[5] D. Dentcheva and A. Ruszczyński, *Optimality and duality theory for stochastic optimization problems with nonlinear dominance constraints*, Math. Program., 99 (2004), pp. 329–350.

[6] D. Dentcheva and A. Ruszczyński, *Semi-infinite probabilistic optimization: First-order stochastic dominance constraints*, Optimization, 53 (2004), pp. 583–601.

[7] D. Dentcheva and A. Ruszczyński, *Portfolio optimization with stochastic dominance constraints*, Journal of Banking & Finance, 30 (2006), pp. 433–451.

[8] C. Fábián, G. Mitra, and D. Roman, *Processing Second-Order Stochastic Dominance Models using Cutting-Plane Representations*, CARISMA Technical report 75, Brunel University, West London, UK, 2008.

[9] J. Forrest, D. de la Nuez, and R. Lougee-Heimer, *CLP User Guide*, 2004.

[10] R. Gollmer, U. Gotzes, F. Neise, and R. Schultz, *Risk Modeling via Stochastic Dominance in Power Systems with Dispersed Generation*, Technical report preprint 651/2007, Department of Mathematics, University of Duisburg-Essen, Duisburg and Essen, Germany, 2007.

[11] R. Gollmer, U. Gotzes, and R. Schultz, *Second-Order Stochastic Dominance Constraints Induced by Mixed-Integer Linear Recourse*, Technical report preprint 644/2007, Department of Mathematics, University of Duisburg-Essen, Duisburg and Essen, Germany, 2007.

[12] R. Gollmer, F. Neise, and R. Schultz, *Stochastic programs with first-order dominance constraints induced by mixed-integer linear recourse*, SIAM J. Optim., 19 (2008), pp. 552–571.

[13] G. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, Cambridge University Press, Cambridge, UK, 1934.

[14] Ilog, *Ilog CPLEX* 9.0 *User's Manual*, 2003.

[15] W. K. Klein Haneveld, *Duality in Stochastic Linear and Dynamic Programming*, Lecture Notes in Econom. and Math. Systems 274, Springer-Verlag, New York, 1986.

[16] W. K. Klein Haneveld and M. H. van der Vlerk, *Integrated chance constraints: Reduced forms and an algorithm*, Comput. Manage. Sci., 3 (2006), pp. 245–269.

[17] T. Kuosmanen, *Efficient diversification according to stochastic dominance criteria*, Man-

age. Sci., 50 (2004), pp. 1390–1406.
[18] E. K. Lee, T. Fox, and I. Crocker, *Integer programming applied to intensity-modulated radiation therapy treatment planning*, Ann. Oper. Res., 119 (2003), pp. 165–181.
[19] J. Luedtke, *Integer Programming Approaches to Some Non-convex and Stochastic Optimization Problems*, Ph.D. thesis, Georgia Institute of Technology, Atlanta, GA, 2007; available online at etd.gatech.edu.
[20] J. Luedtke and S. Ahmed, *A sample approximation approach for optimization with probabilistic constraints*, SIAM J. Optim., 19 (2008), pp. 674–699.
[21] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization*, Academic Press, New York, 1979.
[22] A. Müller and D. Stoyan, *Comparison Methods for Stochastic Models and Risks*, John Wiley & Sons, Ltd., Chichester, UK, 2002.
[23] A. Nemirovski and A. Shapiro, *Scenario approximation of chance constraints*, in Probabilistic and Randomized Methods for Design Under Uncertainty, G. Calafiore and F. Dabbene, eds., Springer, London, 2005, pp. 3–48.
[24] N. Noyan, G. Rudolf, and A. Ruszczyński, *Relaxations of linear programming problems with first order stochastic dominance constraints*, Oper. Res. Lett., 34 (2006), pp. 653–659.
[25] N. Noyan and A. Ruszczyński, *Valid inequalities and restrictions for stochastic programming problems with first order stochastic dominance constraints*, Math. Program., 114 (2008), pp. 249–275.
[26] G. Rudolf and A. Ruszczyński, *Optimization Problems with Second Order Stochastic Dominance Constraints: Duality, Compact Formulations, and Cut Generation methods*, RUTCOR Research Report, Rutgers University, Piscataway, NJ, 2007.
[27] A. Shapiro, *Asymptotic behavior of optimal solutions in stochastic programming*, Math. Oper. Res., 18 (1993), pp. 829–845.
[28] W. Wang and S. Ahmed, *Sample Average Approximation of Expected Value Constrained Stochastic Programs*, preprint, available at www.optimization-online.org, 2007.
[29] G. A. Whitmore and M. C. Findlay, eds., *Stochastic Dominance*, D.C. Heath and Company, Lexington, MA, 1978.

SIAM J. OPTIM.
Vol. 19, No. 3, pp. 1451–1466

THE CONVEX ENVELOPE OF (N–1)-CONVEX FUNCTIONS*

MATTHIAS JACH†, DENNIS MICHAELS†, AND ROBERT WEISMANTEL†

Abstract. The question of determining strong convex underestimators for nonlinear functions is theoretically and practically of major interest. Unfortunately, results along these lines are quite limited as very few general procedures are at hand that can be applied to general classes of functions. In this paper we show how to reduce the question of determining a convex envelope to lower-dimensional optimization problems when the underlying function is indefinite and (n–1)-convex. Our structural result about this reduction technique enables us to give descriptions for the convex envelope of a variety of two-dimensional functions.

Key words. convex envelopes, (n–1)-convex functions, convex relaxations

AMS subject classifications. 52A27, 52A41

DOI. 10.1137/07069359X

1. Introduction. Many approaches for solving mixed-integer nonlinear optimization problems (MINLP) combine local search methods with algorithms to compute global bounds on convex relaxations of underlying feasible regions (e.g., [5, 2, 16, 1, 4, 13, 14]). Typically, convex relaxations are obtained by replacing all nonconvex terms in the original formulation by convex under- and concave overestimators, respectively. In order to derive strong bounds that allow one to verify the quality of known feasible solutions, tight estimators are necessary: the tighter the estimator, the tighter the bound. By definition, the tightest possible convex under- and concave overestimator of a general nonlinear function are called its envelopes. Despite this necessity of having convex and concave envelopes at hand, very little is known in the literature. The limited availability of formulae for the envelopes is due to the fact that the standard representation of envelopes is a nonconvex optimization problem that is intractable, in general. In the special case when a multivariate function is concave in one variable and convex in all the others, Tawarmalani and Sahinidis showed that the nonconvex optimization problem can be tackled using disjunctive programming techniques [11]. To the best of our knowledge, we are not aware of other general techniques in this direction. Indeed, the majority of results about convex envelopes apply to particular functions only. This includes continuously differentiable univariate functions [5], the product terms given by x_1x_2 [5, 2, 4, 10] and $x_1x_2x_3$ [6, 7], and the bivariate function given by x_1/x_2 [16, 11, 12]. The most involved functions for which the convex envelopes are determined analytically are affine transformations of x_1/x_2. The function $f : [\mathbf{l}, \mathbf{u}] \subseteq \mathbf{R}_{\geq 0} \times \mathbf{R}_{>0} \to \mathbf{R}$, $\mathbf{x} \mapsto x_1/x_2$ is convex in x_1 and x_2, but is not convex in both variables simulaneously, i.e., it is indefinite. Our main theorem applies precisely to functions of this type. It allows us to express the value of the convex envelope of such a function at a given point as an (n–1)-dimensional optimization problem. This result can, in turn, be used to determine new formulae for the convex envelopes of generalizations of the function $\mathbf{x} \mapsto x_1x_2$.

*Received by the editors June 4, 2007; accepted for publication (in revised form) July 18, 2008; published electronically December 5, 2008. This work was supported by the Deutsche Forschungsgemeinschaft grant FOR 468.

http://www.siam.org/journals/siopt/19-3/69359.html

†Institut für Mathematische Optimierung, Otto-von-Guericke-Universität Universitätsplatz 2, D-39106 Magdeburg, Germany (mjach@imo.math.uni-magdeburg.de, michaels@imo.math.uni-magdeburg.de, weismantel@imo.math.uni-magdeburg.de).

This paper is structured as follows: We begin by introducing the notation and recall some basic results from convex analysis in section 2. Our main result is given in section 3. Section 4 deals with computational aspects for evaluating the convex envelope for (n–1)-convex functions on a box. Section 5 is devoted to three interesting families of bivariate functions.

2. Preliminaries. Throughout this section we consider twice continuously differentiable functions $f : \mathbf{R}^n \to \mathbf{R}$ over a *convex compact* domain $D \subseteq \mathbf{R}^n$.

DEFINITION 2.1. *Let $D \subseteq \mathbf{R}^n$ be a convex compact subset, and let $f : D \to \mathbf{R}$ be a real-valued function.*

(a) *A function $\eta : \mathbf{R}^n \to \mathbf{R} \cup \{\pm\infty\}$ is called an* underestimator *(*overestimator*) of f on D if $\eta(\mathbf{x}) \le f(\mathbf{x})$ ($\eta(\mathbf{x}) \ge f(\mathbf{x})$) for all $\mathbf{x} \in D$. If η is, in addition, convex (concave) on D, then we call η a convex underestimator (concave overestimator).*

(b) *The tightest convex underestimator of f over D is called the* convex envelope*, denoted by $\mathrm{vex}_D[f]$, while the tightest concave overestimator of f over D is called* concave envelope*, denoted by $\mathrm{cave}_D[f]$. The envelopes are defined pointwise:*

$$\begin{aligned}\mathrm{vex}_D[f](\mathbf{x}) = \max\ \{&\eta(\mathbf{x}) \mid \eta : D \to \mathbf{R} \cup \{\pm\infty\} \text{ with}\\ &\eta(\mathbf{x}) \le f(\mathbf{x}) \text{ for all } \mathbf{x} \in D, \text{ and } \eta \text{ convex}\},\\ \mathrm{cave}_D[f](\mathbf{x}) = \min\ \{&\eta(\mathbf{x}) \mid \eta : D \to \mathbf{R} \cup \{\pm\infty\} \text{ with}\\ &\eta(\mathbf{x}) \ge f(\mathbf{x}) \text{ for all } \mathbf{x} \in D, \text{ and } \eta \text{ concave}\}.\end{aligned}$$

In the following, we recall some basic facts from convex analysis that are required for subsequent sections of this paper.

First, we remark that deriving the envelopes for a function is quite hard, in general. To see this, consider the following representation of the convex envelope that is used in [9]:

$$\mathrm{vex}_D[f](\mathbf{x}) = \min\ \{\mu \mid (\mathbf{x}, \mu) \in \mathrm{conv}(\mathrm{epi}_D[f])\}, \tag{2.1}$$

where $\mathrm{epi}_D[f] := \{(\mathbf{x}, \mu) \in \mathbf{R}^{n+1} \mid \mu \ge f(\mathbf{x}),\ \mathbf{x} \in D\}$ denotes the *epigraph* of f on D.

The representation in formula (2.1) implies that evaluating the convex envelope of a function $f : D \subseteq \mathbf{R}^n \to \mathbf{R}$ at a single point $\mathbf{x} \in D$ requires solving the nonlinear optimization problem

$$\begin{aligned}\mathrm{vex}_D[f](\mathbf{x}) &= \min \mu\\ \text{s. t.} \sum_k \lambda_k \mathbf{x}^{\mathbf{k}} &= \mathbf{x},\\ \sum_k \lambda_k f(\mathbf{x}^{\mathbf{k}}) &= \mu,\\ \sum_k \lambda_k &= 1,\\ \lambda_k &\ge 0 \quad \forall\, k,\\ \mathbf{x}^{\mathbf{k}} &\in D \quad \forall\, k.\end{aligned} \tag{2.2}$$

Problem (2.2) is a highly nonconvex optimization model involving two types of variables: the convex multipliers $\lambda_k \in [0, 1]$ and the variable vectors $\mathbf{x}^{\mathbf{k}}$ representing

suitable points in D. By Caratheodory's theorem about the representation of points in a convex set (cf. Theorem 17.1 in [9]), it suffices to consider convex combinations of at most $(n+1)$ points $(\mathbf{x^k}, f(\mathbf{x^k}))$, $\mathbf{x^k} \in D$. In particular, such sets of points can be chosen to be affinely independent; i.e., to be simplices. In our analysis we are interested in special structured simplices given by the following definition.

DEFINITION 2.2. *Let $f : \mathbf{R}^n \to \mathbf{R}$ be a function restricted to a convex compact domain $D \subseteq \mathbf{R}^n$. For a point $\mathbf{x} \in D$, let $\mathbf{x^k} \in D$ and $\lambda_k \in \mathbf{R}_{\geq 0}$, $k = 1, \ldots, t$, with $\sum_{k=1}^{t} \lambda_k = 1$, be an optimal solution to problem* (2.2) *such that the points $(\mathbf{x^k}, f(\mathbf{x^k}))$ are vertices of a simplex S, and $\lambda_k > 0$ for all $k = 1, \ldots, t$, $t \leq n+1$. Then, for $\mathbf{x} \in D$, S is said to be* minimizing *w.r.t. problem* (2.2).

The next proposition is easy to prove.

PROPOSITION 2.3. *Consider a function $f : \mathbf{R}^n \to \mathbf{R}$ on a convex compact domain D. For a given point $\mathbf{x} \in D$, let S be a minimizing simplex w.r.t. problem* (2.2). *Then, each face $S' \subseteq S$ is a minimizing simplex for all points $\mathbf{x}' \in D$ contained in the relative interior of the projection of S' on D.*

Note that for each point contained in the projection of a simplex S as defined above, there is indeed a face S' whose projection contains the point in its relative interior.

In problem (2.2) we can restrict our attention to those points $(\mathbf{x^k}, f(\mathbf{x^k}))$ which are extreme points of $\mathrm{conv}(\mathrm{epi}_D[f])$. For a continuous function $f : D \subseteq \mathbf{R}^n \to \mathbf{R}$ on a convex compact domain D, let

$$\mathcal{G}_D^{\mathrm{vex}}[f] \;:=\; \{\mathbf{x} \mid (\mathbf{x}, \mathrm{vex}_D[f](\mathbf{x})) \text{ is an extreme point of } \mathrm{conv}(\mathrm{epi}_D[f])\}$$

be the *generating set of the convex envelope of f* on D (cf. [11, 12]). The next observation provides a sufficient condition under which a point $\mathbf{x} \in D$ does not belong to the generating set of its convex envelope.

OBSERVATION 1 (cf. Corollary 5 in [12]). *Let $f : \mathbf{R}^n \to \mathbf{R}$ be restricted to a convex compact subset $D \subseteq \mathbf{R}^n$. If there is a line segment $s \subseteq D$ such that $\mathbf{x}$ is contained in the* relative interior *$ri(s)$ and f is concave over $ri(s)$, then $\mathbf{x} \notin \mathcal{G}_D^{\mathrm{vex}}[f]$.*

Tawarmalani and Sahinidis use this observation in [11] to reduce the complexity of problem (2.2) for those functions

$$f :\; D = [\mathbf{l}, \mathbf{u}] \times [v, w] \;\subseteq\; \mathbf{R}^n \times \mathbf{R} \;\to\; \mathbf{R}, \qquad (\mathbf{x}, z) \;\mapsto\; f(\mathbf{x}, z),$$

which are concave in z, but convex whenever z is fixed. The reduction is possible because

$$\mathcal{G}_D^{\mathrm{vex}}[f] \;\subseteq\; \{(\mathbf{x}, v) \mid \mathbf{x} \in [\mathbf{l}, \mathbf{u}]\} \;\cup\; \{(\mathbf{x}, w) \mid \mathbf{x} \in [\mathbf{l}, \mathbf{u}]\}.$$

That is, the convex hull of the epigraph of f over $[\mathbf{l}, \mathbf{u}] \times [v, w]$ is the convex hull of the union of two convex epigraphs restricted to $z = v$ and $z = w$, respectively. Therefore, in problem (2.2) only one-dimensional simplices, i.e., segments, need to be considered.

Example 1. Consider the function $f : \mathbf{R}^2 \to \mathbf{R}$, $(x_1, x_2) \mapsto \frac{x_1}{x_2}$ on a box $D := [\mathbf{l}, \mathbf{u}] \subseteq \mathbf{R}_{\geq 0} \times \mathbf{R}_{>0}$. A description of its convex envelope has been provided in [11]. To obtain this description, problem (2.2) was reformulated as a univariate minimization problem. An analysis of this optimization problem yields a case distinction with expressions for the individual cases. Using the fact that the set of the concave directions is contained in $(\mathbf{R}^2_{\geq 0}) \cup (\mathbf{R}^2_{\leq 0})$ for every $\mathbf{x} \in D$, the description of

the convex envelope can be summarized in the following formula:

$$\begin{aligned}\operatorname{vex}_D[\tfrac{x_1}{x_2}](\mathbf{x}) = \; & \frac{u_1 - x_1}{u_1 - l_1} \frac{l_1}{\max\left(l_2, \frac{u_2 - x_2}{u_1 - x_1}(l_1 - x_1) + x_2, \frac{x_2\sqrt{l_1}(u_1 - l_1)}{(u_1 - x_1)\sqrt{l_1} + (x_1 - l_1)\sqrt{u_1}}\right)} \\ & + \frac{x_1 - l_1}{u_1 - l_1} \frac{u_1}{\min\left(\frac{x_2 - l_2}{x_1 - l_1}(u_1 - x_1) + x_2, u_2, \frac{x_2\sqrt{u_1}(u_1 - l_1)}{(u_1 - x_1)\sqrt{l_1} + (x_1 - l_1)\sqrt{u_1}}\right)}. \end{aligned} \tag{2.3}$$

From the representation in formula (2.3), one can immediately deduce that the graph of $\operatorname{vex}_D[\frac{x_1}{x_2}]$ is the union of segments. The first and second entries in the max- and min-terms correspond to solutions where one endpoint of the segment is given by a vertex of D, whereas taking the third entry in both terms yields the function first given in Theorem 2 of [15].

Our ability to derive a formula for the convex envelope of the function x_1/x_2 relied on the fact that problem (2.2) in this special case can be reduced to a series of optimization problems in lower dimension, each of which can be solved explicitly.

In the next section we will show that this phenomenon carries over to a much broader family of functions.

DEFINITION 2.4. *Let $f : \mathbf{R}^n \to \mathbf{R}$ be a twice differentiable function.*

(a) *f is said to be (strictly) (n–1)-convex if and only if for all $i \in \{1, \ldots, n\}$ the function $f|_{x_i = \bar{x}_i} : \mathbf{R}^{n-1} \to \mathbf{R}$ is (strictly) convex for each fixed value $\bar{x}_i \in \mathbf{R}$.*

(b) *If f is restricted to a convex domain D, then f is called indefinite (on D) if and only if for each $\mathbf{x} \in D$ the Hessian matrix $\mathcal{H}_f(\mathbf{x})$ is indefinite.*

(c) *The set of all concave directions of f at a point $\mathbf{x} \in D$ is denoted by $\gamma_f(\mathbf{x})$, i.e.,*

$$\gamma_f(\mathbf{x}) \; := \; \{\mathbf{y} \in \mathbf{R}^n \mid \mathbf{y}^\top \mathcal{H}_f(\mathbf{x})\mathbf{y} < 0\}.$$

Example 2.

(a) For integers $n \geq 2$, the family of quadratic functions $f_n : \mathbf{R}^n \to \mathbf{R}$ given by $f_n(x) := x^\top J_n x$, where $J_n \in \mathbf{R}^{n \times n}$ with

$$(J_n) \quad := \quad \begin{pmatrix} 2n-3 & -2 & -2 & \cdots & -2 \\ -2 & 2n-3 & -2 & \cdots & -2 \\ \vdots & & \ddots & & \vdots \\ -2 & \ldots\ldots & -2 & 2n-3 & -2 \\ -2 & \ldots\ldots\ldots & & -2 & 2n-3 \end{pmatrix},$$

is (n–1)-convex and indefinite on $\mathbf{R}^n$.

(b) The bivariate functions $f : \mathbf{R}^2 \to \mathbf{R}$ given by $f(x) := \exp(x_1 x_2)$ and $f(x) := x_1^p x_2^q$ (for $p, q \geq 1$) are 1-convex and indefinite on $\mathbf{R}^2_{\geq 0}$.

3. A structural result. It is the purpose of this section to provide a result which characterizes the convex envelope for further important classes of functions. More precisely, we show that for indefinite (n–1)-convex functions restricted to a box $[\mathbf{l}, \mathbf{u}] \subseteq \mathbf{R}^n$, the minimal value of convex combinations in problem (2.2) is already attained when only segments are chosen. To exclude pathological or trivial cases, we assume in the following that the dimension $n \geq 2$ and the box $[\mathbf{l}, \mathbf{u}] \subseteq \mathbf{R}^n$ is full-dimensional.

THEOREM 3.1. *Let $f : \mathbf{R}^n \to \mathbf{R}$ be an (n–1)-convex and indefinite function on $D := [\mathbf{l}, \mathbf{u}]$, and denote by $\mathcal{F}$ the union of all facets of the box D. Then*

(a) *the concave envelope of f over its domain is polyhedral, i.e., its generating set is finite,*

(b) *the convex envelope of f over its domain is given by the function $g : D \to \mathbf{R}$ which maps each vector $\mathbf{x} \in D$ to*

$$g(\mathbf{x}) := \min\{(1-\lambda)f(\mathbf{x}^1) + \lambda f(\mathbf{x}^2) \mid \mathbf{x}^i \in \mathcal{F},\ i = 1,2, \\ (1-\lambda)\mathbf{x}^1 + \lambda\mathbf{x}^2 = \mathbf{x},\ 0 \le \lambda \le 1\}. \tag{3.1}$$

Part (a) of Theorem 3.1 follows from Observation 1. Our proof for part (b) relies on the fact that concave directions are contained in a fixed pair of orthants for each point $\mathbf{x} \in [\mathbf{l}, \mathbf{u}]$. We show this result first.

LEMMA 3.2. *Let $f : D = [\mathbf{l}, \mathbf{u}] \to \mathbf{R}$ be a twice differentiable function, and let the collection $\{\mathcal{O}_1, \ldots, \mathcal{O}_{2^n}\}$ be the system of open orthants of the space $\mathbf{R}^n$. Then, the function f is (n–1)-convex and indefinite if and only if $\gamma_f(\mathbf{x})$ is nonempty for each $\mathbf{x} \in D$ and there exists an index $i \in \{1, \ldots, 2^n\}$ such that*

$$\forall\, \mathbf{x} \in D: \qquad \gamma_f(\mathbf{x}) \subseteq \mathcal{O}_i \cup (-\mathcal{O}_i).$$

Proof. Sufficiency of the condition is clear: If $\gamma_f(.)$ has the described properties, then f is indefinite and (n–1)-convex on D.

On the other hand, let f be indefinite and (n–1)-convex on D, and let $H = \mathcal{H}_f(\mathbf{x})$ be the Hessian matrix of f at an arbitrary point $\mathbf{x} \in D$.

We claim that H has precisely one negative eigenvalue λ_1. H being indefinite has at least one negative eigenvalue λ_1. If there were another nonpositive eigenvalue λ_2 or if λ_1 were not a simple root, then the space spanned by the corresponding eigenvectors would contain lines of concave directions of f intersecting a hyperplane $x_j = 0$ for some $j = 1, \ldots, n$ in a point different from $\mathbf{0}$. This would contradict the (n–1)-convexity.

Hence, λ_1 is the uniquely determined simple negative eigenvalue of H, to which we can assign the eigenvector $\mathbf{v^1}$, which in turn determines a pair $U(\mathbf{x}) := \mathcal{O}_i \cup (-\mathcal{O}_i)$ for some $i \in \{1, \ldots, 2^n\}$. From the argumentation above it follows that $\gamma_f(\mathbf{x}) \subseteq U(\mathbf{x})$.

We will now show that for all other points $\mathbf{x}' \in D$ the set $\gamma_f(\mathbf{x}')$ is contained in the union $U(\mathbf{x})$ just defined, and assume to the contrary that there exists a point $\mathbf{x}' \in D$ such that $\gamma_f(\mathbf{x}') \not\subseteq U(\mathbf{x})$. Furthermore, we consider the following composition $h : \mathbf{R}^n \to \mathbf{R}$, $\mathbf{y} \mapsto h(\mathbf{y}) := \psi(\phi(\mathbf{y}))$, where ϕ maps a point $\mathbf{y} \in D$ to the eigenvector $\mathbf{v^1}(\mathbf{y})$ associated with the negative eigenvalue of $\mathcal{H}_f(\mathbf{y})$ such that $v_1^1(\mathbf{y}) = 1$, and $\psi : \mathbf{R}^n \to \mathbf{R}$ takes the form

$$\psi : \mathbf{v} \quad \mapsto \quad \begin{cases} \min\{|v_2|, \ldots, |v_n|\} & \text{if } \mathbf{v} \in U(\mathbf{x}), \\ -\min\{|v_2|, \ldots, |v_n|\} & \text{otherwise.} \end{cases}$$

It is straightforward to check that both ϕ and ψ are continuous, and hence, h is continuous, too. For ϕ we point to the continuous dependency of the eigenvalues on the entries of a square matrix (see, e.g., [8]).

If $\mathbf{z}_\lambda := (1-\lambda)\mathbf{x} + \lambda\mathbf{x}'$ with $\lambda \in [0,1]$, then, by assumption, $h(\mathbf{z}_1) > 0$ and $h(\mathbf{z}_0) < 0$, and there is a number $\tilde{\lambda} \in (0,1)$ with $h(\mathbf{z}_{\tilde{\lambda}}) = 0$. This means that the eigenvector corresponding to the negative eigenvalue of $\mathcal{H}_f(\mathbf{z}_{\tilde{\lambda}})$ has a zero component and, hence, is contained in some coordinate plane. This is a contradiction to the property of (n–1)-convexity of f. ∎

We are now ready to prove Theorem 3.1.

Proof of Theorem 3.1. (a) Using Observation 1 and the identity $\mathrm{cave}_D[f] = -\mathrm{vex}_D[-f]$, one can easily check that the generating set of $\mathrm{cave}_D[f]$ is given by the set of the vertices of the domain $[\mathbf{l}, \mathbf{u}]$. Thus, $\mathrm{cave}_D[f]$ is polyhedral.

(b) First, note that the function f is indefinite on D. From Observation 1, it follows that the convex envelope of f is generated by points in $\mathcal{F}$ of D only; i.e., $\mathcal{G}_D^{\mathrm{vex}}[f] \subseteq \mathcal{F}$.

Next, we show the correctness of our claim for strictly $(n–1)$-convex functions. For this, assume that the value of $\mathrm{vex}_D[f]$ at a given point $\bar{\mathbf{x}} \in D$ is attained at a nontrivial convex combination of at least three different points $\mathbf{x^i} \in \mathcal{F}$; i.e.,

$$\mathrm{vex}_D[f](\bar{\mathbf{x}}) = \sum_{1 \le i \le k} \lambda_i f(\mathbf{x}^i), \text{ for some } \lambda_i > 0, \sum_{1 \le i \le k} \lambda_i = 1, \sum_{1 \le i \le k} \lambda_i \mathbf{x}^i = \bar{\mathbf{x}}, \ \mathbf{x}^i \in \mathcal{F},$$

with $k \ge 3$. It suffices to deduce a contradiction for a simplex generated by three vertices since, by Proposition 2.3, each triangular face of a minimizing simplex must be minimizing, too. Without loss of generality, we consider the triangle spanned by $\mathbf{x}^1, \mathbf{x}^2$, and $\mathbf{x}^3$. By symmetry, it suffices to investigate the following two cases:

- $\mathbf{x}^1 = (l_1, x_2^1, x_3^1, \dots)$, $\mathbf{x}^2 = (x_1^2, l_2, x_3^2, \dots)$, $\mathbf{x}^3 = (x_1^3, x_2^3, l_3, \dots)$,
- $\mathbf{x}^1 = (l_1, x_2^1, x_3^1, \dots)$, $\mathbf{x}^2 = (u_1, x_2^2, x_3^2, \dots)$, $\mathbf{x}^3 = (x_1^3, l_2, x_3^3, \dots)$.

Consider the difference vectors $\mathbf{d^{1,2}} := \mathbf{x^1} - \mathbf{x^2}$, $\mathbf{d^{2,3}} := \mathbf{x^2} - \mathbf{x^3}$, and $\mathbf{d^{3,1}} := \mathbf{x^3} - \mathbf{x^1}$. If one of the three difference vectors has a zero component, say $\mathbf{d^{1,2}}$, then the direction $\mathbf{d^{1,2}}$ is contained in a coordinate plane. Thus, f is strictly convex on the segment between $\mathbf{x^1}$ and $\mathbf{x^2}$. Otherwise, it is easy to check that two of the three difference vectors, say $\mathbf{d^{1,2}}$ and $\mathbf{d^{2,3}}$, are contained in two different pairs of open orthants. From Lemma 3.2, it follows that f is strictly convex on at least one of the segments between $\mathbf{x^1}$ and $\mathbf{x^2}$ or between $\mathbf{x^2}$ and $\mathbf{x^3}$.

In both cases there exist two vertices $(\mathbf{x^i}, f(\mathbf{x^i}))$ and $(\mathbf{x^j}, f(\mathbf{x^j}))$ of a minimizing simplex such that on the one hand f is strictly convex on the segment between $\mathbf{x^i}$ and $\mathbf{x^j}$, and on the other hand, the segment spanned by $(\mathbf{x^i}, f(\mathbf{x^i}))$ and $(\mathbf{x^j}, f(\mathbf{x^j}))$ is minimizing, again by Proposition 2.3. This is a contradiction. Thus, we conclude that for strictly $(n–1)$-convex functions there are no minimizing simplices with more than two vertices.

If f is not $(n–1)$-convex in the strict sense, then the sequence of strictly $(n–1)$-convex functions $f_m : \mathbf{R}^n \to \mathbf{R}$, given by $f_m(\mathbf{x}) := f(\mathbf{x}) + \frac{1}{m} \sum_{1 \le i \le n} x_i^2$, uniformly converges to f on D in the $\mathcal{C}^2$ norm. By the compactness of D, the functions f_m are indefinite provided that m is sufficiently large. Let

$$\mathrm{vex}_D[f_m](\mathbf{x}) = (1 - \lambda_m) f_m(\mathbf{x}_m^1) + \lambda_m f_m(\mathbf{x}_m^2), \quad \lambda_m \in [0, 1],$$

then, by the compactness of $\mathcal{F} \times \mathcal{F} \times [0, 1]$, there is a convergent subsequence $\{(\mathbf{x}_{m_k}^1, \mathbf{x}_{m_k}^2, \lambda_{m_k})\}_k$, yielding the statement for f. ∎

It is possible to extend Theorem 3.1 slightly to the case when f is indefinite on the interior of $[\mathbf{l}, \mathbf{u}]$. This extension can be shown by applying standard approximation techniques to the proof of Theorem 3.1. In fact, if we replace the box $[\mathbf{l}, \mathbf{u}]$ in Theorem 3.1 and in its proof by the sequence of boxes $[\mathbf{l} + \frac{1}{n}\mathbf{1}, \mathbf{u} - \frac{1}{n}\mathbf{1}]_n$, where $\mathbf{1} = (1, \dots, 1)^\top$, then we obtain the desired extension. Due to this observation we will use the expression *indefinite on* $[\mathbf{l}, \mathbf{u}]$ also in this more general sense.

4. Computing minimizing segments. In this section we will show that determining the value of the convex envelope of an (n–1)-convex indefinite function $f : [\mathbf{l}, \mathbf{u}] \to \mathbf{R}$ at a given point $\mathbf{x}$ is computationally tractable.

We first remark that if the given point $\mathbf{x}$ is contained in a facet of $[\mathbf{l}, \mathbf{u}]$, i.e., $\mathbf{x} \in \mathcal{F}$ (cf. Theorem 3.1), then, since f is convex on each facet, $\mathrm{vex}_{[\mathbf{l},\mathbf{u}]}[f](\mathbf{x}) = f(\mathbf{x})$.

In the following, let $\mathbf{x} \in [\mathbf{l}, \mathbf{u}] \setminus \mathcal{F}$ be given. By Theorem 3.1, we have to find a segment through $\mathbf{x}$ which has endpoints in $\mathcal{F}$ and provides an optimal solution for problem (2.2). We can classify the set of all segments through $\mathbf{x}$ into subsets where each subset consists of all segments connecting the same two facets of $[\mathbf{l}, \mathbf{u}]$.

The idea is to identify a minimizing segment in each class of segments and to determine the overall minimizing segment afterwards. From both a geometric and a computational point of view we distinguish the following two cases: subsets of segments connecting opposite facets of $[\mathbf{l}, \mathbf{u}]$ (*parallel case*) and subsets of segments connecting adjacent facets of $[\mathbf{l}, \mathbf{u}]$ (*nonparallel case*).

First, we consider the parallel case. Here, computing minimizing segments is similar to the case described in [11] where $f : \mathbf{R}^n \to \mathbf{R}$ is assumed to be convex in the first (n–1) variables, but concave in the last variable.

The proof is based on a reformulation technique following the method used in [11].

Observation 2 (parallel case). *Let $f : [\mathbf{l}, \mathbf{u}] \to \mathbf{R}$ be an (n–1)-convex indefinite function. For a given point $\mathbf{x} \in [\mathbf{l}, \mathbf{u}] \setminus \mathcal{F}$, computing a minimizing segment with endpoints in parallel facets gives rise to a convex optimization problem.*

Proof. Let $\mathbf{x} \in [\mathbf{l}, \mathbf{u}] \setminus \mathcal{F}$ be given. Without loss of generality, we consider the set of all segments through $\mathbf{x}$ that connect the two facets of $[\mathbf{l}, \mathbf{u}]$ given by $x_1 = l_1$ and $x_1 = u_1$. Denote by $\mathbf{a}, \mathbf{b}$ the endpoints of such a segment where $a_1 = l_1$ and $b_1 = u_1$. The corresponding subproblem of problem (2.2) reads as

$$\min\ \lambda f(\mathbf{a}) + (1-\lambda) f(\mathbf{b}) \quad \text{s.t.} \quad \lambda \mathbf{a} + (1-\lambda)\mathbf{b} = \mathbf{x}, \quad \mathbf{l} \le \mathbf{a}, \mathbf{b} \le \mathbf{u}. \tag{4.1}$$

The condition $x_1 = \lambda a_1 + (1-\lambda) b_1 = \lambda l_1 + (1-\lambda) u_1$ implies that $\lambda = (u_1 - x_1)/(u_1 - l_1) \in (0, 1)$. Replacing the variable λ in problem (4.1) by the expression $(u_1 - x_1)/(u_1 - l_1)$ results in a convex objective function which is minimized over linear constraints. ∎

If f is assumed to be strictly (n–1)-convex, then an optimal solution to (4.1) of Observation 2 is easily seen to be uniquely determined.

Next, we consider the nonparallel case. As we will see, in this case the underlying optimization problem can be analyzed elegantly using the notion of unimodal functions.

Definition 4.1. *A differentiable function $f : D \to \mathbf{R}$, D convex, is said to be* unimodal *if there are no critical points other than global minima and if the set of all global minima of f on D is convex.*

In order to prepare the general analysis, we first discuss a convex formulation for the minimization problem in the bivariate case.

Lemma 4.2. *Let $f : [\mathbf{l}, \mathbf{u}] \subseteq \mathbf{R}^2 \to \mathbf{R}$ be a 1-convex indefinite function. For a given point $\mathbf{x} \in [\mathbf{l}, \mathbf{u}] \setminus \mathcal{F}$, computing a minimizing segment with endpoints in adjacent facets can be stated as a minimization problem of a unimodal function.*

Proof. Without loss of generality, we consider a 1-convex function on a box $[\mathbf{l}, \mathbf{u}]$, and assume that a given point $\mathbf{x}$ in the interior of $[\mathbf{l}, \mathbf{u}]$ is contained in a segment with endpoints (y, l_2) and (l_1, z). From the conditions $x_1 = \lambda y + (1-\lambda) l_1$ and $x_2 = \lambda l_2 + (1-\lambda) z$, we can deduce the following formulation of the problem of finding

a minimizing segment:

$$\begin{aligned} \min \quad & \phi(\lambda) := \lambda f\left(\frac{1}{\lambda}(x_1-(1-\lambda)l_1), l_2\right) + (1-\lambda) f\left(l_1, \frac{1}{1-\lambda}(x_2-\lambda l_2)\right) \\ \text{s.t.} \quad & \frac{1}{\lambda}(x_1-(1-\lambda)l_1) \in [l_1,u_1],\ \frac{1}{1-\lambda}(x_2-\lambda l_2) \in [l_2,u_2],\ 0<\lambda<1. \end{aligned} \tag{4.2}$$

The function ϕ is convex on the feasible set of (4.2), since the second derivative satisfies

$$\phi''(\lambda) = \frac{(x_1-l_1)^2}{\lambda^3}\frac{\partial^2 f}{\partial x_1^2}\left(\frac{x_1-(1-\lambda)l_1}{\lambda}, l_2\right) + \frac{(x_2-l_2)^2}{(1-\lambda)^3}\frac{\partial^2 f}{\partial x_2^2}\left(l_1, \frac{x_2-\lambda l_2}{1-\lambda}\right) \geq 0.$$

Considering the parameter λ as a function of y via $\lambda = \frac{x_1-l_1}{y-l_1}$, we obtain

$$\frac{\mathrm{d}\phi}{\mathrm{d}y}(\lambda(y)) = -\frac{\mathrm{d}\phi}{\mathrm{d}\lambda}(\lambda)\frac{x_1-l_1}{(y-l_1)^2},$$

and deduce that the parameter λ and the corresponding value of y minimize the segment value simultaneously. Furthermore, we see that except for global minima, the derivative $\frac{\mathrm{d}\phi}{\mathrm{d}y}$ is not equal to 0, and that the segment value between two global minima remains constant, showing the convexity. ∎

We notice that, if f is strictly 1-convex, then its second partial derivatives $\frac{\partial^2 f}{\partial x_1^2}$ and $\frac{\partial^2 f}{\partial x_2^2}$ do not vanish on any nondegenerate interval. Then the same is true for $\phi''(\lambda)$; i.e., ϕ is strictly convex, and the minimal value is attained at a uniquely determined solution.

In the general $(n$–$1)$-convex case, we consider the problem of finding a minimizing segment with endpoints in adjacent facets $F_1,\ F_2$ for a given point $\mathbf{x} \in [\mathbf{l},\mathbf{u}] \setminus \mathcal{F}$. If we let a point $\mathbf{y} \in F_1$ vary as one endpoint of the segment such that the second one $\mathbf{z} \in F_2$ is uniquely determined, then the minimization problem can be stated as

$$\begin{aligned} \min \quad & \phi(\mathbf{y}) := \lambda(\mathbf{y}) f(\mathbf{y}) + (1-\lambda(\mathbf{y})) f(\mathbf{z}(\mathbf{y})) \\ \text{s.t.} \quad & \mathbf{y} \in F_1, \mathbf{z}(\mathbf{y}) \in F_2, \\ & \mathbf{x} = \lambda(\mathbf{y})\mathbf{y} + (1-\lambda(\mathbf{y}))\mathbf{z}(\mathbf{y}). \end{aligned} \tag{4.3}$$

The set of feasible vectors $\mathbf{y}$ is the intersection of the face F_1 and the pointed cone with vertex $\mathbf{x}$ and directions $\{\mathbf{x}-\mathbf{v} \mid \mathbf{v} \text{ is a vertex of } F_2\}$, and hence a convex set.

We are now prepared to prove the following lemma.

LEMMA 4.3 (nonparallel case). *Let $f : [\mathbf{l},\mathbf{u}] \to \mathbf{R}$ be an $(n$–$1)$-convex, indefinite function. The function ϕ defined in* (4.3) *is unimodal on its domain.*

Proof. First, we show that all critical points are global minima. Let $\mathbf{y^0}$ be a critical point and $\mathbf{y^1}$ be another feasible vector in (4.3). If $\lambda(\mathbf{y^0}) = \lambda(\mathbf{y^1})$, then ϕ is convex on the segment $[\mathbf{y^0},\mathbf{y^1}]$ according to Observation 2; therefore, $\phi(\mathbf{y^0}) \leq \phi(\mathbf{y^1})$. In the case of $\lambda(\mathbf{y^0}) \neq \lambda(\mathbf{y^1})$, we consider ϕ on the one-dimensional segment $[\mathbf{y^0},\mathbf{y^1}]$, which, together with the segment $[\mathbf{z}(\mathbf{y^0}),\mathbf{z}(\mathbf{y^1})]$, can be embedded in $\mathbf{R}^2$ yielding a univariate function $\hat{\phi}(y) := \lambda(y)f(y)+(1-\lambda(y))f(z(y))$ on the segment $[y_0,y_1]$, where y_0 and y_1 correspond to $\mathbf{y_0}$ and $\mathbf{y_1}$, respectively. From the discussion of the bivariate case it follows that $\hat{\phi}(y_0) \leq \hat{\phi}(y_1)$ and hence $\phi(\mathbf{y^0}) \leq \phi(\mathbf{y^1})$.

Next, let $\mathbf{y}^0$ and $\mathbf{y}^1$ be two global minima of ϕ. Performing the case distinction $\lambda(\mathbf{y^0}) = \lambda(\mathbf{y^1})$ and $\lambda(\mathbf{y^0}) \neq \lambda(\mathbf{y^1})$ as outlined before allows us to draw the conclusion that the set of global minima of ϕ is convex. ∎

Observation 2 and Lemma 4.3 show that the value of the convex envelope of an (n–1)-convex indefinite function restricted to a box at a given point can be computed using standard numerical solvers. Moreover, once the endpoints $(\mathbf{y}, f(\mathbf{y}))$, $(\mathbf{z}, f(\mathbf{z}))$ of a minimizing segment for $\mathbf{y} \in [\mathbf{l}, \mathbf{u}]$ have been found, the intersection of the subdifferentials of f at $\mathbf{y}$ and $\mathbf{z}$ when f is restricted to the corresponding facets of $[\mathbf{l}, \mathbf{u}]$ yield feasible normals for any supporting hyperplane at $(\mathbf{x}, \mathrm{vex}_{[\mathbf{l},\mathbf{u}]}[f](\mathbf{x}))$.

In the next section we will describe the convex envelopes of some interesting families of bivariate functions by investigating the reduced nonlinear optimization problem (3.1) of Theorem 3.1.

In doing so, we will implicitly make use of the fact that in these cases for a given point $\mathbf{x} \in [\mathbf{l}, \mathbf{u}]$ a segment can be minimizing only if its direction is contained in $\gamma_f(\mathbf{x})$. By Lemma 3.2, for all $\mathbf{x} \in [\mathbf{l}, \mathbf{u}]$, the sets $\gamma_f(\mathbf{x})$ are contained in one fixed pair of orthants. This allows us to restrict our attention to such segments fitting the underlying orthant pattern only. For the purpose of illustration, consider the bivariate functions on $\mathbf{R}^2_{\geq 0}$ given by $f(\mathbf{x}) := \frac{x_1}{x_2}$ (cf. Example 1) and $f(\mathbf{x}) := \exp(x_1 x_2)$ (cf. Example 3). It is easy to check that

$$\gamma_{\frac{x_1}{x_2}}(\mathbf{x}) \subseteq (\mathbf{R}^2_{\geq 0} \cup \mathbf{R}^2_{\leq 0}) \quad \text{and} \quad \gamma_{\exp(x_1 x_2)}(\mathbf{x}) \subseteq (\mathbf{R}_{\geq 0} \times \mathbf{R}_{\leq 0}) \cup (\mathbf{R}_{\leq 0} \times \mathbf{R}_{\geq 0}).$$

In Figure 4.1(a) and (b), different types of segments through a point $\mathbf{x}$ are shown. The solid segments are possible candidates to minimize since they are contained in the pair of orthants predetermined by $\gamma_f(\mathbf{x})$. The dashed segments do not respect the orthant pattern w.r.t. $\gamma_f(\mathbf{x})$, and can be excluded from further consideration.

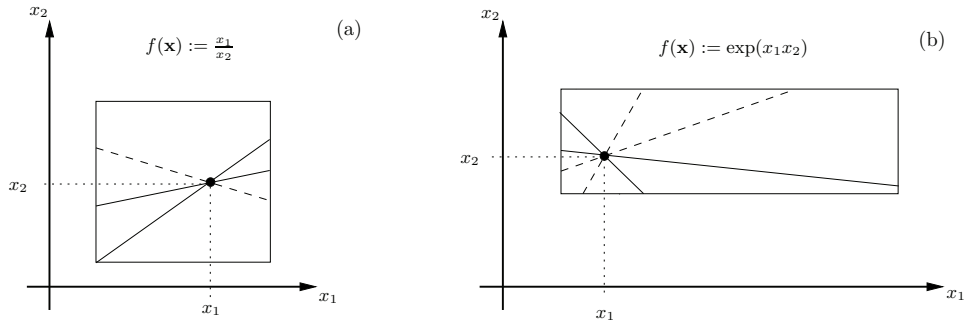


FIG. 4.1. *For the functions given by* $f(\mathbf{x}) := \frac{x_1}{x_2}$ *(cf.* (a) *and Example* 1*) and* $f(\mathbf{x}) := \exp(x_1 x_2)$ *(cf.* (b) *and Example* 3*), the figures show different types of segments connecting points of the facets of the domain* $[\mathbf{l}, \mathbf{u}]$.

5. Application to the bivariate case. Using our results presented in sections 3 and 4, we will investigate three different families of bivariate functions. Each of them can be considered as a generalization of the product term $x_1 x_2$, for which the convex envelope on a box $[\mathbf{l}, \mathbf{u}] \subseteq \mathbf{R}^2$ is well known (cf. [5, 2]):

$$\mathrm{vex}_{[\mathbf{l},\mathbf{u}]}[x_1 x_2](\mathbf{x}) = \max(l_2 x_1 + l_1 x_2 - l_1 l_2, u_2 x_1 + u_1 x_2 - u_1 u_2). \tag{5.1}$$

PROPOSITION 5.1. *Let* $g : \mathbf{R}_{\geq 0} \to \mathbf{R}$ *be a convex and strictly increasing function. Then the function* $f : \mathbf{R}^2_{\geq 0} \to \mathbf{R}$, $\mathbf{x} \mapsto g(x_1 x_2)$, *is* 1*-convex and indefinite. If* f *is*

restricted to a box $[\mathbf{l},\mathbf{u}] \subseteq \mathbf{R}^2_{\geq 0}$*, then on the subset* $R_1 \cup R_2 \subseteq [\mathbf{l},\mathbf{u}]$*, given by*

$$\begin{aligned} R_1 &:= \{(x_1,x_2) \in D \mid l_i x_j \leq -l_j(x_i - l_i) + l_i u_j\}, \\ R_2 &:= \{(x_1,x_2) \in D \mid u_i x_j \geq -u_j(x_i - u_i) + u_i l_j\}, \end{aligned} \tag{5.2}$$

where $l_i u_j \leq u_i l_j$ *for* $(i,j) \in \{(1,2),(2,1)\}$ *(see Figure* 5.1*), we have*

$$\operatorname{vex}_{[\mathbf{l},\mathbf{u}]}[f](\mathbf{x}) = f(\max(l_2 x_1 + l_1 x_2 - l_1 l_2, u_2 x_1 + u_1 x_2 - u_1 u_2)) =: f^*(\mathbf{x}). \tag{5.3}$$

Moreover, for each point $\mathbf{x} \in R_3 := [\mathbf{l},\mathbf{u}] \setminus (R_1 \cup R_2)$*, the relative interior of the projection of the associated minimizing segment is contained in* R_3*, and the endpoints are contained in parallel facets of* $[\mathbf{l},\mathbf{u}]$*.*

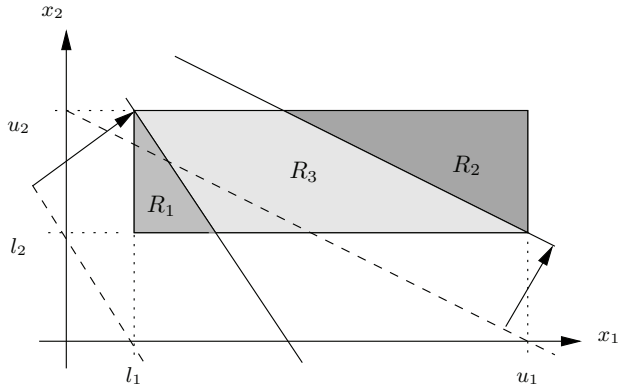


FIG. 5.1. *Subdivision of a box* $[\mathbf{l},\mathbf{u}] \subseteq \mathbf{R}^2_{\geq 0}$ *(with* $l_1 u_2 < u_1 l_2$*) into three regions with respect to different expressions of the convex envelope of functions* $f : \mathbf{R}^2_{\geq 0} \to \mathbf{R}$ *given by* $f(\mathbf{x}) := g(x_1 x_2)$*, where* $g : \mathbf{R}_{\geq 0} \to \mathbf{R}$ *is convex and strictly increasing:* $R_1 := \{\mathbf{x} \in [\mathbf{l},\mathbf{u}] \mid l_1 x_2 \leq -l_2(x_1 - l_1) + l_1 u_2\}$*,* $R_2 := \{\mathbf{x} \in [\mathbf{l},\mathbf{u}] \mid u_1 x_2 \geq -u_2(x_1 - u_1) + u_1 l_2\}$*, and* $R_3 := [\mathbf{l},\mathbf{u}] \setminus (R_1 \cup R_2)$*.*

Proof. A straightforward analysis of the Hessian matrix of f immediately shows the claim of 1-convexity and indefiniteness.

To prove the properties of $f^* : \mathbf{R}^2 \to \mathbf{R}$ as defined in formula (5.3), we first assume that $l_1 u_2 = u_1 l_2$ (i.e., $[\mathbf{l},\mathbf{u}] = R_1 \cup R_2$). Observe that f^* is the composition of the univariate increasing convex function g and the convex envelope of the product term $x_1 x_2$ on $[\mathbf{l},\mathbf{u}]$ (cf. formula (5.1)), and hence a valid convex underestimator of f (cf. [5]). On the other hand, the graph of f^* consists of two families of segments given by the endpoints

$$\begin{aligned} &\big(l_1 + \lambda(u_1 - l_1), l_2,\ f(l_1 + \lambda(u_1 - l_1), l_2)\big), \\ &\qquad \big(l_1, l_2 + \lambda(u_2 - l_2), f(l_1, l_2 + \lambda(u_2 - l_2))\big) \\ &\qquad \text{and} \\ &\big(u_1, l_2 + \mu(u_2 - l_2), f(u_1, l_2 + \mu(u_2 - l_2))\big), \\ &\qquad \big(l_1 + \mu(u_1 - l_1), u_2, f(l_1 + \mu(u_1 - l_1), u_2)\big), \end{aligned} \tag{5.4}$$

where λ, μ range over $[0,1]$. Hence, f^* is also an upper estimate of $\operatorname{vex}_{[\mathbf{l},\mathbf{u}]}[f]$. This proves the correctness of the claim in the special case.

In the general case, we may assume without loss of generality that $l_1 u_2 < u_1 l_2$. Note that R_1 is just a segment if $l_1 = 0$ and $l_2 > 0$. Assume that for a point $\mathbf{x} \in R_1 \cup R_2$

with an associated segment s of a type as described in formula (5.4), there is a segment $\bar{s}$ different from s which yields a lower value at the point $\mathbf{x}$. Then, by increasing l_1 or u_2, or by decreasing u_1 or l_2, we can construct a box $[\bar{\mathbf{l}}, \bar{\mathbf{u}}]$ containing both s and $\bar{s}$ such that $[\bar{\mathbf{l}}, \bar{\mathbf{u}}] = \bar{R}_1 \cup \bar{R}_2$ holds (where $\bar{R}_1$ and $\bar{R}_2$ are given by formula (5.2) using the bounds $\bar{l}_1, \bar{l}_2, \bar{u}_1, \bar{u}_2$). This contradicts the minimality of s w.r.t. $[\bar{\mathbf{l}}, \bar{\mathbf{u}}]$.

Now, let $\mathbf{x} \in R_3 := [\mathbf{l}, \mathbf{u}] \setminus (R_1 \cup R_2)$. Assume first that g is strictly convex and assume further that the relative interior of the projection of an associated minimizing segment $\bar{s}$ is not contained in R_3. By Proposition 2.3, $\bar{s}$ is also minimizing for points in the interior of $R_1 \cup R_2$, but different from those given in formula (5.4). Let $\bar{\mathbf{x}} \in R_1 \cup R_2$ be such a point with a minimizing segment s as defined in formula (5.4). Then the convex hull of s and $\bar{s}$ is a facet of the epigraph of $\mathrm{vex}_{[\mathbf{l},\mathbf{u}]}[f]$. This yields a similar contradiction, as in the proof of Theorem 3.1, or is possible only if $l_1 = l_2 = 0$, but then R_3 was empty.

If R_1 is not a segment, then the minimizing segments associated with points $\mathbf{x} \in R_3$ clearly have the stated property. Otherwise, consider $(l_1^n)_n \to l_1 = 0$ (if $l_2 > 0$) or vice versa. Similarly, the result for functions g that are not necessarily 1-convex in the strict sense follows from an approximation procedure similar to the one used in the proof of Theorem 3.1. □

Example 3. Consider the function $f : \mathbf{R}^2 \to \mathbf{R}$, $\mathbf{x} \mapsto \exp(x_1 x_2)$, restricted to a box $[\mathbf{l}, \mathbf{u}] \subseteq \mathbf{R}^2_{\geq 0}$. Then, by Proposition 5.1, the convex envelope of f on $R_1 \cup R_2$ reads as follows:

$$\mathrm{vex}_{[\mathbf{l},\mathbf{u}]}[\exp(x_1 x_2)] \;=\; \exp(\max(l_2 x_1 + l_1 x_2 - l_1 l_2, u_2 x_1 + u_1 x_2 - u_1 u_2)).$$

For the points $\mathbf{x} \in R_3$, when $l_1 u_2 < u_1 l_2$ (cf. Figure 5.1), it suffices to calculate

$$\min\ \lambda f(a_1, l_2) + (1-\lambda) f(b_1, u_2),$$

where $\lambda = \frac{u_2 - x_2}{u_2 - l_2}$, $l_1 \leq b_1 < a_1 \leq u_1$, and $\lambda a_1 + (1-\lambda) b_1 = x_1$ (see Observation 2). In this example, if the x_1-bounds are neglected, then, using standard analysis means, one can compute that the minimum is attained when

$$\frac{b_1 - a_1}{u_2 - l_2} \;=\; \frac{x_1 - \frac{\ln(l_2/u_2)}{u_2 - l_2}}{x_2 - l_2 - u_2}.$$

If b_1 or a_1 do not respect the bounds, then the endpoints of the segment are shifted such that they are contained in the box $[\mathbf{l}, \mathbf{u}]$.

This results in the following expression:

$$\mathrm{vex}_{[\mathbf{l},\mathbf{u}]}[f](\mathbf{x}) \;=\; \lambda(x_2)\ \exp\big(l_2 g_1(x_1, x_2)\big) + \big(1 - \lambda(x_2)\big)\ \exp\big(u_2 g_2(x_1, x_2)\big)$$

for all $\mathbf{x} \in R_3$, where $\lambda(x_2) := \frac{u_2 - x_2}{u_2 - l_2}$ and

$$g_1(x_1, x_2) \;:=\; \min\left\{u_1,\ \frac{x_1 - l_1}{x_2 - u_2}(l_2 - x_2) + x_1,\ \frac{x_1 - \frac{\ln(l_2/u_2)}{u_2 - l_2}}{x_2 - u_2 - l_2}(l_2 - x_2) + x_1\right\},$$

$$g_2(x_1, x_2) \;:=\; \max\left\{\frac{x_1 - u_1}{x_2 - l_2}(u_2 - x_2) + x_1,\ l_1,\ \frac{x_1 - \frac{\ln(l_2/u_2)}{u_2 - l_2}}{x_2 - u_2 - l_2}(u_2 - x_2) + x_1\right\}.$$

Another generalization is presented in the following proposition.

PROPOSITION 5.2. *Let $f : [\mathbf{l}, \mathbf{u}] \to \mathbf{R}$ be an indefinite function such that $\mathbf{x} \mapsto f(\mathbf{x}) = g(x_1)h(x_2)$, and the functions g, h are nonnegative, convex, and strictly increasing. Assuming $\frac{\mathrm{d}g(l_1)}{\mathrm{d}x_1}h(u_2) < \frac{\mathrm{d}g(u_1)}{\mathrm{d}x_1}h(l_2)$, there are points $v_1, w_1 \in (l_1, u_1)$ such that*

$$\frac{\mathrm{d}g(l_1)}{\mathrm{d}x_1}h(u_2) \;=\; \frac{\mathrm{d}g(y_1)}{\mathrm{d}x_1}h(l_2) \quad \textit{and} \quad \frac{\mathrm{d}g(u_1)}{\mathrm{d}x_1}h(l_2) \;=\; \frac{\mathrm{d}g(z_1)}{\mathrm{d}x_1}h(u_2).$$

Moreover, let $R' \subset [\mathbf{l}, \mathbf{u}]$ be the quadrilateral with vertices (v_1, l_2), (u_1, l_2), (w_1, u_2), (l_1, u_2) as shown in Figure 5.2. Then $\mathrm{vex}_{[\mathbf{l},\mathbf{u}]}[f]$ restricted to R' is given by the union of segments joining points $(y_1, l_2, f(y_1, l_2))$ and $(z_1, u_2, f(z_1, u_2))$, where $v_1 \le y_1 \le u_1$, $l_1 \le z_1 \le w_1$, and $\frac{\mathrm{d}g(y_1)}{\mathrm{d}x_1}h(l_2) = \frac{\mathrm{d}g(z_1)}{\mathrm{d}x_1}h(u_2)$.

For each point $\mathbf{x} \in [\mathbf{l}, \mathbf{u}] \setminus R'$, the relative interior of the projection of the associated minimizing segment is contained in $[\mathbf{l}, \mathbf{u}] \setminus R'$, where the endpoints are contained in orthogonal facets of $[\mathbf{l}, \mathbf{u}]$.

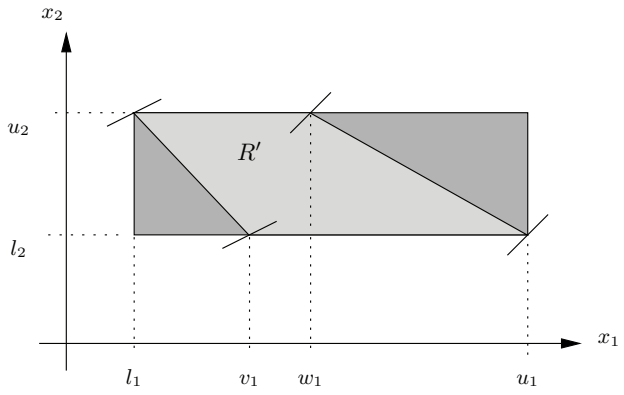


FIG. 5.2. *Subdivision of a box $[\mathbf{l}, \mathbf{u}] \subseteq \mathbf{R}^2_{\ge 0}$ into regions R' and $[\mathbf{l}, \mathbf{u}] \setminus R'$ with respect to different expressions of the convex envelope of functions $f : \mathbf{R}^2_{\ge 0} \to \mathbf{R}$ given by $f(\mathbf{x}) := g(x_1)h(x_2)$, where $g, h : \mathbf{R}_{\ge 0} \to \mathbf{R}$ are convex and strictly increasing. Region R' is a quadrilateral with vertices (l_1, u_2), (u_1, l_2), (v_1, l_2), and (w_1, u_2), where the two pairs of points (l_1, u_2), (v_1, l_2), and (u_1, l_2), (w_1, u_2), respectively, have identical slopes in the direction of the x_1-variable.*

Proof. The first statement follows directly from the properties of f.

Next, let $\mathbf{x} \in R'$, and assume that f is strictly 1-convex on $[\mathbf{l}, \mathbf{u}]$. Then there exist points $(y_1, l_2), (z_1, u_2) \in R'$ with identical slopes

$$\frac{\mathrm{d}g(y_1)}{\mathrm{d}x_1}h(l_2) \;=\; \frac{\mathrm{d}g(z_1)}{\mathrm{d}x_1}h(u_2) \;=:\; m$$

and such that $\mathbf{x}$ is contained in the segment spanned by $(y_1, l_2), (z_1, u_2)$.

We will show that the plane H given by the two tangents at the points $(y_1, l_2, f(y_1, l_2))$ and $(z_1, u_2, f(z_1, u_2))$ with slope m is a valid underestimator of the graph of f on $[\mathbf{l}, \mathbf{u}]$.

Assume to the contrary that there is a point $\hat{\mathbf{x}} \in [\mathbf{l}, \mathbf{u}]$ whose function value of f is smaller than the value given by the plane H. This is equivalent to the fact that there is a point $\tilde{\mathbf{x}}$ in the interior of $[\mathbf{l}, \mathbf{u}]$ with $m = \frac{\mathrm{d}g(\tilde{x_1})}{\mathrm{d}x_1}h(\tilde{x_2})$ whose function value is strictly overestimated by H. We can further assume that the violation among these points is maximal for $\tilde{\mathbf{x}}$.

Now consider the plane H' containing the point $(\tilde{\mathbf{x}}, f(\tilde{\mathbf{x}}))$ obtained from H by shifting it downwards in the direction of the f-axis. Then H' is a valid underestimator of f. By the construction, there is no point $\mathbf{t}$ contained in a facet of $[\mathbf{l}, \mathbf{u}]$ such that $(\mathbf{t}, f(\mathbf{t}))$ is contained in H'. Hence, $(\tilde{\mathbf{x}}, f(\tilde{\mathbf{x}}))$ can be separated by a plane from the graph of f restricted to the facets of $[\mathbf{l}, \mathbf{u}]$. But this contradicts the property that $(\tilde{\mathbf{x}}, f(\tilde{\mathbf{x}}))$ is contained in the convex hull of the graph of f restricted to the facets of $[\mathbf{l}, \mathbf{u}]$. Thus, the plane H is indeed valid for the graph of f proving that the segment given by $(y_1, l_2, f(y_1, l_2))$ and $(z_1, u_2, f(z_1, u_2))$ is minimizing.

The statement for points $\mathbf{x} \in [\mathbf{l}, \mathbf{u}] \setminus R'$ and the case when f is not necessarily strictly 1-convex can be proven with arguments similar to the proof of Proposition 5.1. □

Example 4. For numbers $p, q > 1$, consider the function $f : \mathbf{R}^2 \to \mathbf{R}$, $\mathbf{x} \mapsto x_1^p x_2^q$, on $[\mathbf{l}, \mathbf{u}] \subseteq \mathbf{R}^2_{\geq 0}$. If $pl_1^{p-1}u_2^q < pu_1^{p-1}l_2^q$, then the minimizing segments on R' as defined in Proposition 5.2 (see also Figure 5.2) are given by

$$(y_1, y_2) := \left(\frac{x_1}{\lambda + \mu - \lambda\mu}, l_2 \right) \quad \textit{and} \quad (z_1, z_2) := \left(\frac{\mu x_1}{\lambda + \mu - \lambda\mu}, u_2 \right),$$

where $\lambda := \frac{u_2 - x_2}{u_2 - l_2}$ and $\mu := \left(\frac{l_2}{u_2}\right)^{\frac{q}{p-1}}$.

If $p = q$, then Proposition 5.1 can also be applied to obtain the description for the regions R_1 and R_2 defined in formula (5.2). (Note that by Proposition 5.1, R_1 and R_2 do not overlap with R'.) For points in $[\mathbf{l}, \mathbf{u}] \setminus (R' \cup R_1 \cup R_2)$, the projections of the minimizing segments have a vertex of $[\mathbf{l}, \mathbf{u}]$ as an endpoint.

Next, we focus on the convex and concave envelopes of an arbitrary bivariate quadratic polynomial. We remark that Anstreicher and Burer have recently given a semidefinite representation for the envelopes of this class of functions (see [3]). However, our results developed in sections 3 and 4 enable us to derive explicit formulae via a case distinction.

Example 5. Consider an arbitrary bivariate quadratic polynomial $f : [\mathbf{l}, \mathbf{u}] \subseteq \mathbf{R}^2 \to \mathbf{R}$ defined by $f(\mathbf{x}) := a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2 + b_1x_1 + b_2x_2 + c$, with all coefficients being real.

If $a_{11}^2 + a_{22}^2 = 0$, then the convex and concave envelopes are generated by the four vertices of $[\mathbf{l}, \mathbf{u}]$ and are, hence, polyhedral. In this case, the formulae of the envelopes are easy to compute.

Let $a_{11}^2 + a_{22}^2 > 0$. Setting $a_{21} := a_{12}$, we obtain the Hessian matrix H_f of f as $H_f(\mathbf{x}) = 2(a_{ij})_{1 \leq i,j \leq 2}$. This leads to the following cases:

- *Case* 1. H_f is positive (negative) (semi-)definite. Then f is convex (concave), and the envelopes are trivial to compute. In the convex case,

$$\mathrm{vex}_{[\mathbf{l},\mathbf{u}]}[f](\mathbf{x}) = f(\mathbf{x}),$$

  and $\mathrm{cave}_{[\mathbf{l},\mathbf{u}]}[f](\mathbf{x})$ is generated by the four vertices of $[\mathbf{l}, \mathbf{u}]$.
- *Case* 2. H_f is indefinite, and $a_{11}a_{22} < 0$. Without loss of generality, let $a_{11} > 0 > a_{22}$. Then, f is convex in x_1 and concave in x_2. For determining a formula for the envelopes we can make use of Observation 2 or the reduction technique introduced in [11].
  First, assume that the eigenvector related to the negative eigenvalue of H_f has entries different in sign. An analysis of the reduced underlying optimization problem (4.1) defined in the proof of Observation 2 shows that the envelopes

can be computed as

$$\begin{aligned}\operatorname{vex}_{[\mathbf{l},\mathbf{u}]}[f](\mathbf{x}) &= \lambda f(y_1,l_2)+(1-\lambda)f(z_1,u_2),\\ \operatorname{cave}_{[\mathbf{l},\mathbf{u}]}[f](\mathbf{x}) &= \mu f(l_1,y_2)+(1-\mu)f(u_1,z_2),\end{aligned}$$

where $\lambda=\frac{u_2-x_2}{u_2-l_2}$, $\mu=\frac{u_1-x_1}{u_1-l_1}$,

$$\begin{aligned}y_1 &= \min\left(\tfrac{a_{12}}{a_{11}}(1-\lambda)(u_2-l_2)+x_1,\ u_1,\ \tfrac{x_1-l_1}{x_2-u_2}(l_2-x_2)+x_1\right),\\ z_1 &= \max\left(\tfrac{a_{12}}{a_{11}}\lambda(l_2-u_2)+x_1,\ \tfrac{x_1-u_1}{x_2-l_2}(u_2-x_2)+x_1,\ l_1\right),\\ y_2 &= \max\left(\tfrac{a_{12}}{a_{22}}(1-\mu)(u_1-l_1)+x_2,\ l_2,\ \tfrac{u_2-x_2}{u_1-x_1}(l_1-x_1)+x_2\right),\\ z_2 &= \min\left(\tfrac{a_{12}}{a_{22}}\mu(l_1-u_1)+x_2,\ \tfrac{x_2-l_2}{x_1-l_1}(u_1-x_1)+x_2,\ u_2\right).\end{aligned}\tag{5.5}$$

If the eigenvector to the negative eigenvalue of H_f does not have entries different in sign, then we can choose $(1,0)$ and $(0,1)$, respectively, as the eigenvectors associated with the positive and negative eigenvalues of H_f, respectively. In this case $a_{12}=0$ holds, so formula (5.5) can be simplified to

$$y_1=z_1=x_1,\quad y_2=z_2=x_2.$$

If $a_{11}a_{22}=0$ and $a_{11}>a_{22}=0$, then f is strictly convex in x_1 and affine in x_2. Hence, the convex envelope is given as in formula (5.5), whereas the concave envelope is polyhedral.

- *Case* 3. H_f is indefinite and $a_{11}a_{22}>0$. We may assume that $a_{11},a_{22}>0$ (otherwise, consider $-f$ and use $\operatorname{vex}_{[\mathbf{l},\mathbf{u}]}[-f](\mathbf{x})=-\operatorname{cave}_{[\mathbf{l},\mathbf{u}]}[f](\mathbf{x})$). Then f is 1-convex, i.e., the concave envelope is polyhedral. For the convex envelope, we assume that the eigenvector to the negative eigenvalue of H_f has entries different in sign (otherwise, consider $f(x_1,l_2+u_2-x_2)$). Moreover, let

$$\tfrac{\partial f}{\partial x_2}(l_1,u_2)(l_2-u_2)+f(l_1,u_2)\ \geq\ \tfrac{\partial f}{\partial x_1}(u_1,l_2)(l_1-u_1)+f(u_1,l_2).$$

Set $\mathbf{v}:=\left(-\sqrt{a_{22}},\sqrt{a_{11}}\right)$. The box $[\mathbf{l},\mathbf{u}]$ can be separated into three regions using the two lines with direction $\mathbf{v}$ through the corner points (l_1,u_2) and (u_1,l_2), respectively (see Figure 5.3). In the regions D_1 and D_2, the convex envelope is given by segments whose projections to $\mathbf{R}^2$ are parallel to the vector $\mathbf{v}$. Thus, on D_1 and D_2, the convex envelope reads as

$$\begin{aligned}\operatorname{vex}_{[\mathbf{l},\mathbf{u}]}[f](\mathbf{x})\,|_{D_1} &= \lambda f(y_1,l_2)+(1-\lambda)f(l_1,z_2),\\ \operatorname{vex}_{[\mathbf{l},\mathbf{u}]}[f](\mathbf{x})\,|_{D_2} &= \mu f(v_1,u_2)+(1-\mu)f(u_1,w_2),\end{aligned}$$

where $\lambda=\frac{x_1-l_1}{y_1-l_1}$, $\mu=\frac{x_1-u_1}{y_1-u_1}$,

$$\begin{aligned}y_1&=-\sqrt{\tfrac{a_{22}}{a_{11}}}(l_2-x_2)+x_1, & z_2&=-\sqrt{\tfrac{a_{11}}{a_{22}}}(l_1-x_1)+x_2,\\ v_1&=-\sqrt{\tfrac{a_{22}}{a_{11}}}(u_2-x_2)+x_1, & w_2&=-\sqrt{\tfrac{a_{11}}{a_{22}}}(u_1-x_1)+x_2.\end{aligned}$$

Validity is seen by expanding these formulae yielding the expressions

$$\begin{aligned}g(x_1,x_2)&=a_{11}x_1^2+2\sqrt{a_{11}a_{22}}x_1x_2+a_{22}x_2^2+(b_1+2l_2(a_{12}-\sqrt{a_{11}a_{22}}))x_1\\ &\quad+(b_2+2l_1(a_{12}-\sqrt{a_{11}a_{22}}))x_2+c-2l_1l_2(a_{12}-\sqrt{a_{11}a_{22}}),\\ h(x_1,x_2)&=a_{11}x_1^2+2\sqrt{a_{11}a_{22}}x_1x_2+a_{22}x_2^2+(b_1+2u_2(a_{12}-\sqrt{a_{11}a_{22}}))x_1\\ &\quad+(b_2+2u_1(a_{12}-\sqrt{a_{11}a_{22}}))x_2+c-2u_1u_2(a_{12}-\sqrt{a_{11}a_{22}}).\end{aligned}$$

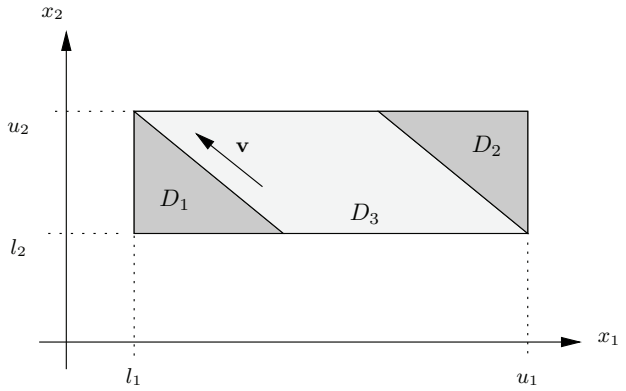


FIG. 5.3. *Subdivision of a box* $[\mathbf{l}, \mathbf{u}] \subseteq \mathbf{R}^2_{\geq 0}$ *into regions* D_1, D_2, *and* D_3 *with respect to different expressions of the convex envelope of a bivariate quadratic polynomial* $a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2 + b_1x_1 + b_2x_2 + c$ *(with* $a_{11} \geq a_{22} > 0$*) that is* 1*-convex and indefinite. The subdivision is defined by the two lines with direction* $\mathbf{v} = (-\sqrt{a_{22}}, \sqrt{a_{11}})$ *through the corner points* (l_1, u_2) *and* (u_1, l_2)*, respectively.*

The functions g and h are clearly convex on their domains D_1 and D_2, respectively. With an argument as in the proof of Proposition 5.1, the validity follows. Moreover, it can be deduced that in the middle region D_3, the envelope is described by the formulae of Case 2.

6. Conclusions. In this paper we showed that evaluating the value of the convex envelope of an (n–1)-convex indefinite function on a box is a computationally tractable task, provided that the number of variables involved is not too large. However, for deriving analytical formulae based on our structural result (cf. Theorem 3.1), it is required to have strong analytical and geometrical properties at hand. We are not aware of practically relevant families of functions with more than three variables for which we could provide formulae for the envelope following the approach outlined in Examples 1–5. Indeed, this lack of geometric information limits our ability to turn the knowledge of Theorem 3.1 into an analytical formula. Still, Observation 2 and Lemma 4.3 may be applied and can be utilized by a numerical solver for evaluating the value of the convex envelope at a given point and a normal vector of a supporting hyperplane.

Acknowledgment. We greatly appreciate the suggestions of two anonymous referees.

REFERENCES

[1] C. S. ADJIMAN, S. DALLWIG, C. A. FLOUDAS, AND A. NEUMAIER, *A global optimization method, αBB, for general twice-differentiable constrained NLPs—*I. *Theoretical advanced*, Computers Chem. Engrg., 22 (1998), pp. 1137–1158.

[2] F. A. AL-KHAYYAL AND J. E. FALK, *Jointly constrained biconvex programming*, Math. Oper. Res., 8 (1983), pp. 273–286.

[3] K. ANSTREICHER AND S. BURER, *Computable Representations for Convex Hulls of Low-Dimensional Quadratic Forms*, Technical report, Department of Management Sciences, University of Iowa, Iowa City, Iowa, 2007, available online at http://www.optimization-online.org/DB_FILE/2007/02/1586.pdf.

[4] J. LINDEROTH, *A simplicial branch-and-bound algorithm for solving quadratically constrained quadratic programs*, Math. Program., 103 (2005), pp. 251–282.

[5] G. P. McCormick, *Computability of global solutions to factorable nonconvex programs.* I. *Convex underestimating problems*, Math. Programing, 10 (1976), pp. 147–175.
[6] C. A. Meyer and C. A. Floudas, *Trilinear monomials with positive or negative domains: Facets of the convex and concave envelopes*, in Frontiers in Global Optimization, C. A. Floudas and P. M. Pardalos, eds., Kluwer Academic Publishers, Boston, 2004, pp. 327–352.
[7] C. A. Meyer and C. A. Floudas, *Trilinear monomials with mixed sign domains: Facets of the convex and concave envelopes*, J. Global Optim., 29 (2004), pp. 125–155.
[8] A. Ostrowski, *Über die Stetigkeit von charakteristischen Wurzeln in Abhängigkeit von den Matrizenelementen*, Jber. Deutsch. Math. Verein., 60 (1957), pp. 40–42.
[9] R. T. Rockafellar, *Convex Analysis*, Princeton Landmarks in Mathematics, Princeton University Press, Princeton, NJ, 1970.
[10] H. D. Sherali and A. Alameddine, *An explicit characterization of the convex envelope of a bivariate bilinear function over special polytopes*, Ann. Oper. Res., 25 (1990), pp. 197–209.
[11] M. Tawarmalani and N. V. Sahinidis, *Semidefinite relaxations of fractional programs via novel convexification techniques*, J. Global Optim., 20 (2001), pp. 137–158.
[12] M. Tawarmalani and N. V. Sahinidis, *Convex extensions and envelopes of lower semi-continuous functions*, Math. Program., 93 (2002), pp. 247–263.
[13] M. Tawarmalani and N. V. Sahinidis, *Global optimization of mixed-integer nonlinear programs: A theoretical and computational study*, Math. Program., 99 (2004), pp. 563–591.
[14] M. Tawarmalani and N. V. Sahinidis, *A polyhedral branch-and-cut approach to global optimization*, Math. Program., 103 (2004), pp. 225–249.
[15] J. M. Zamora and I. E. Grossmann, *A global MINLP optimization algorithm for the synthesis of heat exchanger networks with no stream splits*, Computers Chem. Engng., 22 (1998), pp. 367–384.
[16] J. M. Zamora and I. E. Grossmann, *A branch and contract algorithm for problems with concave univariate, bilinear and linear fractional terms*, J. Global Optim., 14 (1999), pp. 217–249.

SIAM J. OPTIM.
Vol. 19, No. 3, pp. 1467–1488

UNIVERSAL CONFIDENCE SETS FOR SOLUTIONS OF OPTIMIZATION PROBLEMS*

SILVIA VOGEL†

Abstract. We consider random approximations to deterministic optimization problems. The objective function and the constraint set can be approximated simultaneously. Relying on concentration-of-measure results we derive universal confidence sets for the constraint set, the optimal value, and the solution set. Special attention is paid to solution sets which are not single-valued. With many statistical estimators being solutions to random optimization problems, the approach can also be employed to derive confidence sets for constrained estimation problems.

Key words. random optimization problems, universal confidence sets, convergence rate, constrained estimation

AMS subject classifications. 90C15, 90C31, 62F25, 62F30

DOI. 10.1137/070680023

1. Introduction.

Random approximations of deterministic or random optimization problems come into play if unknown quantities are replaced with estimates or for numerical reasons. One solves the approximate problem and hopes that the solution is a good surrogate for the solution of the true problem. Hence there is the need for methods that help to evaluate the goodness of the surrogate solution.

Usually the approximating problems can be arranged to a sequence $(P_n)_{n\in N}$ which approximates the true problem in a suitable sense. Often n can be regarded as the size of the sample on which the estimates are based. Qualitative stability results, which make assertions on the (semi-)convergence of the constraint sets, the optimal values, and the solution sets, are available for convergence almost surely, in probability and in distribution, cf. [12], [17], [8], [7], [19], [3], and the references therein. Furthermore, there are quantitative results which estimate the distance between the optimal values and/or solutions sets by suitable probability metrics; see [14] for an overview.

Confidence bounds for optimal values and solution sets provide valuable additional information. In parametric statistics confidence sets are standard tools. They contain the true value of a parameter at least with a prescribed probability and should satisfy some quality criteria, e.g., minimal size. In the traditional way they are derived from the distribution of a suitable point estimator, which often turns out to be the solution to an appropriate optimization problem. However, the exact distribution of the estimator for a given sample size n is available only in rare cases. Therefore, usually the limit distribution is used as a surrogate, i.e., one deals with asymptotic confidence sets. As indicated in [12] and [3], qualitative stability results for convergence in distribution can also be employed to derive asymptotic confidence sets.

In this paper we will consider a general method which provides *for each* n a set which covers the true solution at least with the prescribed probability. Hence we speak of *universal* confidence sets. These universal confidence sets have a structure which resembles that of many confidence sets in statistics. They are suitable neighborhoods of the solutions to the approximate problems. However, in order to derive such sets,

*Received by the editors January 12, 2007; accepted for publication (in revised form) May 14, 2008; published electronically December 17, 2008.
http://www.siam.org/journals/siopt/19-3/68002.html
†Mathematics and Sciences, Ilmenau Technical University, Weimarer Straße 25, 98693 Ilmenau, Germany (Silvia.Vogel@tu-ilmenau.de).

we do not try to obtain full knowledge about the true distribution of some statistic. Instead we rely on uniform concentration-of-measure results and some assumptions about the true model.

We will also give conditions under which the confidence sets do not only cover the true set, but converge to it in an appropriate sense. Since in general the diameters of the neighborhoods will converge to zero with increasing n, universal confidence sets provide a valuable aid for the decision whether the solution to an approximate problem is good enough or should be improved choosing a larger n.

Confidence bounds for constraint sets and optimal values will be treated in a similar way. Confidence sets for the optimal values can help to assess the quality of the solution to an approximate problem and may be of interest also for model selection. Confidence bounds for constraint sets are of independent interest if sets have to be approximated which are defined by inequality constraints. It is obvious that the method also can be employed for the derivation of confidence sets in statistics if the quantity under consideration can be obtained as solution to an optimization problem.

We adopt ideas from the paper [13]. In [13] Pflug shows how results about uniform convergence in probability, supplemented with known convergence rate and tail behavior, together with a growth condition for the true objective function can be used to derive different kinds of confidence sets. Confidence sets in the sense described above are obtained under the additional assumption that the solution set is single-valued. Furthermore, sufficient conditions for the assumed convergence conditions are discussed in [13].

We will pursue the way proposed in [13] farther, take into account also the approximation of the constraint set, and show how one can proceed if the solution set is not single-valued. The results are formulated in a general way, allowing, e.g., for "relaxed" constraint sets and "κ-optimal" solutions.

We will assume that suitable assertions on the (one-sided) uniform convergence in probability of the objective functions and/or the constraint functions with a convergence rate and tail behavior function are available. Furthermore, we assume some knowledge about the true model, such as a growth condition for the objective function.

Emphasis in this paper is on the concepts in a general form. Three simple examples at the end of the paper are in the first instance meant for illustration. Applications of our results to more complex problems require additional sufficient conditions for the convergence conditions and methods to estimate the parameters of the true model. These topics will be discussed elsewhere.

In order to derive confidence sets for each n, we assume full knowledge about the tail behavior function. If this function is not completely known, the proposed approach can still be employed to derive asymptotic confidence sets.

The paper is organized as follows. In section 2 we introduce the mathematical model and show how universal confidence sets can be derived from suitable convergence results. In section 3 and section 4 we prove the needed convergence assertions for the constraint sets, the optimal values, and the solutions sets. Section 5 contains the examples. The first example is to show how one can deal with the uniform convergence assumptions in a simple case. Approximation of a chance constraint is dealt with in the second example. The third example was chosen to demonstrate the applicability of our results in statistics. We will provide universal confidence bounds for quantiles, allowing for distribution functions which are not continuous.

2. Universal confidence sets. Let (E, d) be a complete separable metric space and $[\Omega, \Sigma, P]$ a complete probability space. We assume that a deterministic optimiza-

tion problem

$$(P_0) \quad \min_{x\in\Gamma_0} f_0(x)$$

is approximated by a sequence of random problems

$$(P_n) \quad \min_{x\in\Gamma_n(\omega)} f_n(x,\omega), \quad n\in N.$$

Additionally to (P_n), for a given $\kappa > 0$, we consider so-called κ-relaxations

$$(P_{n,\kappa}) \quad \min_{x\in\Gamma_{n,\kappa}(\omega)} f_{n,\kappa}(x,\omega), \quad n\in N.$$

The relaxed problems offer the possibility to deal with "relaxed" constraint sets, objective functions and/or solution sets, which are accurate only up to a small parameter that depends on n and κ and tends to zero for each κ if $n\to\infty$. Consequently, the approach can be applied, e.g., to constraint sets and functions which are obtained by Monte Carlo methods (cf. [15], [11]), or to methods which use plug-in estimators, and to ε_n-optimal solutions. Moreover, the relaxed problems are crucial in our approach for the derivation of outer approximations for constraint sets and solution set (see section 3).

The following results will be formulated for $(P_{n,\kappa})$. The problem (P_n) is then regarded as a special case of $(P_{n,\kappa})$ with objective functions and constraint sets that do not depend on κ.

Γ_0 is a nonempty closed subset of E, and the function f_0, which maps into the extended reals $\bar{R}^1 := R^1\cup\{-\infty\}\cup\{+\infty\}$, is a lower semicontinuous function. For each $n\in N$ and $\kappa>0$, $\Gamma_{n,\kappa}|\Omega\to 2^E$ is a closed-valued measurable multifunction, and $f_{n,\kappa}|E\times\Omega\to\bar{R}^1$ is a lower semicontinuous random function, which is supposed to be $(\mathcal{B}(E)\otimes\Sigma,\bar{\mathcal{B}}^1)$ measurable. $\mathcal{B}(E)$ denotes the Borel-σ-field of E and $\bar{\mathcal{B}}^1$ the σ-field which is generated by the Borel sigma field $\mathcal{B}^1$ of R^1 and $\{+\infty\}$, $\{-\infty\}$. Furthermore, we assume that all objective functions are (almost surely) proper functions, i.e., functions with values in $(-\infty,+\infty]$ which are not identically ∞.

The measurability conditions imposed here do not have the weakest form. We use them for sake of simplicity. They are satisfied in many applications and guarantee that all functions of ω needed in the following have the necessary measurability properties. Moreover, the lower semicontinuity assumption of the objective functions $f_{n,\kappa}$ can be dropped. Imposing this condition, however, we can omit some technical details in the proofs.

In the following, the optimal values are denoted by Φ. $\Phi_{n,\kappa}(\omega) := \inf_{x\in\Gamma_{n,\kappa}(\omega)} f_{n,\kappa}(x,\omega)$ is the optimal value for the realization $(P_{n,\kappa}(\omega))$ of the approximate problem, while $\Phi_0 := \inf_{x\in\Gamma_0} f_0(x)$ is the optimal value to (P_0). $\Psi_{n,\kappa}(\omega)$ and Ψ_0 denote the corresponding solution sets.

Our main concern will be with the solution sets Ψ_0 and $\Psi_{n,\kappa}$. We aim at proving assertions of the form

$$(1) \qquad \forall\kappa>0: \sup_{n\ge n_0(\kappa)} P\{\omega:\ \Psi_{n,\kappa}(\omega)\setminus U_{\beta_{n,\kappa}}\Psi_0\neq\emptyset\}\le\mathcal{H}(\kappa)$$

and

$$(2) \qquad \forall\kappa>0: \sup_{n\ge n_0(\kappa)} P\{\omega:\ \Psi_0\setminus U_{\beta_{n,\kappa}}\Psi_{n,\kappa}(\omega)\neq\emptyset\}\le\mathcal{H}(\kappa).$$

Here $(\beta_{n,\kappa})_{n\in N}$ is a sequence of nonnegative numbers which tends to zero for each $\kappa > 0$, and $\mathcal{H}|R^+ \to R^+$ is a function with $\lim_{\kappa\to\infty} \mathcal{H}(\kappa) = 0$. $U_\alpha X$ denotes an open neighborhood of the set X with radius α: $U_\alpha X := \{x \in E : \ d(x, X) < \alpha\}$. $\bar{U}_\alpha X$ means its closure.

In stochastic programming, objective functions and constraint functions which are expectations of random functions are of special interest. If the true, but unknown distribution is replaced with the empirical measure, one often obtains a convergence rate of the form $\tilde{\beta}_{n,\kappa} = \frac{\kappa}{\sqrt{n}}$ for these functions. As we shall see later, the rate $\beta_{n,\kappa}$ which occurs in (1) and (2) is a nondecreasing function of $\tilde{\beta}_{n,\kappa}$. Hence the neighborhoods grow with increasing κ and become smaller with increasing n, cf. section 5. Examples of $\mathcal{H}$ are given in section 5 and [13].

It is desirable that the following assumption (C) is satisfied. Let $\kappa_n(\varepsilon) := \max\{\kappa : \beta_{n,\kappa} \le \varepsilon\}$.

(C) For all $\kappa > 0$ $(\beta_{n,\kappa})_{n\in N}$ converges monotonically to zero, and for all $\varepsilon > 0$ $\lim_{n\to\infty} \kappa_n(\varepsilon) \to \infty$ is satisfied.

If (C) is fulfilled, inequality (1) implies the following property of $(\Psi_n)_{n\in N}$:

For each $\varepsilon > 0$ and an arbitrary compact set $K \subset E$ we have for sufficiently large n,

$$P\{\omega : \ (\Psi_n(\omega) \setminus U_\varepsilon \Psi_0) \cap K \neq \emptyset\} \le P\{\omega : \ \Psi_n(\omega) \setminus U_{\beta_{n,\kappa_n(\varepsilon)}} \Psi_0 \neq \emptyset\} \le \mathcal{H}(\kappa_n(\varepsilon)).$$

Hence the condition

$$\forall \varepsilon > 0 \ \forall \text{ compact } K : \ \lim_{n\to\infty} P\{\omega : \ (\Psi_n(\omega) \setminus U_\varepsilon \Psi_0) \cap K \neq \emptyset\} = 0 \tag{3}$$

holds.

Sequences $(\Psi_n)_{n\in N}$ satisfying condition (3) are called inner approximations in probability to Ψ_0 (cf. [8], [20]). Roughly spoken, inner approximations tend to a subset of Ψ_0 in the particular convergence mode under consideration. For the aims of this paper, convergence in probability is the appropriate mode. In a corresponding way one can show that a sequence which fulfills condition (2) for a sequence $(\beta_{n,\kappa})_{n\in N}$ with the above properties is a so-called outer approximation in probability to Ψ_0. Outer approximations tend to cover Ψ_0. A sequence which is an inner and an outer approximation in probability to Ψ_0 is (Kuratowski–Painlevé)-convergent in probability to Ψ_0.

Because of the relationship to inner and outer approximations in probability, we will call a sequence $(\Psi_{n,\kappa})_{n\in N}$ fulfilling relation (1) an *inner approximation in probability to* Ψ_0 *with convergence rate* $\beta_{n,\kappa}$ *and tail behavior function* $\mathcal{H}$ (in short, an *inner* $(\beta_{n,\kappa}, \mathcal{H})$*-approximation*) and a sequence $(\Psi_{n,\kappa})_{n\in N}$ fulfilling (2) an *outer approximation in probability to* Ψ_0 *with convergence rate* $\beta_{n,\kappa}$ *and tail behavior function* $\mathcal{H}$ (in short, an *outer* $(\beta_{n,\kappa}, \mathcal{H})$*-approximation*). Sequences $(\Psi_{n,\kappa})_{n\in N}$ which are inner and outer $(\beta_{n,\kappa}, \mathcal{H})$-approximations in probability to Ψ_0, such that $(\beta_{n,\kappa})_{n\in N}$ satisfies condition (C), are called $(\beta_{n,\kappa}, \mathcal{H})$*-convergent in probability to* Ψ_0. $(\beta_{n,\kappa}, \mathcal{H})$-convergence in probability is closely related to the normalized convergence investigated in [2].

In order to derive *universal confidence sets to the level* ε_0, i.e., a sequence of random sets $(C_n)_{n\in N}$ with the property $\sup_{n\ge n_0} P\{\omega : \ \Psi_0 \setminus C_n(\omega) \neq \emptyset\} \le \varepsilon_0$, we can proceed as follows.

Suppose that an outer $(\beta_{n,\kappa}, \mathcal{H})$-approximation $(\Psi_{n,\kappa})_{n\in N}$ to Ψ_0 is available and choose to $\varepsilon_0 > 0$ a $\kappa_0 > 0$ such that $\mathcal{H}(\kappa_0) \le \varepsilon_0$. The sets

$$C_n := U_{\beta_{n,\kappa_0}} \Psi_{n,\kappa_0} \tag{4}$$

have the desired property. Of course, one is interested in small confidence sets; hence $(\beta_{n,\kappa})_{n\in N}$ should go to zero as fast as possible and $\mathcal{H}(\kappa)$ should converge to zero as fast as possible if κ tends to infinity.

Unfortunately, under reasonable conditions one obtains inner approximations for the solution set only; see, for instance, the following example.

Example E1. Consider the family of minimization problems $\{\tilde{P}_n,\ n \in N_0\}$ with $E = R^1,\ \Gamma_0 = \Gamma_n = [-1,+1],\ f_0(x) \equiv 0$, and $f_n(x) = \min\{|x|, \frac{1}{n}\}$. Then $(f_n)_{n\in N}$ converges uniformly to f_0, but $\Psi_n = \{0\}$ for all $n \in N$, while $\Psi_0 = [-1,+1]$.

Inner approximations will serve our purpose if all approximating problems (P_n) have solutions which are uniformly bounded and the solution set to the problem (P_0) is single valued, because in this case inner approximations are also outer approximations, and we can proceed as above.

What can be done if the solution set to (P_0) is not single valued? Taking into account that we need knowledge about convergence rates anyway, we can exploit this knowledge to determine suitable relaxing sequences $(\rho_{n,\kappa})_{n\in N}$, which tend to zero for each $\kappa > 0$ and consider $\rho_{n,\kappa}$-optimal solutions, denoted by $\Psi^r_{n,\kappa}$.

The problem, that in general only inner approximations can be obtained, is also apparent for the constraint sets. An example will be considered in section 5.

As mentioned, for reasonable confidence sets one would like to have $\lim_{n\to\infty} \beta^{(i)}_{n,\kappa} = 0$ and $\lim_{\kappa\to\infty} \mathcal{H}_i(\kappa) = 0$ for all sequences $(\beta^{(i)}_{n,\kappa})_{n\in N}$ and functions $\mathcal{H}_i$ which occur in the following. These properties are, however, not needed to prove the results in section 3 and section 4. We only assume throughout the paper that the sequences $(\beta^{(i)}_{n,\kappa})_{n\in N}$ belong to the class B of nonincreasing sequences of positive numbers and the functions $\mathcal{H}_i$ belong to the class H of nonincreasing functions which are defined on R^+ and map into R^+.

3. Approximation of the constraint set. In this section we consider constraint sets, which are given by inequality constraints, and their approximations. Results of that kind are, of course, needed if approximation of the constraint set is inherent in the problem under consideration. Moreover, because of the equations

$$\Psi_0 = \{x \in \Gamma_0 :\ f_0(x) - \Phi_0 \le 0\} \text{ and}$$
$$\Psi_n(\omega) = \{x \in \Gamma_n :\ f_n(x,\omega) - \Phi_n(\omega) \le 0\},$$

the statements can be employed to derive assertions on the behavior of the solution sets, regarding the difference between the true objective function and the optimal value as constraint function. As the optimal values for problems with $\rho_{n,\kappa}$-relaxed constraint sets may depend on κ, the resulting "constraint function" $\tilde{g}_{n,\kappa} := f_n - \Phi_{n,\kappa}$ may depend on κ, too. Furthermore, when applying Theorem 1 below to the solution sets, the multifunctions Q_n, which occur in this theorem, will be interpreted as constraint sets; hence we have to allow that they depend on κ, too.

We assume that the feasibility set Γ_0 in (P_0) can be written as

$$\Gamma_0 = \left\{x :\ g^j_0(x) \le 0,\ j \in J\right\} \cap Q_0,$$

where $J = \{1,\dots,j_M\}$ is a finite index set, the functions $g^j_0 | E \to R^1,\ j \in J$, are lower semicontinuous in all points $x \in E$, and Q_0 is a closed nonempty subset of E. Furthermore, we assume that Γ_0 is nonempty.

For each $\kappa > 0$, the set Q_0 is approximated by a sequence $(Q_{n,\kappa})_{n\in N}$ of closed-valued measurable multifunctions, and the functions $g^j_0,\ j \in J$, are approximated by

sequences $(g^j_{n,\kappa})_{n\in N}$ of functions $g^j_{n,\kappa}|E \to R^1,\ j \in J$, which are $(\mathcal{B}(E)\otimes\Sigma,\mathcal{B}^1)$-measurable. Furthermore, we assume that the functions $g_{n,\kappa}(\cdot,\omega)$ are lower semicontinuous for all $\omega\in\Omega$. Also these measurability and semicontinuity properties could be weakened.

Hence the approximate constraint set $\Gamma_{n,\kappa}$ has the form

$$\Gamma_{n,\kappa}(\omega) = \left\{x \in E :\ g^j_{n,\kappa}(x,\omega) \le 0,\ j \in J\right\} \cap Q_{n,\kappa}(\omega).$$

Under our assumptions $\Gamma_{n,\kappa}$ is a closed-valued measurable multifunction.

If we have $Q_{n,\kappa}(\omega) \equiv Q_0 = E$, we will use the denotation $\hat\Gamma_0$ and $\hat\Gamma_{n,\kappa}$, respectively:

$$\hat\Gamma_0 = \left\{x \in E :\ g^j_0(x) \le 0,\ j \in J\right\}$$

and

$$\hat\Gamma_{n,\kappa}(\omega) := \left\{x \in E :\ g^j_{n,\kappa}(x,\omega) \le 0,\ j \in J\right\}.$$

In the following we employ functions $\nu,\ \mu$, and λ. They are assumed to belong to the set Λ of functions $\tilde\lambda|R^1 \to R^1$ which are right-continuous, nondecreasing, nonconstant, and have the property $\tilde\lambda(0) = 0$. By the superscript $^{-1}$ we denote their inverses: $\tilde\lambda^{-1}(y) := \inf\{x \in R :\ \tilde\lambda(x) \ge y\}$.

THEOREM 1 (inner approximation of the constraint set). *Assume that the following conditions are satisfied:*

(CI1) *There exists a function $\mathcal{H}_1 \in H$ and for all $\kappa > 0$ a sequence $(\beta^{(1)}_{n,\kappa})_{n\in N} \in B$ such that*

$$\sup_{n\in N} P\{\omega :\ Q_{n,\kappa}(\omega) \setminus U_{\beta^{(1)}_{n,\kappa}} Q_0 \ne \emptyset\} \le \mathcal{H}_1(\kappa).$$

(CI2) *There exists a function $\mathcal{H}_2 \in H$ and for all $\kappa > 0$ a sequence $(\beta^{(2)}_{n,\kappa})_{n\in N} \in B$ such that*

$$\sup_{j\in J}\ \sup_{n\in N} P\{\omega :\ \inf_{x\in UQ_0\setminus\Gamma_0} (g^j_{n,\kappa}(x,\omega) - g^j_0(x)) \le -\beta^{(2)}_{n,\kappa}\} \le \mathcal{H}_2(\kappa)$$

for a suitable neighborhood UQ_0.

(CI3) *There exists a function $\nu \in \Lambda$ such that for all $\varepsilon > 0$*

$$U_\varepsilon\Gamma_0 \supset U_{\nu(\varepsilon)}Q_0 \cap U_{\nu(\varepsilon)}\hat\Gamma_0.$$

(CI4) *There exists a function $\mu \in \Lambda$ such that for all $\varepsilon > 0$*

$$\forall\, x \in U_\varepsilon Q_0 \setminus U_\varepsilon\hat\Gamma_0\ \ \exists j \in J :\ g^j_0(x) \ge \mu(\varepsilon).$$

Then for all $\kappa > 0$, $\beta^{(3)}_{n,\kappa} = \max\{\nu^{-1}(\beta^{(1)}_{n,\kappa}), \nu^{-1}(\mu^{-1}(\beta^{(2)}_{n,\kappa}))\}$, and $n_0(\kappa) = \min\{l :\ U_{\nu(\beta^{(3)}_{l,\kappa})}Q_0 \subset UQ_0\}$ the relation $\sup_{n\ge n_0(\kappa)} P\{\omega :\ \Gamma_{n,\kappa}(\omega) \setminus U_{\beta^{(3)}_{n,\kappa}}\Gamma_0 \ne \emptyset\} \le \mathcal{H}_1(\kappa) + j_M\mathcal{H}_2(\kappa)$ holds.

Proof. Assume that for given $\kappa > 0,\ n \ge n_0(\kappa)$, and $\omega \in \Omega$ the relation $\Gamma_{n,\kappa}(\omega) \setminus U_{\beta^{(3)}_{n,\kappa}}\Gamma_0 \ne \emptyset$ holds. Then there is $x_{n,\kappa}(\omega) \in \Gamma_{n,\kappa}(\omega)$ which does not belong to $U_{\beta^{(3)}_{n,\kappa}}\Gamma_0$. Because of (CI3) we have $x_{n,\kappa}(\omega) \notin U_{\nu(\beta^{(3)}_{n,\kappa})}Q_0$ or $x_{n,\kappa}(\omega) \in U_{\nu(\beta^{(3)}_{n,\kappa})}Q_0$ and $x_{n,\kappa}(\omega) \notin U_{\nu(\beta^{(3)}_{n,\kappa})}\hat\Gamma_0$.

In the first case we obtain $Q_{n,\kappa} \backslash U_{\nu(\beta^{(3)}_{n,\kappa})} Q_0 \neq \emptyset$. Hence, because of $\nu(\beta^{(3)}_{n,\kappa}) \geq \beta^{(1)}_{n,\kappa}$, $Q_{n,\kappa}(\omega) \setminus U_{\beta^{(1)}_{n,\kappa}} Q_0 \neq \emptyset$, and we can employ (CI1).

In the second case we obtain by (CI4) for at least one $j \in J$, $g_0^j(x_{n,\kappa}(\omega)) \geq \mu(\nu(\beta^{(3)}_{n,\kappa})) \geq \beta^{(2)}_{n,\kappa}$; hence, because of $U_{\nu(\beta^{(3)}_{n,\kappa})} Q_0 \subset UQ_0$, $\inf_{x \in UQ_0 \setminus \Gamma_0} (g^j_{n,\kappa}(x,\omega) - g_0^j(x)) \leq -\beta^{(2)}_{n,\kappa}$. It remains to employ (CI2). □

Condition (CI1) says that the sequence $(Q_{n,\kappa})_{n \in N}$ is an inner $(\beta^{(1)}_{n,\kappa}, \mathcal{H}_1)$-approximation to Q_0. Condition (CI2) claims—roughly speaking—that the functions $g^j_{n,\kappa}$ outside Γ_0 do not take values that are "essentially" less than the value of g_0^j at the same point. It is a relaxed one-sided version of the uniform concentration-of-measure assumption that was used in [13] for the objective functions. Because of its relationship to a lower semicontinuous approximation in probability (cf. [8], [20]), each sequence $(g^j_{n,\kappa})_{n \in N}$ which satisfies condition (CI2) could be called a *lower semicontinuous approximation in probability to* g_0^j *at* $UQ_0 \setminus \Gamma_0$ *with convergence rate* $\beta^{(2)}_{n,\kappa}$ *and tail behavior function* $\mathcal{H}_2$. (CI2) is a crucial assumption in our approach. Unfortunately, so far there are only a few sufficient conditions which are directly applicable; see, for instance, [13] and the examples at the end of the paper. A general approach for the derivation of sufficient conditions which can serve as a bridge to the concentration-of-measure results for sequences of random variables will be provided elsewhere.

The assumptions (CI3) and (CI4) are conditions about the true model. (CI4) requires that at least one constraint function g_0^j grows outside $\hat{\Gamma}_0$ with a certain rate. This rate can sometimes be derived from the model; see section 5. In general cases it will have to be estimated. Replacing μ with estimates, however, changes the overall convergence rate and will be discussed elsewhere. Condition (CI3) is an assumption about the mutual position and the curvature of Q_0 and $\hat{\Gamma}_0$. It can be dispensed with if $Q_{n,\kappa} \equiv Q_0$ and (CI4) is sharpened to (CI4-W); see Corollary 1 below. (CI2) can then be slightly weakened to (CI2-W). (CI4-W) refers to a function $\tilde{\mu}$ which is determined with respect to the distance to Γ_0, not to $\hat{\Gamma}_0$. $\tilde{\mu}$ will in general be different from μ; compare the following example.

Example E2. Let $E = R^2$, $Q_0 = \{(x,y) : \ y = 2x\}$, $j_M = 1$, and $g_0^1(x,y) = x$. Then $\hat{\Gamma}_0 = (-\infty, 0] \times R^1$ and $\Gamma_0 = \{(x,y) : \ x \leq 0, y = 2x\}$. Consequently we obtain for all $\varepsilon > 0$

$$\forall x \in U_\varepsilon Q_0 \setminus U_\varepsilon \hat{\Gamma}_0 : \ g_0^1(x) \geq \varepsilon, \text{ but}$$
$$\forall x \in Q_0 \setminus U_\varepsilon \Gamma_0 : \ g_0^1(x) \geq \frac{\varepsilon}{\sqrt{5}}.$$

COROLLARY 1. *Assume that* $Q_{n,\kappa} \equiv Q_0$ *holds and the following assumptions are satisfied:*

(CI2-W) *There exists a function* $\mathcal{H}_2 \in H$ *and for all* $\kappa > 0$ *a sequence* $(\beta^{(2)}_{n,\kappa})_{n \in N} \in B$ *such that*

$$\sup_{j \in J} \sup_{n \in N} P\{\omega : \inf_{x \in Q_0 \setminus \Gamma_0} (g^j_{n,\kappa}(x,\omega) - g_0^j(x)) \leq -\beta^{(2)}_{n,\kappa}\} \leq \mathcal{H}_2(\kappa).$$

(CI4-W) *There exists a function* $\tilde{\mu} \in \Lambda$ *such that for all* $\varepsilon > 0$

$$\forall \ x \in Q_0 \setminus U_\varepsilon \Gamma_0 \ \ \exists j \in J : \ g_0^j(x) \geq \tilde{\mu}(\varepsilon).$$

Then for all $\kappa > 0$ *and* $\beta^{(3)}_{n,\kappa} = \tilde{\mu}^{-1}(\beta^{(2)}_{n,\kappa})$ *the relation* $\sup_{n \in N} P\{\omega : \ \Gamma_{n,\kappa}(\omega) \setminus U_{\beta^{(3)}_{n,\kappa}} \Gamma_0 \neq \emptyset\} \leq j_M \mathcal{H}_2(\kappa)$ *holds.*

Proof. Assume that for given $\kappa > 0$, $n \in N$, and $\omega \in \Omega$ the relation $\Gamma_{n,\kappa}(\omega) \setminus U_{\beta^{(3)}_{n,\kappa}}\Gamma_0 \neq \emptyset$ holds. Then there is $x_{n,\kappa}(\omega) \in \Gamma_{n,\kappa}(\omega)$ which does not belong to $U_{\beta^{(3)}_{n,\kappa}}\Gamma_0$. Because of $Q_{n,\kappa} \equiv Q_0$, $x_{n,\kappa}(\omega)$ belongs to Q_0, and we can immediately make use of (CI4-W), which gives $g_0^j(x_{n,\kappa}(\omega)) \geq \tilde{\mu}(\beta^{(3)}_{n,\kappa}) \geq \beta^{(2)}_{n,\kappa}$. Consequently $\inf_{x\in Q_0\setminus\Gamma_0}(g^j_{n,\kappa}(x,\omega) - g_0^j(x)) \leq -\beta^{(2)}_{n,\kappa}$. It remains to employ (CI2-W). □

Later on, Corollary 1 can immediately be applied to obtain results about the approximation of the solution set if the constraint set remains fixed.

Remark 1. Sometimes, concentration-of-measure results are formulated for strict inequalities. Therefore we would like to mention that the conclusion of Theorem 1 does not change if condition (CI2) is replaced by the weaker condition (CI2-S) below and at the same time the stronger condition (CI4-S) is imposed instead of (CI4). This holds correspondingly for further assertions.

(CI2-S) There exists a function $\mathcal{H}_2 \in H$ and for all $\kappa > 0$ a sequence $(\beta^{(2)}_{n,\kappa})_{n\in N} \in B$ such that

$$\sup_{j\in J}\sup_{n\in N} P\{\omega : \inf_{x\in UQ_0\setminus\Gamma_0}(g^j_{n,\kappa}(x,\omega) - g_0^j(x)) < -\beta^{(2)}_{n,\kappa}\} \leq \mathcal{H}_2(\kappa)$$

for a suitable neighborhood UQ_0.

(CI4-S) There exists a function $\mu \in \Lambda$ such that for all $\varepsilon > 0$

$$\forall\, x \in U_\varepsilon Q_0 \setminus U_\varepsilon\hat{\Gamma}_0 \;\; \exists j \in J :\; g_0^j(x) > \mu(\varepsilon).$$

If we want to exploit Theorem 1 for solution sets we have to deal with one constraint function only. Therefore we provide a further specialization of Theorem 1, namely for $j_M = 1$ and $g_0^1 =: g_0$, $g^1_{n,\kappa} =: g_{n,\kappa}$. Additionally, in order to give an example for the function μ in (CI4), we replace (CI4) with a special growth condition (Gr-g_0), which is inspired by the growth condition in [13]. (Gr-g_0) specifies μ as $\mu(\varepsilon) = c_1\varepsilon^{\delta_1}$ for all $\varepsilon > 0$ with $U_\varepsilon Q_0 \subset U_{\theta_1}Q_0$.

COROLLARY 2. *Assume that* (CI1), (CI2), (CI3), *and the following condition* (Gr-g_0) *are satisfied.*

(Gr-g_0) *There exist constants* $c_1 > 0$, $\delta_1 > 0$, *and* $\theta_1 > 0$ *such that*

$$\forall x \in \bar{U}_{\theta_1}Q_0 :\; g_0(x) \geq c_1 \cdot d\big(x,\hat{\Gamma}_0\big)^{\delta_1}.$$

Then for all $\kappa > 0$, $\beta^{(3)}_{n,\kappa} = \max\{\nu^{-1}(\beta^{(1)}_{n,\kappa}), \nu^{-1}((\frac{\beta^{(2)}_{n,\kappa}}{c_1})^{\frac{1}{\delta_1}})\}$, *and* $n_0(\kappa) = \min\{l : U_{\beta^{(3)}_{l,\kappa}}Q_0 \subset \bar{U}_{\theta_1}Q_0 \cap UQ_0\}$ *the relation*

$$\sup_{n\geq n_0(\kappa)} P\{\omega :\; \Gamma_{n,\kappa}(\omega) \setminus U_{\beta^{(3)}_{n,\kappa}}\Gamma_0 \neq \emptyset\} \leq \mathcal{H}_1(\kappa) + \mathcal{H}_2(\kappa)$$

holds.

An important reason for considering constraint functions $g^j_{n,\kappa}$ which depend on κ is the intended application of Theorem 1 to the derivation of assertions about the solution set. If one is interested in the constraint set only, the following special case of Theorem 1 will be applicable in many cases. Often one has inequality constraints and perhaps the intersection with a fixed set Q_0. Usually the functions g_n do not depend on κ, and a convergence rate of the form $\beta^{(2)}_{n,\kappa} = \frac{\kappa}{\gamma_n}$ can be assumed. Then instead of (CI2) the following condition can be used and the growth condition can be given with respect to Γ_0.

(CI2') There exists a function $\mathcal{H}_2 \in H$ and a sequence $(\gamma_n)_{n\in N}$ which tends to ∞ such that for all $\kappa > 0$

$$\sup_{n\in N} P\{\omega :\ \gamma_n \left(\inf_{x\in Q_0\setminus\Gamma_0} (g_n(x,\omega) - g_0(x)\right) \le -\kappa\} \le \mathcal{H}_2(\kappa).$$

(Gr-g_0-Γ_0) There exist constants $c_1 > 0$, $\delta_1 > 0$, and $\theta_1 > 0$ such that

$$\forall x \in Q_0 \cap \bar{U}_{\theta_1}\Gamma_0 :\ g_0(x) \ge c_1 \cdot d(x,\Gamma_0)^{\delta_1}.$$

In this case Corollary 2 can be simplified in the following way:

COROLLARY 3. *Assume that $j_M = 1$ and for all $n \in N$ $g_{n,\kappa} \equiv g_n$ and $Q_{n,\kappa} \equiv Q_0$ holds. Additionally, suppose that* (CI2') *and* (Gr-g_0-Γ_0) *are satisfied. Then for all $\kappa > 0$ and $n_0(\kappa) = \min\{l : \gamma_l \ge (\frac{\kappa}{\theta_1})^{\delta_1}\}$ the relation*

$$\sup_{n\ge n_0(\kappa)} P\left\{\omega :\ \Gamma_n(\omega) \setminus U_{\frac{\kappa}{\gamma_n^{\frac{1}{\delta_1}}}}\Gamma_0 \ne \emptyset\right\} \le \mathcal{H}_2\left(c_1\kappa^{\delta_1}\right)$$

holds.

Proof. Let $\tilde{\kappa} := c_1\kappa^{\delta_1}$. With $\beta^{(2)}_{n,\tilde{\kappa}} = \frac{\tilde{\kappa}}{\gamma_n}$ and $\tilde{\mu}(\varepsilon) = c_1\varepsilon^{\delta_1}$ for $0 < \varepsilon \le \theta_1$ we can apply Corollary 1, taking into account that, additionally, $\beta^{(3)}_{n,\tilde{\kappa}} \le \theta_1$ has to be satisfied. We obtain for $\tilde{\kappa}$, $\beta^{(3)}_{n,\tilde{\kappa}} = (\frac{\tilde{\kappa}}{c_1\gamma_n})^{\frac{1}{\delta_1}} = \frac{\kappa}{\gamma_n^{\frac{1}{\delta_1}}}$, and $n_0(\kappa) = \min\{l : \beta^{(3)}_{l,\tilde{\kappa}} \le \theta_1\} = \min\{l : \gamma_l \ge (\frac{\kappa}{\theta_1})^{\delta_1}\}$ the inequality $\sup_{n\ge n_0(\kappa)} P\{\omega :\ \Gamma_n(\omega) \setminus U_{\beta^{(3)}_{n,\tilde{\kappa}}}\Gamma_0 \ne \emptyset\} \le \mathcal{H}_2(\tilde{\kappa})$, which yields the conclusion. □

As mentioned in the introduction, in general, the sequence $(\Gamma_n)_{n\in N}$ approximates a subset of Γ_0 only. In order to obtain outer approximations, additional assumptions have to be imposed. Qualitative stability theory usually assumes that the condition $\Gamma_0 \subset \mathrm{cl}\{x \in Q_0 :\ g_0^j(x) < 0,\ \forall j \in J\}$ is fulfilled where cl denotes the closure. In order to obtain also a convergence rate and a tail behavior function, we impose a "quantified version" of this assumption with a (negative) growth function μ; see (CO3) in Theorem 2. Unfortunately, a condition of that kind is useless if one intends to employ the result for the solution set, because (CO3) can not be satisfied by $\tilde{g}_0 := f_0 - \Phi_0$.

Hence in Theorem 3 we will provide a second approach which uses "relaxed" inequality constraints. The simple principle can be explained using a modified version of Example E1.

Example E3. Let $E = R^1$, $\Gamma_0 = \Gamma_n = [-1,+1]$, $g_0(x) \equiv 0$, and $g_n(x) = \min\{|x|, \frac{1}{n}\}$. Then $(g_n)_{n\in N}$ converges uniformly to g_0. We consider $\Gamma_0 := \{x \in R^1 : g_0(x) \le 0\}$ and $\Gamma_n := \{x \in R^1 : g_n(x) \le 0\}$. Obviously, $\Gamma_n = \{0\}$ for all $n \in N$, while $\Gamma_0 = [-1,+1]$; i.e., only a subset of Γ_0 is approximated. If we, however, consider the sets $\tilde{\Gamma}_n := \{x \in R^1 :\ g_n(x) \le \frac{1}{n}\}$, which are defined by the "relaxed" inequality constraint $g_n(x) \le \frac{1}{n}$, we see that even $\tilde{\Gamma}_n = \Gamma_0$ holds for all $n \in N$.

Similarly, we weaken the "≤ 0" inequality constraints by "$\le \rho_{n,\kappa}$" with suitably chosen sequences $(\rho_{n,\kappa})_{n\in N}$ of positive reals which satisfy $\lim_{n\to\infty} \rho_{n,\kappa} = 0\ \forall \kappa > 0$.

For the formulation of (CO3) we need the "ε-interior" of Γ_0. Let, for a given $\varepsilon > 0$, $CI(\varepsilon) := \Gamma_0 \setminus U_\varepsilon(E \setminus \Gamma_0)$. $\bar{U}_\varepsilon$ denotes the closure of U_ε. The requirement $\Gamma_0 \subset \bar{U}_\varepsilon CI(\varepsilon)$ is needed since we allow for rather general sets Γ_0.

Condition (CO1) below is the "outer" counterpart to (CI1), and condition (CO2) is the "upper semicontinuous" counterpart to (CI2).

THEOREM 2 (outer approximation of the constraint set (CO3)). *Assume that the following conditions are satisfied:*

(CO1) *There exists a function* $\mathcal{H}_1 \in H$ *and for all* $\kappa > 0$ *a sequence* $(\beta^{(1)}_{n,\kappa})_{n\in N} \in B$ *such that*

$$\sup_{n\in N} P\left\{\omega :\ Q_0 \setminus U_{\beta^{(1)}_{n,\kappa}} Q_{n,\kappa}(\omega) \neq \emptyset\right\} \leq \mathcal{H}_1(\kappa).$$

(CO2) *There exists a function* $\mathcal{H}_2 \in H$ *and for all* $\kappa > 0$ *a sequence* $(\beta^{(2)}_{n,\kappa})_{n\in N} \in B$ *such that*

$$\sup_{j\in J}\ \sup_{n\in N} P\left\{\omega :\ \sup_{x\in\Gamma_0}\left(g^j_{n,\kappa}(x,\omega) - g^j_0(x)\right) \geq \beta^{(2)}_{n,\kappa}\right\} \leq \mathcal{H}_2(\kappa).$$

(CO3) *There exist* $\tilde{\varepsilon} > 0$ *and a function* $\mu \in \Lambda$ *such that for all* $0 < \varepsilon \leq \tilde{\varepsilon}$ $\Gamma_0 \subset \bar{U}_\varepsilon CI(\varepsilon)$ *and*

$$\forall x \in CI(\varepsilon)\ \forall j \in J :\ g^j_0(x) \leq -\mu(\varepsilon).$$

Then for all $\kappa > 0$, $\beta^{(3)}_{n,\kappa} = \max\{\beta^{(1)}_{n,\kappa}, \mu^{-1}(2\beta^{(2)}_{n,\kappa})\}$, *and* $n_0(\kappa) = \min\{l : \beta^{(3)}_{l,\kappa} \leq 2\tilde{\varepsilon}\}$ *the relation*

$$\sup_{n\geq n_0(\kappa)} P\left\{\omega :\ \Gamma_0 \setminus \left(U_{\beta^{(3)}_{n,\kappa}} \hat{\Gamma}_{n,\kappa}(\omega) \cap U_{\beta^{(1)}_{n,\kappa}} Q_{n,\kappa}(\omega)\right) \neq \emptyset\right\} \leq \mathcal{H}_1(\kappa) + j_M \mathcal{H}_2(\kappa)$$

holds.

Proof. Assume that for given $\kappa > 0$, $n \geq n_0(\kappa)$, and $\omega \in \Omega$ the relation $\Gamma_0 \setminus (U_{\beta^{(3)}_{n,\kappa}} \hat{\Gamma}_{n,\kappa}(\omega) \cap U_{\beta^{(1)}_{n,\kappa}} Q_{n,\kappa}(\omega)) \neq \emptyset$ holds. Then there is $x_{n,\kappa}(\omega) \in \Gamma_0$ which does not belong to $U_{\beta^{(3)}_{n,\kappa}} \hat{\Gamma}_{n,\kappa}(\omega) \cap U_{\beta^{(1)}_{n,\kappa}} Q_{n,\kappa}(\omega)$.

If $x_{n,\kappa}(\omega) \notin U_{\beta^{(1)}_{n,\kappa}} Q_{n,\kappa}(\omega)$, we can employ (CO1).

Now suppose that $x_{n,\kappa}(\omega) \notin U_{\beta^{(3)}_{n,\kappa}} \hat{\Gamma}_{n,\kappa}(\omega)$. Because of $\beta^{(3)}_{n,\kappa} \leq 2\tilde{\varepsilon}$ and the first part of condition (CO3), we find $\tilde{x}_{n,\kappa}(\omega) \in CI(\frac{\beta^{(3)}_{n,\kappa}}{2})$ with $\tilde{x}_{n,\kappa}(\omega) \notin \hat{\Gamma}_{n,\kappa}(\omega)$; i.e., $g^{j_0}_{n,\kappa}(\tilde{x}_{n,\kappa}(\omega), \omega) > 0$ for at least one $j_0 \in J$. The second part of (CO3) implies $g^{j_0}_0(\tilde{x}_{n,\kappa}(\omega)) \leq -\mu(\frac{\beta^{(3)}_{n,\kappa}}{2}) \leq -\beta^{(2)}_{n,\kappa}$. It remains to employ (CO2). ∎

Unfortunately, the result is not the "symmetric" counterpart to the statement of Theorem 1. In order to obtain an assertion of the form

$$\sup_{n\geq n_0(\kappa)} P\left\{\omega :\ \Gamma_0 \setminus U_{\beta^{(3)}_{n,\kappa}} \Gamma_{n,\kappa}(\omega) \neq \emptyset\right\} \leq \mathcal{H}(\kappa),$$

we would need functions $\nu_{n,\kappa}$ and conditions similar to (CI3) for each n and κ.

Now we consider $\rho_{n,\kappa}$-relaxed inequality constraints. Let

$$\hat{\Gamma}^r_{n,\kappa}(\omega) := \{x \in E :\ g^j_{n,\kappa}(x,\omega) \leq \rho_{n,\kappa},\ j \in J\}$$

and

$$\Gamma^r_{n,\kappa}(\omega) = \hat{\Gamma}^r_{n,\kappa}(\omega) \cap Q_{n,\kappa}(\omega).$$

THEOREM 3 (outer approximation of the constraint set, relaxation). *Assume that* (CO1) *and* (CO2) *are satisfied. Then for all* $\kappa > 0$ *and* $\rho_{n,\kappa} = \beta^{(2)}_{n,\kappa}$ *the relation*

$$\sup_{n\in N} P\left\{\omega :\ \Gamma_0 \setminus \left(\hat{\Gamma}^r_{n,\kappa}(\omega) \cap U_{\beta^{(1)}_{n,\kappa}} Q_{n,\kappa}(\omega)\right) \neq \emptyset\right\} \leq \mathcal{H}_1(\kappa) + j_M \mathcal{H}_2(\kappa)$$

holds.

Proof. Assume that for given $\kappa > 0$, $n \in N$, and $\omega \in \Omega$ the relation $\Gamma_0 \setminus (\hat{\Gamma}^r_{n,\kappa}(\omega) \cap U_{\beta^{(1)}_{n,\kappa}} Q_{n,\kappa}(\omega)) \neq \emptyset$ is fulfilled. Then there is $x_{n,\kappa}(\omega) \in \Gamma_0$ which does not belong to $\hat{\Gamma}^r_{n,\kappa}(\omega) \cap U_{\beta^{(1)}_{n,\kappa}} Q_{n,\kappa}(\omega)$. Hence $g^j_0(x_{n,\kappa}(\omega)) \leq 0 \ \forall j \in J$ and $x_{n,\kappa}(\omega) \in Q_0$, but either $x_{n,\kappa}(\omega) \notin U_{\beta^{(1)}_{n,\kappa}} Q_{n,\kappa}(\omega)$ or $g^j_{n,\kappa}(x_{n,\kappa}(\omega), \omega) > \beta^{(2)}_{n,\kappa} = \rho_{n,\kappa}$ for at least one $j \in J$.

In the first case we obtain $Q_0 \setminus U_{\beta^{(1)}_{n,\kappa}} Q_{n,\kappa}(\omega) \neq \emptyset$ and employ (CO1).

The second case yields $\sup_{x \in \Gamma_0}(g^j_{n,\kappa}(x, \omega) - g^j_0(x)) \geq \beta^{(2)}_{n,\kappa}$ for at least one $j \in J$, and we make use of (CO2). □

Remark 2. A corresponding result holds if $U_{\beta^{(1)}_{n,\kappa}} Q_{n,\kappa}$ is replaced with $Q_{n,\kappa}$ in condition (CO1) and in the assertion. This observation is of importance if we apply the result to the solution set and have to deal with a constraint set which was obtained via relaxation: The constraint set will then play the role of Q in Theorem 3, and an additional enlargement of $Q_{n,\kappa}$ by a neighborhood should be avoided.

Relaxing the constraints means enlarging the approximating sets. Hence the question arises under what conditions $(\Gamma^r_{n,\kappa})_{n \in N}$ is also an inner approximation. Results of that kind will help to assess the quality of an outer approximation. An inspection of the proof to Theorem 1 shows that with $\rho_{n,\kappa} = \beta^{(2)}_{n,\kappa}$ the following statement can be obtained. The "price" for the relaxation is the additional factor 2 in the definition of $\beta^{(2)}_{n,\kappa}$.

THEOREM 4 (inner approximation of the constraint set, relaxation). *Assume that* (CI1), (CI2), (CI3), *and* (CI4) *are satisfied. Then for all* $\kappa > 0$, $\rho_{n,\kappa} = \beta^{(2)}_{n,\kappa}$, $\beta^{(3)}_{n,\kappa} = \max\{\nu^{-1}(\beta^{(1)}_{n,\kappa}), \nu^{-1}(\mu^{-1}(2\beta^{(2)}_{n,\kappa}))\}$, *and* $n_0(\kappa) = \min\{l : U_{\nu(\beta^{(3)}_{l,\kappa})} Q_0 \subset U Q_0\}$ *the relation*

$$\sup_{n \geq n_0(\kappa)} P\left\{\omega : \Gamma^r_{n,\kappa}(\omega) \setminus U_{\beta^{(3)}_{n,\kappa}} \Gamma_0 \neq \emptyset\right\} \leq \mathcal{H}_1(\kappa) + j_M \mathcal{H}_2(\kappa)$$

holds.

The assertions about inner and outer approximations provided so far can be combined to several convergence statements. As an example, we will summarize what we obtain for only one constraint function under a growth condition. In the following we combine conditions like (CI1) and (CO1). The assumption that in both conditions the same function $\mathcal{H}_1$ and the same sequences $(\beta^{(1)}_{n,\kappa})_{n \in N}$ occur is no restriction. If the original functions or sequences are different we can always take the maximum.

THEOREM 5 (approximation of the constraint set, relaxation). *Assume that* $j_M = 1$, (CI3), *and* (Gr-g_0) *are satisfied. Furthermore, suppose that* (CI1) *and* (CO1) *are fulfilled with the same function* $\mathcal{H}_1 \in H$ *and the same sequences* $(\beta^{(1)}_{n,\kappa})_{n \in N} \in B$, *and* (CI1) *and* (CO2) *are fulfilled with the same function* $\mathcal{H}_2 \in H$ *and the same sequences* $(\beta^{(2)}_{n,\kappa})_{n \in N} \in B$.

Then for all $\kappa > 0$, $\rho_{n,\kappa} = \beta^{(2)}_{n,\kappa}$, $\beta^{(3)}_{n,\kappa} = \max\{\beta^{(1)}_{n,\kappa}, \beta^{(2)}_{n,\kappa}, \nu^{-1}(\beta^{(1)}_{n,\kappa}), \nu^{-1}((\frac{2\beta^{(2)}_{n,\kappa}}{c_1})^{\frac{1}{\delta_1}})\}$, *and* $n_0(\kappa) = \min\{l : U_{\nu(\beta^{(3)}_{l,\kappa})} Q_0 \subset U Q_0\}$ *the relation*

$$\sup_{n \geq n_0(\kappa)} P\left\{\omega : \left(\Gamma^r_{n,\kappa}(\omega) \setminus U_{\beta^{(3)}_{n,\kappa}} \Gamma_0\right) \cup \left(\Gamma_0 \setminus \left(\hat{\Gamma}^r_{n,\kappa}(\omega) \cap Q_{n,\kappa}(\omega)\right)\right) \neq \emptyset\right\} \leq 2\mathcal{H}_1(\kappa) + 2\mathcal{H}_2(\kappa)$$

holds.

Proof. We have

$$P\{\omega:\ (\Gamma^r_{n,\kappa}(\omega)\setminus U_{\beta^{(3)}_{n,\kappa}}\Gamma_0)\cup(\Gamma_0\setminus(\hat\Gamma^r_{n,\kappa}(\omega)\cap Q_{n,\kappa}(\omega))\neq\emptyset\}$$
$$\leq P\{\omega:\ \Gamma^r_{n,\kappa}(\omega)\setminus U_{\beta^{(3)}_{n,\kappa}}\Gamma_0\neq\emptyset\}+P\{\omega:\ \Gamma_0\setminus(\hat\Gamma^r_{n,\kappa}(\omega)\cap Q_{n,\kappa}(\omega))\neq\emptyset\}.$$

The assumptions of Theorem 3 and Theorem 4 are satisfied with $\mu(\varepsilon)=c_1\varepsilon^{\delta_1}$ and $\tilde\beta^{(1)}_{n,\kappa}=\tilde\beta^{(2)}_{n,\kappa}=\beta^{(3)}_{n,\kappa}$. □

Finally, for sake of convenience, we give a special case of Theorem 5 which uses assumptions that correspond to those of Corollary 3. The following condition will be imposed:

(CO2') There exist a function $\mathcal{H}_2\in H$ and a sequence $(\gamma_n)_{n\in N}$, which tends to ∞, such that for all $\kappa>0$

$$\sup_{n\in N}P\left\{\omega:\ \gamma_n\sup_{x\in\Gamma_0}(g_n(x,\omega)-g_0(x))\geq\kappa\right\}\leq\mathcal{H}_2(\kappa).$$

Corollary 4. *Assume that $j_M=1$, for all $n\in N$ $g_{n,\kappa}\equiv g_n$ holds, and* (Gr-g_0-Γ_0) *is satisfied. Additionally, suppose that* (CI2') *and* (CO2') *are fulfilled with the same function $\mathcal{H}_2\in H$ and the same sequences $(\beta^{(2)}_{n,\kappa})_{n\in N}\in B$. Then for all $\kappa>0$, $\Gamma_0=\{x\in Q_0:\ g_0(x)\leq 0\}$, $\Gamma^r_{n,\kappa}(\omega)=\{x\in Q_0:\ g_n(x,\omega))\leq\frac{\kappa}{\sqrt{n}}\}$, $\beta^{(3)}_{n,\kappa}=(\frac{2\kappa}{c_1\gamma_n})^{\frac{1}{\delta_1}}$, and $n_0(\kappa)=\min\{l:\gamma_l\geq\frac{2\kappa}{c_1\theta_1^{\delta_1}}\}$ the relation*

$$\sup_{n\geq n_0(\kappa)}P\left\{\omega:\ \left(\Gamma^r_{n,\kappa}(\omega)\setminus U_{\beta^{(3)}_{n,\kappa}}\Gamma_0\right)\cup\left(\Gamma_0\setminus\Gamma^r_{n,\kappa}(\omega)\right)\neq\emptyset\right\}\leq 2\mathcal{H}_2(\kappa)$$

holds.

Proof. We have

$$P\left\{\omega:\ \left(\Gamma^r_{n,\kappa}(\omega)\setminus U_{\beta^{(3)}_{n,\kappa}}\Gamma_0\right)\cup\left(\Gamma_0\setminus\Gamma^r_{n,\kappa}(\omega)\right)\neq\emptyset\right\}$$
$$\leq P\left\{\omega:\ \Gamma^r_{n,\kappa}(\omega)\setminus U_{\beta^{(3)}_{n,\kappa}}\Gamma_0\neq\emptyset\right\}+P\{\omega:\ \Gamma_0\setminus\Gamma^r_{n,\kappa}(\omega)\neq\emptyset\}.$$

Firstly, assume that for given $\kappa>0$, $n\geq n_0(\kappa)$ and $\omega\in\Omega$ the relation $\Gamma^r_{n,\kappa}(\omega)\setminus U_{\beta^{(3)}_{n,\kappa}}\Gamma_0\neq\emptyset$ is satisfied. Then there is $x_{n,\kappa}(\omega)\in\Gamma^r_{n,\kappa}(\omega)$ which does not belong to $U_{\beta^{(3)}_{n,\kappa}}\Gamma_0$. Because of $n\geq n_0(\kappa)$ we have $\beta^{(3)}_{n,\kappa}\leq\theta_1$. Consequently, by (Gr-$g_0$-$\Gamma_0$), $g_0(x_{n,\kappa}(\omega)\geq c_1(\beta^{(3)}_{n,\kappa})^{\delta_1}=\frac{2\kappa}{\gamma_n}$. Hence $\inf_{x\in Q_o\setminus\Gamma_0}(g_n(x,\omega)-g_0(x))\leq\frac{\kappa}{\gamma_n}-\frac{2\kappa}{\gamma_n}=-\frac{\kappa}{\gamma_n}$, and we can employ (CI2').

Now, assume that for given $\kappa>0$, $n\geq n_0(\kappa)$, and $\omega\in\Omega$ the relation $P\{\omega:\Gamma_0\setminus\Gamma^r_{n,\kappa}(\omega)\neq\emptyset\}$ is satisfied. Then there is $x_{n,\kappa}(\omega)\in\Gamma_0$ which does not belong to $\Gamma^r_{n,\kappa}(\omega)$. Because of $x_{n,\kappa}(\omega)\in Q_0$ we obtain $g_n(x_{n,\kappa}(\omega),\omega)>\frac{\kappa}{\sqrt{n}}$, and finally $\sup_{x\in\Gamma_0}(g_n(x,\omega)-g_0(x))\geq\frac{\kappa}{\gamma_n}$. It remains to employ (CO2'). □

4. Approximation of the optimal values and the solution sets. We turn to the optimal values and the solutions sets of the problems (P_0) and $(P_{n,\kappa})$.

In the following, the constraint sets and their approximations are not supposed to have a special form. Especially, $\Gamma_{n,\kappa}$ can be described by inequality constraints as in section 3, but it can also denote a set originating from a relaxation like $\Gamma^r_{n,\kappa}$.

We do not impose compactness conditions on Γ_0 and $\Gamma_{n,\kappa}$. Instead, for the sake of simplicity, we assume that the original and the approximating problems have a solution.

We start with the consideration of the optimal values. Results of this kind are, among others, needed for assertions about the solution sets. The assertion of Theorem 6 says, roughly speaking, that the optimal values of the approximate problem are greater (in the in-probability sense) than the true optimal value. The denotation "lower approximation" in the theorem is chosen because of the relationship to lower semicontinuity of a function of one variable—in our case the "variable" n.

THEOREM 6 (lower approximation of the optimal value). *Assume that the following conditions are satisfied:*

(VL1) *There exists a function* $\mathcal{H}_1 \in H$ *and for all* $\kappa > 0$ *a sequence* $(\beta_{n,\kappa}^{(1)})_{n\in N} \in B$ *such that*

$$\sup_{n\in N} P\left\{\omega :\ \Gamma_{n,\kappa}(\omega) \setminus U_{\beta_{n,\kappa}^{(1)}}\Gamma_0 \neq \emptyset\right\} \leq \mathcal{H}_1(\kappa).$$

(VL2) *There exists a function* $\mathcal{H}_2 \in H$ *and to all* $\kappa > 0$ *a sequence* $(\beta_{n,\kappa}^{(2)})_{n\in N} \in B$ *such that*

$$\sup_{n\in N} P\left\{\omega :\ \inf_{x\in U\Gamma_0}(f_{n,\kappa}(x,\omega) - f_0(x)) \leq -\beta_{n,\kappa}^{(2)}\right\} \leq \mathcal{H}_2(\kappa)$$

for a suitable neighborhood $U\Gamma_0$.

(VL3) *There exists a function* $\lambda \in \Lambda$ *such that for all* $\varepsilon > 0$

$$\forall\, x \in U_{\lambda(\varepsilon)}\Gamma_0 \cap U\Gamma_0 :\ f_0(x) \geq \Phi_0 - \varepsilon.$$

Then for all $\kappa > 0$, $\beta_{n,\kappa}^{(3)} = \max\{2\lambda^{-1}(\beta_{n,\kappa}^{(1)}), 2\beta_{n,\kappa}^{(2)}\}$, *and* $n_0(\kappa) = \min\{l :\ U_{\lambda(\frac{\beta_{l,\kappa}^{(3)}}{2})}\Gamma_0 \subset U\Gamma_0\}$ *the relation*

$$\sup_{n\geq n_0(\kappa)} P\left\{\omega :\ \Phi_{n,\kappa}(\omega) - \Phi_0 \leq -\beta_{n,\kappa}^{(3)}\right\} \leq \mathcal{H}_1(\kappa) + \mathcal{H}_2(\kappa)$$

holds.

Proof. Assume that for given $\kappa > 0$, $n \geq n_0(\kappa)$, and $\omega \in \Omega$ the relation $\Phi_{n,\kappa}(\omega) \leq \Phi_0 - \beta_{n,\kappa}^{(3)}$ holds. Then there exists $x_{n,\kappa}(\omega) \in \Gamma_{n,\kappa}(\omega)$ such that $f_{n,\kappa}(x_{n,\kappa}(\omega),\omega) = \Phi_{n,\kappa}(\omega) \leq \Phi_0 - \beta_{n,\kappa}^{(3)}$.

Firstly, let $x_{n,\kappa}(\omega) \in U_{\lambda(\frac{\beta_{n,\kappa}^{(3)}}{2})}\Gamma_0$. Then $\inf_{x\in U\Gamma_0}(f_{n,\kappa}(x,\omega) - f_0(x)) \leq f_{n,\kappa}(x_{n,\kappa}(\omega),\omega) - f_0(x_{n,\kappa}(\omega)) \leq \Phi_{n,\kappa}(\omega) - \Phi_0 + \frac{\beta_{n,\kappa}^{(3)}}{2} \leq -\frac{\beta_{n,\kappa}^{(3)}}{2} \leq -\beta_{n,\kappa}^{(2)}$, and we can employ (VL2).

Secondly, if $x_{n,\kappa}(\omega) \notin U_{\lambda(\frac{\beta_{n,\kappa}^{(3)}}{2})}\Gamma_0$, we have $\Gamma_{n,\kappa}(\omega) \setminus U_{\beta_{n,\kappa}^{(1)}}\Gamma_0 \neq \emptyset$. It remains to employ (VL1). □

The proof shows that the following assertion also holds, which applies to the important special case $\Gamma_{n,\kappa} \equiv \Gamma_0$.

COROLLARY 5. *Assume that* $\Gamma_{n,\kappa} \equiv \Gamma_0$ *and the following condition is satisfied:*

(VL2') *There exists a function* $\mathcal{H}_2 \in H$ *and to all* $\kappa > 0$ *a sequence* $(\beta_{n,\kappa}^{(2)})_{n\in N} \in B$ *such that*

$$\sup_{n\in N} P\left\{\omega :\ \inf_{x\in\Gamma_0}(f_{n,\kappa}(x,\omega) - f_0(x)) \leq -\beta_{n,\kappa}^{(2)}\right\} \leq \mathcal{H}_2(\kappa).$$

Then for all $\kappa > 0$ *the relation*

$$\sup_{n\in N} P\left\{\omega :\ \Phi_{n,\kappa}(\omega) - \Phi_0 \leq -\beta_{n,\kappa}^{(2)}\right\} \leq \mathcal{H}_2(\kappa)$$

holds.

In the following we consider the counterpart, so-called upper approximations, and distinguish two cases according to whether

$$\forall \kappa > 0: \sup_{n \geq n_0(\kappa)} P\left\{\omega:\ \Gamma_0 \setminus U_{\beta^{(1)}_{n,\kappa}} \Gamma_{n,\kappa}(\omega) \neq \emptyset\right\} \leq \mathcal{H}_1(\kappa) \text{ or}$$
$$\forall \kappa > 0: \sup_{n \geq n_0(\kappa)} P\left\{\omega:\ \Gamma_0 \setminus \Gamma_{n,\kappa}(\omega) \neq \emptyset\right\} \leq \mathcal{H}_1(\kappa)$$

is imposed. The first case, which is dealt with in Theorem 7, occurs if outer approximations are derived under condition (CO3). The second case, which is considered in Theorem 8, comes into play if $\Gamma_{n,\kappa}$ is obtained via relaxation; compare Remark 2. Recall that minimization is always taken with respect to $\Gamma_{n,\kappa}$. Hence, in Theorem 7 and Theorem 8 we will find different convergence rates.

THEOREM 7 (upper approximation of the optimal value I). *Assume that the following conditions are satisfied:*

(VU1) *There exists a function $\mathcal{H}_1 \in H$ and for all $\kappa > 0$ a sequence $(\beta^{(1)}_{n,\kappa})_{n\in N} \in B$ such that*

$$\sup_{n\in N} P\left\{\omega:\ \Gamma_0 \setminus U_{\beta^{(1)}_{n,\kappa}} \Gamma_{n,\kappa}(\omega) \neq \emptyset\right\} \leq \mathcal{H}_1(\kappa).$$

(VU2) *There exists a function $\mathcal{H}_2 \in H$ and for all $\kappa > 0$ a sequence $(\beta^{(2)}_{n,\kappa})_{n\in N} \in B$ such that*

$$\sup_{n\in N} P\left\{\omega:\ \sup_{x\in U\Psi_0} (f_{n,\kappa}(x,\omega) - f_0(x)) \geq \beta^{(2)}_{n,\kappa}\right\} \leq \mathcal{H}_2(\kappa)$$

for a suitable neighborhood $U\Psi_0$.

(VU3) *There exists a function $\lambda \in \Lambda$ such that for all $\varepsilon > 0$*

$$\forall\, x \in U_{\lambda(\varepsilon)}\Psi_0 \cap U\Psi_0:\ f_0(x) \leq \Phi_0 + \varepsilon.$$

Then for all $\kappa > 0$, $\beta^{(3)}_{n,\kappa} = \max\{2\lambda^{-1}(\beta^{(1)}_{n,\kappa}), 2\beta^{(2)}_{n,\kappa}\}$, and $n_0(\kappa) = \min\{l:\ U_{\lambda(\frac{\beta^{(3)}_{l,\kappa}}{2})}\Psi_0 \subset U\Psi_0\}$ the relation

$$\sup_{n\geq n_0(\kappa)} P\left\{\omega:\ \Phi_{n,\kappa}(\omega) - \Phi_0 \geq \beta^{(3)}_{n,\kappa}\right\} \leq \mathcal{H}_1(\kappa) + \mathcal{H}_2(\kappa)$$

holds.

Proof. Assume that for given $\kappa > 0$, $n \geq n_0(\kappa)$, and $\omega \in \Omega$ the relation $\Phi_{n,\kappa}(\omega) \geq \Phi_0 + \beta^{(3)}_{n,\kappa}$ holds. Then there exists $x_{n,\kappa}(\omega) \in \Gamma_{n,\kappa}(\omega)$ such that $f_{n,\kappa}(x_{n,\kappa}(\omega),\omega) = \Phi_{n,\kappa}(\omega) \geq \Phi_0 + \beta^{(3)}_{n,\kappa}$. To $x_{n,\kappa}(\omega)$ we select $\tilde{x}_{n,\kappa}(\omega) \in \Gamma_{n,\kappa}(\omega)$ such that $d(\tilde{x}_{n,\kappa}(\omega), \Psi_0) = \min_{x\in\Gamma_{n,\kappa}(\omega)} d(x,\Psi_0)$.

Firstly, assume that $\tilde{x}_{n,\kappa}(\omega) \in U_{\lambda(\frac{\beta^{(3)}_{n,\kappa}}{2})}\Psi_0$. Then $f_{n,\kappa}(\tilde{x}_{n,\kappa}(\omega),\omega) \geq \Phi_{n,\kappa}(\omega) \geq \Phi_0 + \beta^{(3)}_{n,\kappa} \geq f_0(\tilde{x}_{n,\kappa}(\omega)) + \frac{\beta^{(3)}_{n,\kappa}}{2}$ and consequently, $\sup_{x\in U\Psi_0}(f_{n,\kappa}(x,\omega) - f_0(x)) \geq \frac{\beta^{(3)}_{n,\kappa}}{2} \geq \beta^{(2)}_{n,\kappa}$. Hence we can make use of (VU2).

If $\tilde{x}_{n,\kappa}(\omega) \notin U_{\lambda(\frac{\beta^{(3)}_{n,\kappa}}{2})}\Psi_0$, we have $\Gamma_0 \setminus U_{\beta^{(1)}_{n,\kappa}} \Gamma_{n,\kappa}(\omega) \neq \emptyset$ and we can employ (VU1). □

If we impose the special upper semicontinuity condition (UCon) for f_0, we obtain the following corollary.

COROLLARY 6 (upper approximation of the optimal value I). *Assume that* (VU1), (VU2), *and the following condition* (UCon) *are satisfied:*

(UCon) *There exist constants* $c_2 > 0$, $\delta_2 > 0$, *and* $\theta_2 > 0$ *such that*

$$\forall x \in \bar{U}_{\theta_2}\Psi_0 : \ f_0(x) \leq \Phi_0 + c_2 d(x, \Psi_0)^{\delta_2}.$$

Then, for all $\kappa > 0$, $\beta_{n,\kappa}^{(3)} = \max\{2c_2(\beta_{n,\kappa}^{(1)})^{\delta_2}, 2\beta_{n,\kappa}^{(2)}\}$, *and* $n_0(\kappa) = \min\{l : \ U_{\lambda_{l,\kappa}}\Psi_0 \subset U\Psi_0 \cap \bar{U}_{\theta_2}\Psi_0\}$ *with* $\lambda_{l,\kappa} = (\frac{\beta_{l,\kappa}^{(3)}}{2c_2})^{\frac{1}{\delta_2}}$ *the relation*

$$\sup_{n \geq n_0(\kappa)} P\left\{\omega : \ \Phi_{n,\kappa}(\omega) - \Phi_0 \geq \beta_{n,\kappa}^{(3)}\right\} \leq \mathcal{H}_1(\kappa) + \mathcal{H}_2(\kappa)$$

holds.

Proof. We employ Theorem 7. Because of (UCon) we can choose $\lambda(\varepsilon) = (\frac{\varepsilon}{c_2})^{\frac{1}{\delta_2}}$ and consequently $\lambda^{-1}(\varepsilon) = c_2\varepsilon^{\delta_2}$. □

A similar corollary can be proved for the lower approximation of the optimal values. We give only the result for the upper approximation because we will use it in the following. Furthermore, the assertions can be supplemented by results similar to Corollary 3 and Corollary 4.

We could also consider a variant of (UCon) which refers to $\hat{\Psi}_0$ instead of Ψ_0. We refrain from giving a corresponding statement because (UCon) requires only f_0 not to vary too much outside Ψ_0. It is in its importance not comparable to a growth condition which requires that f_0 has to grow outside the reference set with a certain rate.

THEOREM 8 (upper approximation of the optimal value II). *Assume that the following conditions are satisfied:*

(VU1-R) *There exists a function* $\mathcal{H}_1 \in H$ *such that*

$$\sup_{n \in N} P\{\omega : \ \Gamma_0 \setminus \Gamma_{n,\kappa}(\omega) \neq \emptyset\} \leq \mathcal{H}_1(\kappa).$$

(VU2-R) *There exists a function* $\mathcal{H}_2 \in H$ *and for all* $\kappa > 0$ *a sequence* $(\beta_{n,\kappa}^{(2)})_{n \in N} \in B$ *such that*

$$\sup_{n \in N} P\left\{\omega : \ \sup_{x \in \Psi_0}(f_{n,\kappa}(x, \omega) - f_0(x)) \geq \beta_{n,\kappa}^{(2)}\right\} \leq \mathcal{H}_2(\kappa).$$

Then for all $\kappa > 0$ *the relation*

$$\sup_{n \in N} P\left\{\omega : \ \Phi_{n,\kappa}(\omega) - \Phi_0 \geq \beta_{n,\kappa}^{(2)}\right\} \leq \mathcal{H}_1(\kappa) + \mathcal{H}_2(\kappa)$$

holds.

Proof. Assume that for given $\kappa > 0$, $n \in N$, and $\omega \in \Omega$ the relation $\Phi_{n,\kappa}(\omega) \geq \Phi_0 + \beta_{n,\kappa}^{(2)}$ holds. Then there exists $x_{n,\kappa}(\omega) \in \Gamma_{n,\kappa}(\omega)$ such that $f_{n,\kappa}(x_{n,\kappa}(\omega), \omega) = \Phi_{n,\kappa}(\omega) \geq \Phi_0 + \beta_{n,\kappa}^{(2)}$. To $x_{n,\kappa}(\omega)$ we select $\tilde{x}_{n,\kappa}(\omega) \in \Gamma_{n,\kappa}(\omega)$ such that $d(\tilde{x}_{n,\kappa}(\omega), \Psi_0) = \min_{x \in \Gamma_{n,\kappa}(\omega)} d(x, \Psi_0)$.

If $\tilde{x}_{n,\kappa}(\omega) \in \Psi_0$ we have

$$f_{n,\kappa}(\tilde{x}_{n,\kappa}(\omega), \omega) \geq \Phi_{n,\kappa}(\omega) \geq \Phi_0 + \beta_{n,\kappa}^{(2)} = f_0(\tilde{x}_{n,\kappa}(\omega)) + \beta_{n,\kappa}^{(2)}$$

and consequently, $\sup_{x \in \Psi_0}(f_{n,\kappa}(x, \omega) - f_0(x)) \geq \beta_{n,\kappa}^{(2)}$. Hence (VU2-R) can be utilized. Otherwise we have $\Gamma_0 \setminus \Gamma_{n,\kappa}(\omega) \neq \emptyset$ and can employ the first assumption. □

Now we turn to the solution sets. We use the abbreviation

$$\hat{\Psi}_0 = \{x \in E :\ f_0(x) \le \Phi_0\}.$$

THEOREM 9 (inner approximation of the solution set). *Assume that* (VL1), (VL2), *and the following assumptions are satisfied:*

(SI3) *There exists a function* $\mathcal{H}_3 \in H$ *and for all* $\kappa > 0$ *a sequence* $(\hat{\beta}^{(2)}_{n,\kappa})_{n\in N} \in B$ *and* $n_0(\kappa)$ *such that*

$$\sup_{n\ge n_0(\kappa)} P\left\{\omega :\ \Phi_{n,\kappa}(\omega) - \Phi_0 \ge \hat{\beta}^{(2)}_{n,\kappa}\right\} \le \mathcal{H}_3(\kappa).$$

(SI4) *There exists a function* $\nu \in \Lambda$ *such that for all* $\varepsilon > 0$

$$U_\varepsilon \Psi_0 \supset U_{\nu(\varepsilon)}\Gamma_0 \cap U_{\nu(\varepsilon)}\hat{\Psi}_0.$$

(SI5) *There exists a function* $\mu \in \Lambda$ *such that for all* $\varepsilon > 0$

$$\forall\, x \in U_\varepsilon\Gamma_0 \setminus U_\varepsilon\hat{\Psi}_0 :\ f_0(x) \ge \Phi_0 + \mu(\varepsilon).$$

Then for all $\kappa > 0$, $\beta^{(3)}_{n,\kappa} = \max\{\nu^{-1}(\beta^{(1)}_{n,\kappa}), \nu^{-1}(\mu^{-1}(\beta^{(2)}_{n,\kappa} + \hat{\beta}^{(2)}_{n,\kappa}))\}$, *and* $n_1(\kappa) = \min\{l \ge n_0(\kappa) :\ U_{\nu(\beta^{(3)}_{l,\kappa})}\Gamma_0 \subset U\Gamma_0\}$ *the relation*

$$\sup_{n\ge n_1(\kappa)} P\left\{\omega :\ \Psi_{n,\kappa}(\omega) \setminus U_{\beta^{(3)}_{n,\kappa}}\Psi_0 \ne \emptyset\right\} \le \mathcal{H}_1(\kappa) + \mathcal{H}_2(\kappa) + \mathcal{H}_3(\kappa)$$

holds.

Proof. Let $\tilde{g}_{n,\kappa}(x,\omega) := f_{n,\kappa}(x,\omega) - \Phi_{n,\kappa}(\omega)$, $\tilde{g}_0(x) := f_0(x) - \Phi_0$. Then $\Psi_{n,\kappa}(\omega) = \Gamma_{n,\kappa}(\omega) \cap \{x \in E :\ \tilde{g}_{n,\kappa}(x,\omega) \le 0\}$ and $\Psi_0 = \Gamma_0 \cap \{x \in E :\ \tilde{g}_0(x) \le 0\}$. Furthermore,

$$\begin{aligned}
&\sup_{n\ge n_0(\kappa)} P\left\{\omega :\ \inf_{x\in U\Gamma_0\setminus\Psi_0} (\tilde{g}_{n,\kappa}(x,\omega) - \tilde{g}_0(x)) \le -\beta^{(2)}_{n,\kappa} - \hat{\beta}^{(2)}_{n,\kappa}\right\}\\
&\le \sup_{n\ge n_0(\kappa)} P\left\{\omega :\ \inf_{x\in U\Gamma_0\setminus\Psi_0} (f_{n,\kappa}(x,\omega) - f_0(x)) \le -\beta^{(2)}_{n,\kappa}\right\}\\
&+ \sup_{n\ge n_0(\kappa)} P\left\{\omega :\ -\Phi_{n,\kappa}(\omega) + \Phi_0 \le -\hat{\beta}^{(2)}_{n,\kappa}\right\} \le \mathcal{H}_2(\kappa) + \mathcal{H}_3(\kappa) =: \tilde{\mathcal{H}}_2(\kappa).
\end{aligned}$$

It remains to apply Theorem 1 with $\tilde{\beta}^{(2)}_{n,\kappa} = \beta^{(2)}_{n,\kappa} + \hat{\beta}^{(2)}_{n,\kappa}$ and $\tilde{\mathcal{H}}_2$. □

When applying our results to problems in decision theory or estimation theory, the most critical assumption is probably (SI4). Fortunately, there are several important applications where some quantities do not vary with n and ν can be avoided or, as in our third example, (SI4) is easy to verify.

We emphasize that we can choose $\nu(\varepsilon) = \varepsilon$ if $\Gamma_{n,\kappa} \equiv \Gamma_0$ and the growth condition is given with respect to Ψ_0.

(SI5-W) There exists a function $\tilde{\mu} \in \Lambda$ such that for all $\varepsilon > 0$

$$\forall\, x \in \Gamma_0 \setminus U_\varepsilon\Psi_0 :\ f_0(x) \ge \Phi_0 + \tilde{\mu}(\varepsilon).$$

This case is considered in [13]. Imposing the special form of $\tilde{\mu}$ in [13], Pflug's result can be derived from Corollary 3.

Furthermore, if $U_{\tilde{\varepsilon}}\Psi_0 \subset \Gamma_0$ for a suitable $\tilde{\varepsilon} > 0$, we can also deal with $\nu(\varepsilon) = \varepsilon$ for all $\varepsilon \leq \tilde{\varepsilon}$.

In the general case, however, if the constraint set and the objective function are approximated simultaneously and the solution lies on the boundary of the constraint set, ν cannot be ignored. Only in rare cases one should have enough information to determine it exactly. One way out are adaptive methods for successive approximation of ν. However, even if one does not succeed in determining ν with satisfactory accuracy, our results still yield assertions on the convergence rate, albeit without a reliable constant. Results of that kind can be used to derive asymptotic confidence sets if a limiting distribution is not available.

The following corollary incorporates sufficient conditions for the needed approximation of the optimal value. It combines Theorem 9 and Corollary 6.

COROLLARY 7 (inner approximation of the solution set). *Assume that* (VL1) *and* (VU1) *are fulfilled with the same function $\mathcal{H}_1$ and the same sequences $(\beta^{(1)}_{n,\kappa})_{n\in N}$, and* (VL2) *and* (VU2) *are fulfilled with the same function $\mathcal{H}_2$ and the same sequences $(\beta^{(2)}_{n,\kappa})_{n\in N}$. Furthermore, suppose that* (SI4), (UCon), *and the following condition are satisfied:*

(Gr-f_0) There exist constants $c_1 > 0$, $\delta_1 > 0$, and $\theta_1 > 0$ such that

$$\forall x \in \bar{U}_{\theta_1}\Gamma_0 :\ f_0(x) - \Phi_0 \geq c_1 \cdot d(x, \hat{\Psi}_0)^{\delta_1}.$$

Then for all $\kappa > 0$, $\beta^{(3)}_{n,\kappa} = \max\{\nu^{-1}(\beta^{(1)}_{n,\kappa}), \nu^{-1}((\frac{\hat{\beta}^{(2)}_{n,\kappa}+\beta^{(2)}_{n,\kappa}}{c_1})^{\frac{1}{\delta_1}})\}$, $\hat{\beta}^{(2)}_{n,\kappa} = \max\{2c_2 (\beta^{(1)}_{n,\kappa})^{\delta_2}, 2\beta^{(2)}_{n,\kappa}\}$, $n_1(\kappa) = \min\{l :\ U_{\beta^{(3)}_{l,\kappa}}\Gamma_0 \subset U\Gamma_0,\ U_{\hat{\lambda}_{l,\kappa}}\Psi_0 \subset U\Psi_0,\ \beta^{(3)}_{l,\kappa} \leq \theta_1,\ \hat{\lambda}_{l,\kappa} \leq \theta_2\}\}$, and $\hat{\lambda}_{n,\kappa} = (\frac{\hat{\beta}^{(2)}_{n,\kappa}}{2c_2})^{\frac{1}{\delta_1}}$ the relation

$$\sup_{n\geq n_1(\kappa)} P\left\{\omega :\ \Psi_{n,\kappa}(\omega) \setminus U_{\beta^{(3)}_{n,\kappa}}\Psi_0 \neq \emptyset\right\} \leq 2\mathcal{H}_1(\kappa) + 2\mathcal{H}_2(\kappa)$$

holds.

Proof. We apply Theorem 9 together with Corollary 6. (SI3) is satisfied with $\hat{\beta}^{(2)}_{n,\kappa}$, $\mathcal{H}_3 = \mathcal{H}_1 + \mathcal{H}_2$, and $\hat{n}_0(\kappa) = \min\{l :\ \hat{\lambda}_{l,\kappa} \leq \theta_2,\ U_{\hat{\lambda}_{l,\kappa}}\Psi_0 \subset U\Psi_0\}$. Theorem 9 with $\mu^{-1}(\varepsilon) = (\frac{\varepsilon}{c_1})^{\frac{1}{\delta_1}}$ yields the conclusion. □

The following condition covers the cases dealt with in Theorem 5 or Corollary 4.

(CK-R) There exists a function $\mathcal{H}_1 \in H$ and for all $\kappa > 0$ a sequence $(\beta^{(1)}_{n,\kappa})_{n\in N} \in B$ such that

$$\sup_{n\in N} P\left\{\omega :\ \left(\Gamma_{n,\kappa}(\omega) \setminus U_{\beta^{(1)}_{n,\kappa}}\Gamma_0\right) \cup (\Gamma_0 \setminus \Gamma_{n,\kappa}(\omega)) \neq \emptyset\right\} \leq \mathcal{H}_1(\kappa).$$

If (CK-R) is satisfied, also (VL1) and (VU1) are fulfilled with $\mathcal{H}_1$ and $\beta^{(1)}_{n,\kappa}$. Consequently, if $\Gamma_{n,\kappa} = \Gamma^r_{n,\kappa}$, we can employ Theorem 5 or Corollary 4 in order to determine a suitable $\rho_{n,\kappa}$ and formulate sufficient conditions for (VL1) and (VU1).

Finally, we consider outer approximations of the solution set via $\rho_{n,\kappa}$-optimal solutions of the approximating problems.

Let

$$\hat{\Psi}^r_{n,\kappa}(\omega) := \{x \in E :\ f_{n,\kappa}(x,\omega) \leq \Phi_{n,\kappa}(\omega) + \rho_{n,\kappa}\}.$$

$\Gamma_{n,\kappa}$ can, e.g., be specified as $U_{\beta^{(1)}_{n,\kappa}}\Gamma_n$ or as $\Gamma^r_{n,\kappa}$.

THEOREM 10 (outer approximation of the solution set, relaxation). *Assume that* (VU1), (VU2), *and the following assumption are satisfied:*

(SO3) *There exists a function* $\mathcal{H}_3 \in H$ *and for all* $\kappa > 0$ *a sequence* $(\hat{\beta}^{(2)}_{n,\kappa})_{n\in N} \in B$ *and* $n_0(\kappa)$ *such that*

$$\sup_{n\geq n_0(\kappa)} P\left\{\omega:\ \Phi_{n,\kappa}(\omega) - \Phi_0 \leq -\hat{\beta}^{(2)}_{n,\kappa}\right\} \leq \mathcal{H}_3(\kappa).$$

Then for all $\kappa > 0$, $\rho_{n,\kappa} = \beta^{(2)}_{n,\kappa} + \hat{\beta}^{(2)}_{n,\kappa}$, *and* $\beta^{(3)}_{n,\kappa} = \max\{\beta^{(1)}_{n,\kappa}, \beta^{(2)}_{n,\kappa} + \hat{\beta}^{(2)}_{n,\kappa}\}$ *the relation*

$$\sup_{n\geq n_0(\kappa)} P\left\{\omega:\ \Psi_0 \setminus (\hat{\Psi}^r_{n,\kappa}(\omega) \cap U_{\beta^{(1)}_{n,\kappa}}\Gamma_{n,\kappa}(\omega)) \neq \emptyset\right\} \leq \mathcal{H}_1(\kappa) + \mathcal{H}_2(\kappa) + \mathcal{H}_3(\kappa)$$

holds.

Proof. Let $n \geq n_0(\kappa)$. We apply Theorem 3 and underline the denotations that correspond to that in Theorem 3. With $\underline{g}_{n,\kappa}(x,\omega) := f_{n,\kappa}(x,\omega) - \Phi_{n,\kappa}(\omega)$ and $\underline{g}_0(x) := f_0(x) - \Phi_0$ we have $\Psi_0 = \{x \in \Gamma_0 :\ \underline{g}_0(x) \leq 0\}$ and $\hat{\Psi}^r_{n,\kappa}(\omega) := \{x \in E :\ \underline{g}_{n,\kappa}(x,\omega) \leq \rho_{n,\kappa}\}$. Furthermore, let $\underline{Q}_{n,\kappa} = \Gamma_{n,\kappa}$, $\hat{\underline{\Gamma}}^r_{n,\kappa} = \hat{\Psi}^r_{n,\kappa}$, $\underline{Q}_0 = \Gamma_0$, and $\hat{\underline{\Gamma}}_0 = \hat{\Psi}_0$.

Because of

$$\begin{aligned}
&P\left\{\omega:\ \inf_{x\in\Psi_0}((f_{n,\kappa}(x,\omega) - \Phi_{n,\kappa}(\omega)) - (f_0(x) - \Phi_0)) \geq \beta^{(2)}_{n,\kappa} + \hat{\beta}^{(2)}_{n,\kappa}\right\}\\
&\leq P\left\{\omega:\ \inf_{x\in\Psi_0}(f_{n,\kappa}(x,\omega) - f_0(x)) \geq \beta^{(2)}_{n,\kappa}\right\} + P\left\{\omega:\ -\Phi_{n,\kappa}(\omega) + \Phi_0 \geq \hat{\beta}^{(2)}_{n,\kappa}\right\}\\
&\leq \mathcal{H}_2(\kappa) + \mathcal{H}_3(\kappa) =: \underline{\mathcal{H}}_2(\kappa)
\end{aligned}$$

condition (CO2) is satisfied with $\underline{\beta}^{(2)}_{n,\kappa} = \beta^{(2)}_{n,\kappa} + \hat{\beta}^{(2)}_{n,\kappa}$ and $\underline{\mathcal{H}}_2$. □

A similar result can be obtained if we impose (VU1-R) and (VU2-R); compare the remark after Theorem 3.

5. Examples.

5.1. Example 5.1—Approximation of the objective functions. Firstly, we will discuss a simple example in order to show how one can deal with the uniform convergence assumption for the objective functions. At the first glance this assumption seems to be rather restrictive. There is, however, a growing number of results from probability theory yielding assertions of that kind; cf. [16], [1], [13]. Nevertheless there is still the need for sufficient conditions for the convergence assumptions. The following example, though very simple, is intended to show how one can proceed in principle also in more involved cases. The approach will be further developed elsewhere.

We assume $E = R^p$ and consider a fixed compact constraint set K and a linear objective function $q(z)^T x$ with $x = (x_1, \ldots, x_p)^T$ $q = (q_1, \ldots, q_p)^T$, and $q_i | R^m \to R^1$. z is the realization of a random vector Z with a given distribution P_Z on the sigma-field of Borel sets of R^m. The range of $q(Z)$ is supposed to be bounded. The problem

$$(P_0) \quad \min_{x\in K} \mathrm{E} q(Z)^T x$$

is approximated, replacing the expectation E with respect to P_Z by the expectation with respect to the empirical distribution based on a sequence $(Z^{(j)})_{j\in N}$ of indepen-

dent random vectors which are distributed according to P_Z:

$$(P_n) \quad \min_{x\in K} \frac{1}{n}\sum_{j=1}^{n} q\big(Z^{(j)}\big)^T x.$$

We assume $\mathrm{E}q(Z) \neq \mathbf{0}$, because otherwise the problem becomes trivial, and abbreviate $m := \max_{i=1,\dots,p} \sup_\omega |q_i(Z(\omega))|$. We consider the sets $K_k = \{x \in K :\ k-1 < ||x|| \leq k\},\ k = 1, 2, \dots$, where $||\cdot||$ denotes the Euclidean norm, and a suitable neighborhood $\bar{U}K$. Let $I_K := \{k :\ K_k \cap \bar{U}K \neq \emptyset\}$. Hence we obtain by Hoeffding's inequality ([4], [1]):

$$\begin{aligned}
&P\left\{\omega :\ \sup_{x\in\bar{U}K} |\frac{1}{n}\sum_{j=1}^{n} q(Z^{(j)}(\omega))^T x - \mathrm{E}q(Z)^T x| \geq \frac{\kappa}{\sqrt{n}}\right\}\\
&\leq \sum_{k\in I_K} P\left\{\omega :\ \sup_{x\in K_k} |\frac{1}{n}\sum_{j=1}^{n} q(Z^{(j)}(\omega))^T x - \mathrm{E}q(Z)^T x| \geq \frac{\kappa}{\sqrt{n}}\right\}\\
&\leq \sum_{k\in I_K} P\left\{\omega :\ \max_{i=1,\dots,p} |\frac{1}{n}\sum_{j=1}^{n} q_i(Z^{(j)}(\omega)) - \mathrm{E}q_i(Z)| \geq \frac{\kappa}{k\sqrt{n}}\right\}\\
&\leq 2p \sum_{k\in I_K} e^{-\frac{\kappa^2}{2k^2m^2}} =: \mathcal{H}_2(\kappa).
\end{aligned}$$

Of course this inequality can be further improved. For example, due to the linearity, it is enough to consider the boundary of K instead of the whole set K. Employing other concentration-of-measure inequalities, the boundedness condition for $q(Z)$ can also be weakened. We give the rough estimation above, because the basic idea of the approach can often be utilized, even if the functions have a more involved form.

In order to derive assertions about the solution set, we can employ Theorem 9 and (in case of a nonunique solution) Theorem 10. (VL1), (VU1), and (VU1-R) are satisfied with $\mathcal{H}_1 \equiv 0$. (VL2), (VU2), and (VU2-R) are fulfilled with $\beta^{(2)}_{n,\kappa} = \frac{\kappa}{\sqrt{n}}$ and $\mathcal{H}_2$. The function ν, which occurs in condition (SI4), can be determined if we assume a special form of K. It remains to investigate the semicontinuity condition (UCon) and the growth condition (Gr-f_0), where we can replace $\bar{U}_{\theta_1}\Gamma_0$ with Γ_0.

We have $f_0(x) - \Phi_0 \leq ||\mathrm{E}q(Z)||d(x, \Psi_0)$; hence (UCon) is satisfied with $c_2 = ||\mathrm{E}q(Z)||$ and $\delta_2 = 1$. (Gr-f_0) refers to the distance to $\hat{\Psi}_0$. To $x \in \Gamma_0$ we find $\hat{x}_0 \in \{x \in R^p :\ \mathrm{E}q(Z)^T x = \Phi_0\}$ such that $|f_0(x) - \Phi_0| = |\mathrm{E}q(Z)^T(x - \hat{x}_0)| = ||\mathrm{E}q(Z)|| \cdot ||x - \hat{x}_0|| = ||\mathrm{E}q(Z)||d(x, \hat{\Psi}_0)$. Consequently (Gr-$f_0$) is satisfied with $c_1 = ||\mathrm{E}q(Z)||$ and $\delta_1 = 1$.

5.2. Example 5.2—Approximation of a probabilistic constraint. Secondly, we consider the approximation of a constraint set which is determined by a probabilistic constraint. Replacing the true probability measure with the empirical measure, we obtain a sequence of approximating constraint functions.

In detail, we assume that the constraint function has the special form

$$g_0(x) = \alpha - P_Z((-\infty, \gamma(x)]) = \alpha - F_Z(\gamma(x)).$$

Z is a real-valued random variable with given distribution P_Z on the σ-field $\mathcal{B}^1$. $\alpha \in (0,1)$ denotes a probability level and $\gamma | E \to R^1$ a given concave function. The inequality constraint $g_0(x) \leq 0$ then reads as $P\{\omega : Z(\omega) \leq \gamma(x)\} \geq \alpha$. We assume that

$$\Gamma_0 = \hat{\Gamma}_0 = \{x \in E : \ g_0(x) \leq 0\} \neq \emptyset.$$

The approximating constraint set has the form

$$\Gamma_n(\omega) = \{x \in E : \ \alpha - F_n(\gamma(x), \omega) \leq 0\}$$

with the empirical distribution function F_n .

In order to fulfill (CI2') and (CO2') we can directly apply the Dvoretzky–Kiefer–Wolfowitz inequality with Massart's bound ([10], [1]), and we obtain $P\{\omega : \sqrt{n} \sup_{x \in R^1} |(\alpha - F_n(\gamma(x), \omega)) - (\alpha - F_Z(\gamma(x))| > \kappa\} \leq 2e^{-2\kappa^2}$.

(CI3) is not needed. In order to fulfill (Gr-g_0), we will impose growth conditions for F_Z and γ.

Assume that, for the given probability level α, the α-quantile q_α of F_Z is unique and consider a compact set $\tilde{K}$ such that $q_\alpha \in \text{int}\tilde{K}$. Furthermore, let $X_{\tilde{K}} := \{x \in E : \gamma(x) \in \tilde{K}\}$ and suppose that the following conditions are satisfied:

(IG) There exist positive constants $c_{1,\gamma}$, $c_{1,F}$, $\delta_{1,\gamma}$, and $\delta_{1,F}$ such that $\forall y \in \tilde{K}$ with $y < q_\alpha$: $\alpha - F_Z(y) > c_{1,F} d(y, q_\alpha)^{\delta_{1,F}}$ and $\forall x \in X_{\tilde{K}}$: $\gamma(x) < q_\alpha - c_{1,\gamma} d(x, \Gamma_0)^{\delta_{1,\gamma}}$.

(Gr-g_0) with $\tilde{K}$ instead of UQ_0 and a strict inequality is then satisfied with $\tilde{c}_1 = c_F (c_{1,\gamma})^{\delta_F}$ and $\tilde{\delta}_1 = \delta_{1,\gamma} \cdot \delta_F$.

If Γ_0 is single-valued, it remains to apply Theorem 1. Otherwise we employ Theorem 2 and assume that the following condition is satisfied:

(OG) There exist positive constants $c_{2,\gamma}$, $c_{2,F}$, $\delta_{2,\gamma}$, and $\delta_{2,F}$ such that $\forall y \in \tilde{K}$ with $y > q_\alpha$: $F_Z(y) - \alpha > c_{2,F} d(y, q_\alpha)^{\delta_{2,F}}$. Furthermore, there exists an $\tilde{\varepsilon} > 0$ such that $CI(\tilde{\varepsilon}) \neq \emptyset$ and $\forall x \in \Gamma_0 \setminus CI(\tilde{\varepsilon}) : \ \gamma(x) > q_\alpha + c_{2,\gamma} d(x, (E \setminus \Gamma_0)^{\delta_{2,\gamma}}$.

Hence, with respect to (CO3) we obtain for all $0 < \varepsilon \leq \tilde{\varepsilon}$ $\Gamma_0 \subset \bar{U}_\varepsilon CI(\varepsilon)$ and $\forall x \in \Gamma_0 \setminus CI(\tilde{\varepsilon}) : \ g_0(x) < -\tilde{c}_2 d(x, E \setminus \Gamma_0)^{\tilde{\delta}_2}$ with $\tilde{c}_2 = c_F (c_{2,\gamma})^{\delta_F}$ and $\tilde{\delta}_2 = \delta_{2,\gamma} \cdot \delta_F$. Thus in Theorem 2 we can choose $\mu(\varepsilon) = \tilde{c}_2 \varepsilon^{\tilde{\delta}_2}$.

Consequently, for all $\kappa \leq \tilde{\varepsilon}$, $\beta^{(3)}_{n,\kappa} = \max\{(\frac{\kappa}{\tilde{c}_1})^{\frac{1}{\tilde{\delta}_1}} n^{-\frac{1}{2\tilde{\delta}_1}}, (2\frac{\kappa}{\tilde{c}_2})^{\frac{1}{\tilde{\delta}_2}} n^{-\frac{1}{2\tilde{\delta}_2}}\}$, and $n_0(\kappa) = \min\{l : \beta^{(3)}_{l,\kappa} \leq 2\tilde{\kappa}, \ \gamma(\Gamma_0 \setminus CI(\frac{\beta^{(3)}_{l,\kappa}}{2})) \subset K\}$ the relation $\sup_{n \geq n_0(\kappa)} P\{\omega : \ (\Gamma_n(\omega) \setminus U_{\beta^{(3)}_{n,\kappa}} \Gamma_0) \cup (\Gamma_0 \setminus (U_{\beta^{(3)}_{n,\kappa}} \Gamma_n(\omega)) \neq \emptyset\} \leq 4e^{-2\kappa^2}$ holds.

5.3. Example 5.3—Quantile estimation. Finally, we consider quantile estimation because here relaxation of the constraint set comes into play in a natural way. Papers dealing with quantile estimation usually assume that the distribution function is strictly increasing in a neighborhood of the quantile (cf. [5], [6]). There are, however, applications where one cannot a priori assume that the lower and the upper quantile coincide.

We consider, as in the foregoing example, a real-valued random variable Z with distribution P_Z and distribution function F_Z. We will, for a fixed $\alpha \in (0,1)$, investigate the lower α-quantile

$$q^l_\alpha := \inf \left\{x \in R^1 : F_Z(x) \geq \alpha\right\}.$$

We consider the constraint set

$$\Gamma_0 := \{x \in R : F_Z(x) \geq \alpha\}$$

and the optimization problem

$$(P_0) \quad \min_{x \in \Gamma_0} x.$$

As F_Z is upper semicontinuous by definition, the set Γ_0 is closed and the minimum q_α^l will be attained.

(P_0) could be approximated replacing F_Z by the empirical distribution function F_n. Unfortunately, the set $\{x \in R^1 : F_n(x) \geq \alpha\}$, in general, does not approximate the whole set Γ_0. In [20] we showed that with a suitable relaxation $\rho_{n,\kappa}$ the solutions to the approximate problems convergence in probability to the desired quantile. Here we can proceed in a similar way; consider the modified constraint set $\Gamma_{n,\kappa}$ with

$$\Gamma_{n,\kappa}(\omega) := \left\{x \in R : F_n(x,\omega) > \alpha - \frac{\kappa}{\sqrt{n}}\right\}$$

and investigate the approximating optimization problems

$$(P_{n,\kappa}) \quad \min_{x \in \Gamma_{n,\kappa}} x.$$

$(P_{n,\kappa})$ has a unique solution, too.

In order to obtain a convergence rate, we need some knowledge about F_Z, e.g., a growth condition.

THEOREM 11 (quantile estimation). *Assume that there exist constants $c > 0$, $\delta > 0$, and $\theta > 0$ such that $\forall\, x \in \bar{U}_\theta \Gamma_0 :\ F_Z(x) < \alpha - cd(x,\Gamma_0)^\delta$. Then for all $\kappa > 0$, $\beta_{n,\kappa}^{(3)} = (\frac{2\kappa}{c})^{\frac{1}{\delta}} n^{-\frac{1}{2\delta}}$, and $n_0(\kappa) = \min\{l : \beta_{l,\kappa}^{(3)} \leq \theta\}$ the relations*

$$\sup_{n \geq n_0(\kappa)} P\{\omega : (\Gamma_{n,\kappa}(\omega) \setminus U_{\beta_{n,\kappa}^{(3)}} \Gamma_0) \cup (\Gamma_0 \setminus \Gamma_{n,\kappa}(\omega)) \neq \emptyset\} \leq 2e^{-2\kappa^2} \text{ and}$$

$$\sup_{n \geq n_0(\kappa)} P\{\omega : (\Psi_{n,\kappa}(\omega) \setminus U_{\beta_{n,\kappa}^{(3)}} \Psi_0) \cup (\Psi_0 \setminus U_{\beta_{n,\kappa}^{(3)}} \Psi_{n,\kappa}(\omega)) \neq \emptyset\} \leq 2e^{-2\kappa^2}$$

hold.

Proof. In order to prove the first assertion, we employ Corollary 4 with strict inequalities; compare Remark 1. Condition (Gr-g_0-Γ_0) is fulfilled by assumption. Due to the Dvoretzky–Kiefer–Wolfowitz inequality with Massart's bound ([10], [1]), "strict" variants of (CI2') and (CO2') are satisfied with $\gamma_n = n^{\frac{1}{2}}$ and $\mathcal{H}_2(\kappa) = e^{-2\kappa^2}$.

The second assertion could be derived via a variant of Theorem 9, which takes into account that the objective function is not approximated. We will instead give a direct proof.

Assume that for given $\kappa > 0$, $n \geq n_0(\kappa)$, and $\omega \in \Omega$ the relation $(\Psi_{n,\kappa}(\omega) \setminus U_{\beta_{n,\kappa}^{(3)}} \Psi_0) \cup (\Psi_0 \setminus U_{\beta_{n,\kappa}^{(3)}} \Psi_{n,\kappa}(\omega)) \neq \emptyset$ holds. Then there are $x_{n,\kappa}(\omega) \in \Psi_{n,\kappa}(\omega)$ and $x_0 \in \Psi_0$ such that either $x_{n,\kappa}(\omega) \leq x_0 - \beta_{n,\kappa}^{(3)}$ or $x_0 \leq x_{n,\kappa} - \beta_{n,\kappa}^{(3)}$ is satisfied. Since the solutions Ψ_0 and $\Psi_{n,\kappa}$ consist of the left boundary points of the constraint sets, we obtain $(\Gamma_{n,\kappa}(\omega) \setminus U_{\beta_{n,\kappa}^{(3)}} \Gamma_0) \cup (\Gamma_0 \setminus \Gamma_{n,\kappa}(\omega)) \neq \emptyset$, and the conclusion follows by the first assertion. □

Acknowledgment. The author is grateful to two anonymous referees for constructive comments and suggestions.

REFERENCES

[1] L. Devroye and G. Lugosi, *Combinatorical Methods in Density Estimation*, Springer, New York, 2001.

[2] Y. M. Ermoliev and V. I. Norkin, *Normalized convergence in stochastic optimization*, Ann. Oper. Res., 30 (1991), pp. 187–198.

[3] O. Gersch, *Convergence in Distribution of Random Closed Sets Applications in Stability Theory of Stochastic Optimisation*, Dissertation thesis, Technical University Ilmenau, 2007.

[4] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, J. Amer. Statist. Assoc., 58 (1963), pp. 13–30.

[5] A. I. Kibzun and Y. S. Kan, *Stochastic Programming Problems*, Wiley, New York, 1996.

[6] K. Knight, *What are the limiting distibutions of quantile estimators?* Technical report, University of Toronto, Toronto, ON, 1999.

[7] P. Lachout, E. Liebscher, and S. Vogel, *Strong convergence of estimators as ε_n-estimators of optimization problems*, Ann. Inst. Statist. Math., 57 (2005), pp. 291–313.

[8] P. Lachout and S. Vogel, *On continuous convergence and epi-convergence of random functions. Part* I: *Theory and relations*, Kybernetika, 39 (2003), pp. 75–98.

[9] P. Lachout and S. Vogel, *On continuous convergence and epi-convergence of random functions. Part* II: *Sufficient conditions and applications*, Kybernetika, 39 (2003), pp. 99–118.

[10] P. Massart, *The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality*, Ann. Probab., 18 (1990), pp. 1269–1293.

[11] A. Nemirovski and A. Shapiro, *Scenario approximations of chance constraints*, SIAM J. Optim., 17 (2006), pp. 969–996.

[12] G. C. Pflug, *Asymptotic dominance and confidence for solutions of stochastic programs*, Czechoslovak J. Oper. Res., 1 (1992), pp. 21–30.

[13] G. C. Pflug, *Stochastic optimization and statistical inference*, in Stochastic Programming Handbooks in Operations Research and Mangement Science 10, A. Ruszczyński and A. Shapiro, eds., Elsevier, New York, 2003, pp. 427–482.

[14] W. Römisch, *Stability of stochastic programming*, in Stochastic Programming Handbooks in Operations Research and Mangement Science Vol 10, A. Ruszczyński and A. Shapiro, eds., Elsevier, New York, 2003, pp. 483–554.

[15] A. Shapiro, *Monte Carlo sampling methods*, in Stochastic Programming Handbooks in Operations Research and Mangement Science 10, A. Ruszczyński and A. Shapiro, eds., Elsevier, New York, 2003, pp. 353–425.

[16] A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes*, Springer, New York, 1996.

[17] S. Vogel, *A stochastic approach to stability in stochastic programming*, J. Comput. Appl. Math., Series Appl. Analysis and Stochastics, 56 (1994), pp. 65–96.

[18] S. Vogel, *On Stability in Stochastic Programming - Sufficient Conditions for Continuous Convergence and Epi-Convergence*, preprint, TU Ilmenau, 1995.

[19] S. Vogel, *On semicontinuous approximations of random closed sets with application to random optimization problems*, Ann. Oper. Res., 142 (2006), pp. 169–282.

[20] S. Vogel, *Qualitative stability of stochastic programs with applications in asymptotic statistics*, Statist. Decisions, 23 (2005), pp. 219–248.

SIAM J. OPTIM.
Vol. 19, No. 3, pp. 1489–1509

ON EXTENSION OF FENCHEL DUALITY AND ITS APPLICATION*

LI GUOYIN† AND NG KUNG FU†

Abstract. By considering the epigraphs of conjugate functions, we extend the Fenchel duality, applicable to a (possibly infinite) family of proper lower semicontinuous convex functions on a Banach space. Applications are given in providing fuzzy KKT conditions for semi-infinite programming.

Key words. Fenchel duality, epigraph, Karush–Kuhn–Tucker (KKT) conditions, semi-infinite programming

AMS subject classifications. Primary, 90C34, 90C25; Secondary, 52A07, 41A29, 90C46

DOI. 10.1137/080716803

1. Introduction.

The famous Fenchel duality theorem can be stated as follows (cf. [30, Corollary 2.8.5]): For any family of finitely many proper lower semicontinuous convex functions $f_0, f_1, \ldots, f_n$ on a Banach space X, if $\mathrm{dom} f_{i_0} \cap \mathrm{int}\left(\bigcap_{i\neq i_0} \mathrm{dom} f_i\right) \neq \emptyset$ for some $i_0 \in \{0, 1, \ldots, n\}$, then their conjugate functions $f_0^*, f_1^*, \ldots, f_n^*$ satisfy the relation

$$\inf_{x\in X}\left(\sum_{i=0}^{n} f_i(x)\right) = \max\left\{-\sum_{i=0}^{n} f_i^*(x_i^*) : \sum_{i=0}^{n} x_i^* = 0\right\}, \tag{1.1}$$

and, in fact, the following stronger relation holds for any $x^* \in X^*$:

$$\inf_{x\in X}\left\{\sum_{i=0}^{n} f_i(x) - \langle x^*, x\rangle\right\} = \max\left\{-\sum_{i=0}^{n} f_i^*(x_i^*) : \sum_{i=0}^{n} x_i^* = x^*\right\}. \tag{1.2}$$

Background information on the Fenchel duality theory can be found in Rockafellar [28] (see also [1, 2, 16, 27, 30]). This theory is a fundamental tool for establishing penalty results in nonlinear programming (cf. [8]). Moreover, it also plays an important role in the theory of best approximation (cf. [14, 20]), error bound analysis [12], in the study of monotone operators [25], and also in the KKT theory in connection with the following convex programming:

$$\begin{aligned} &\min_{x\in X} && f_0(x) \\ &\text{subject to (s.t.)} && f_i(x) \le 0 \quad (i = 1, \ldots, n). \end{aligned}$$

The Fenchel duality enables us to transform the original problem (primal problem) into an optimization problem on the dual space (dual problem). In some cases, especially in optimal control problems, the dual problems are easier to handle than the original ones (see [13, Example 25.2], [15]). Stimulated by the study of semi-infinite programming problems (see [17, 21] and the references therein), it is both interesting and useful to extend the Fenchel duality applicable to a family $\{f_i\}_{i\in I}$ of proper lower semicontinuous convex functions on a Banach space with the index set

*Received by the editors February 26, 2008; accepted for publication (in revised form) July 22, 2008; published electronically December 17, 2008. The research of the authors was supported by an Earmarked Grant from the Research Grant Council of Hong Kong.
http://www.siam.org/journals/siopt/19-3/71680.html
†Department of Mathematics, The Chinese University of Hong Kong, Shatin, New Territory, Hong Kong (gyli@math.cuhk.edu.hk, kfng@math.cuhk.edu.hk).

I, which is allowed to be infinite. In this present paper, much of our study is based on the consideration of the epigraphs of the conjugate functions and is motivated by the recent work of Jeyakumar and his collaborators (see [9, 10, 19], for example); we provide characterizations (and sufficient conditions) for the following property: For any $x^* \in X^*$,

$$\begin{aligned}&\inf_{x\in X}\{f(x)-\langle x^*,x\rangle\}\\ &=\max\left\{-\sum_{i\in I}f_i^*(x_i^*):\ x_i^*\in X^* \text{ and } \sum_{i\in I}\langle x_i^*,x\rangle=\langle x^*,x\rangle \text{ for any } x\in X\right\},\end{aligned}$$

where f is the sum function of $\{f_i : i \in I\}$, that is, $f(x) = \sum_{i\in I} f_i(x)$ for all $x \in X$. As an application, we present a fuzzy KKT condition in section 5 for the semi-infinite programming problem.

2. Preliminaries. Throughout this paper, X denotes a Banach space and X^* denotes its topological dual. We use $B(x, \epsilon)$ (resp. $\overline{B}(x, \epsilon)$) to denote the open (resp. closed) ball of X with center x and radius ϵ. For a set A in X, the interior (resp. relative interior, closure, convex hull, affine hull, linear span) of A is denoted by $\mathrm{int}A$ (resp. $\mathrm{ri}A$, $\overline{A}$, $\mathrm{co}A$, $\mathrm{aff}A$, $\mathrm{span}A$) (if A is a subset of X^*, its weak* closure is denoted by $\overline{A}^{w^*}$). Let A be a nonempty subset of X. The indicator function $\delta_A : X \to \mathbb{R}\cup\{+\infty\}$ and the support function $\sigma_A : X^* \to \mathbb{R}\cup\{+\infty\}$ of A are, respectively, defined by

$$\delta_A(x) := \begin{cases} 0 & \text{if} \quad x \in A,\\ +\infty & \text{otherwise,}\end{cases} \tag{2.1}$$

and $\sigma_A(x^*) = \sup_{x\in A}\langle x^*, x\rangle$ for all $x^* \in X^*$. Let $\Gamma(X)$ denote the class of proper lower semicontinuous convex functions on X, $\Gamma_c(X) := \{f \in \Gamma(X) : f$ is continuous and real-valued on $X\}$, and $\Gamma_+(X) := \{f \in \Gamma(X) : f$ is nonnegative on $X\}$. For a proper function f on X, the effective domain and the epigraph are, respectively, defined by $\mathrm{dom}f := \{x \in X : f(x) < +\infty\}$ and $\mathrm{epi}f := \{(x, r) \in X \times \mathbb{R} : f(x) \le r\}$. The subdifferential of f at $x \in X$ is defined by

$$\partial f(x) = \begin{cases} \{x^* \in X^* : \langle x^*, y - x\rangle \le f(y) - f(x) \text{ for all } y \in X\} & \text{if } x \in \mathrm{dom}f,\\ \emptyset & \text{otherwise.}\end{cases} \tag{2.2}$$

More generally, for any $\epsilon \ge 0$, the ϵ-subdifferential of f at $x \in X$ is defined by

$$\partial_\epsilon f(x) = \begin{cases} \{x^* \in X^* : \langle x^*, y - x\rangle \le f(y) - f(x) + \epsilon \text{ for all } y \in X\} & \text{if } x \in \mathrm{dom}f,\\ \emptyset & \text{otherwise.}\end{cases} \tag{2.3}$$

As usual, for a proper function f on X, its conjugate function $f^* : X^* \to \mathbb{R}\cup\{+\infty\}$ is defined by $f^*(x^*) = \sup_{x\in X}\{\langle x^*, x\rangle - f(x)\}$ for all $x^* \in X^*$. In particular, one has

$$(\delta_A)^*(x^*) = \sigma_A(x^*) \text{ for all } x^* \in X^*. \tag{2.4}$$

The definition of f^* entails that $\langle x^*, x\rangle \le f^*(x^*) + f(x)$ (Young's inequality) for any $x \in X$ and $x^* \in X^*$. Moreover, for any $\epsilon \ge 0$ and $x \in \mathrm{dom}f$,

$$x^* \in \partial_\epsilon f(x) \ \Leftrightarrow\ f^*(x^*)+f(x) \le \langle x^*, x\rangle+\epsilon \ \Leftrightarrow\ (x^*, \epsilon+\langle x^*, x\rangle-f(x)) \in \mathrm{epi}f^*. \tag{2.5}$$

In particular, we have the following Young's equality:

$$x^* \in \partial f(x) \ \Leftrightarrow\ \langle x^*, x\rangle = f^*(x^*) + f(x).$$

From the definitions, it is clear that, for any proper lower semicontinuous convex functions f_1, f_2 on X,

$$f_1 \le f_2 \;\Leftrightarrow\; f_1^* \ge f_2^* \;\Leftrightarrow\; \mathrm{epi} f_1^* \subseteq \mathrm{epi} f_2^*. \tag{2.6}$$

Moreover, it is known that $f^* \in \Gamma(X^*)$ for any $f \in \Gamma(X)$ (cf. [30, Theorem 2.3.3]). As usual, $X^* \times \mathbb{R}$ and $(X \times \mathbb{R})^*$ are identified and, for convenience, we use the norms defined by

$$\|(x, \alpha)\| = \max\{\|x\|, |\alpha|\} \text{ for all } (x, \alpha) \in X \times \mathbb{R}$$

and

$$\|(x^*, \alpha)\| = \|x^*\| + |\alpha| \text{ for all } (x^*, \alpha) \in X^* \times \mathbb{R}.$$

If H is a subspace of X, the restrictions and the corresponding norms of the restrictions are defined as follows: $x^*|_H \in H^*$, $(x^*|_H, \alpha) \in H^* \times \mathbb{R} = (H \times \mathbb{R})^*$, $\|x^*\|_H := \sup\{\langle x^*, x\rangle : x \in H, \|x\| \le 1\}$, and

$$\|(x^*|_H, \alpha)\| = \|x^*\|_H + |\alpha|. \tag{2.7}$$

Let I be an index set, and let $\mathcal{F}(I)$ denote the collection of all finite subsets of I (thus $\mathcal{F}(I)$ is a directed set ordered under the inclusion relation). Let $\{a_i : i \in I\} \subseteq \mathbb{R} \cup \{+\infty\}$. We define the sum of $\{a_i : i \in I\}$ by

$$\sum_{i \in I} a_i = \lim_{A \in \mathcal{F}(I)} \sum_{i \in A} a_i,$$

provided that the (unconditional) limit $\lim_{A \in \mathcal{F}(I)} \sum_{i \in A} a_i$ exists as a member of $\mathbb{R} \cup \{+\infty\}$. In particular, if $a_i \ge 0$ for all $i \in I$, then $\sum_{i \in I} a_i$ exists and

$$\sum_{i \in I} a_i = \sup_{A \in \mathcal{F}(I)} \sum_{i \in A} a_i \le +\infty. \tag{2.8}$$

Remark 2.1. Let $\{a_i, b_i, c_i\}_{i \in I} \subseteq \mathbb{R}$ be such that $a_i \le b_i \le c_i$ for all $i \in I$. Suppose that $\sum_{i \in I} a_i$ and $\sum_{i \in I} c_i$ exist in $\mathbb{R}$. Then $\sum_{i \in I} b_i$ also exists in $\mathbb{R}$ (because $0 \le b_i - a_i \le c_i - a_i$ and $\sum_{i \in I}(c_i - a_i) < +\infty$).

Let $\{f_i : i \in I\}$ be a family of extended real-valued functions on X. We define their sum function f as follows: Let $D_f := \{x \in X : \sum_{i \in I} f_i(x) \text{ exists in } \mathbb{R} \cup \{+\infty\}\}$; we define

$$f(x) = \sum_{i \in I} f_i(x) \text{ for all } x \in D_f.$$

In particular, if $f_i \in \Gamma_+(X)$ for all $i \in I$, then $D_f = X$ and

$$\left(\sum_{i \in I} f_i\right)(x) = \sup_{A \in \mathcal{F}(I)} \sum_{i \in A} f_i(x) \text{ for all } x \in X. \tag{2.9}$$

For $x^* \in X^*$ and a family $\{x_i^*\}_{i \in I}$ of elements in X^*, the notation

$$x^* = \sum_{i \in I}^{*} x_i^* \tag{2.10}$$

means that $\langle x^*, h\rangle = \lim_{A\in\mathcal{F}(I)} \sum_{i\in A}\langle x_i^*, h\rangle$ for each $h \in X$. Let $\{A_i\}_{i\in I}$ be a family of the subsets of X^*. The set $\{x^* \in X^* : \exists x_i^* \in A_i$ for all $i \in I$ such that $x^* = \sum_{i\in I}^* x_i^*\}$ will be denoted by $\sum_{i\in I}^* A_i$. It is easy to check that $\sum_{i\in I}^* A_i$ is convex if each A_i is convex and that $\sum_{i\in I}^* A_i = \sum_{i\in I} A_i$ if I is a finite set. Moreover, $\{A_i\}_{i\in I}$ is said to be weak* summable if $\sum_{i\in I}^* x_i^*$ exists in X^* (that is, (2.10) holds for some $x^* \in X^*$) whenever $x_i^* \in A_i$ for each $i \in I$.

Remark 2.2. The above definition is slightly different from [32]: Our notation $\sum_{i\in I}^* A_i$ does not require the family $\{A_i\}_{i\in I}$ to be weak* summable.

A useful relationship between $\mathrm{epi}f^*$ and $\partial_\epsilon f$ is given in the following formula observed by Burachik and Jeyakumar in [9] (we note that, as observed in [3], this formula works even when f is merely a proper function):

(2.11)
$$\mathrm{epi}f^* = \bigcup_{\epsilon\geq 0}\{(x^*, \epsilon + \langle x^*, x\rangle - f(x)) : x^* \in \partial_\epsilon f(x)\} \text{ for all } f \in \Gamma(X),\ x \in \mathrm{dom}f.$$

Throughout this paper, unless explicitly mentioned otherwise, I is an arbitrary index set (that is, the cardinality $|I| \leq +\infty$). For convenience, we list below several known results that will be useful for us.

LEMMA 2.1 (cf. [30]). *Let I be a finite set, and let $\{f, f_i : i \in I\} \subseteq \Gamma(X)$ be such that $f(x) = \sum_{i\in I} f_i(x)$ for all $x \in X$. Then $\mathrm{epi}f^* = \overline{\sum_{i\in I}\mathrm{epi}f_i^*}^{w^*}$, and, moreover, the result can be strengthened to $\mathrm{epi}f^* = \sum_{i\in I}\mathrm{epi}f_i^*$ if there exists $i_0 \in I$ such that $\mathrm{dom}f_{i_0} \cap \mathrm{int}\big(\bigcap_{i\neq i_0}(\mathrm{dom}f_i)\big) \neq \emptyset$.*

Remark 2.3. Let I be a finite set, and let C be a closed convex subset of X. Recall that $\mathrm{sqri}C := \{x \in C : \bigcup_{\lambda\geq 0}\lambda(C - x)$ is a closed subspace$\}$. A weaker generalized interior point regularity condition ensuring $\mathrm{epi}f^* = \sum_{i\in I}\mathrm{epi}f_i^*$ is as follows (cf. [5, 23]): There exists $i_0 \in I$ such that
$$0 \in \mathrm{sqri}\prod_{i\neq i_0}(\mathrm{dom}f_i - \mathrm{dom}f_{i_0}).$$

The following lemma can be found in [20, Lemma 2.3]. We note that it has been also derived in [4, section 4.3] via a different approach.

LEMMA 2.2. *Let $\{f_i : i \in I\} \subseteq \Gamma(X)$. Suppose that there exists $x_0 \in X$ such that $\sup_{i\in I} f_i(x_0) < \infty$. Then*
$$\mathrm{epi}(\sup_{i\in I} f_i)^* = \overline{\mathrm{co}\bigcup_{i\in I}\mathrm{epi}f_i^*}^{w^*},$$
where $\sup_{i\in I} f_i : X \to \mathbb{R}\cup\{+\infty\}$ is defined by $(\sup_{i\in I} f_i)(x) = \sup_{i\in I} f_i(x)$ for all $x \in X$.

Remark 2.4. Let $f \in \Gamma(X)$ and $A := \{x : f(x) \leq 0\} \neq \emptyset$. Then $\delta_A = \sup_{\lambda>0}\lambda f$, and it follows from Lemma 2.2 that

$$\mathrm{epi}(\delta_A)^* = \overline{\mathrm{co}\bigcup_{\lambda>0}\mathrm{epi}(\lambda f)^*}^{w^*} = \overline{\bigcup_{\lambda>0}\mathrm{epi}(\lambda f)^*}^{w^*}, \tag{2.12}$$

where the last equality holds because $\bigcup_{\lambda>0}\mathrm{epi}(\lambda f)^*$ is a convex set.

For continuous functions, the following result in [31] will play an important role.

LEMMA 2.3. *Let $\{f, f_i : i \in I\} \subseteq \Gamma_c(X)$ be such that $f(x) = \sum_{i\in I} f_i(x)$ for all $x \in X$. Then $\{\partial f_i(x)\}_{i\in I}$ is weak* summable, and the following relation holds:*
$$\partial f(x) = \overline{\sum_{i\in I}^*\partial f_i(x)}^{w^*} \qquad \text{for all } x \in X.$$

Moreover, if I is countable, then $\sum_{i\in I}{}^{*}\partial f_i(x)$ is weak closed, and hence*

$$\partial f(x) = \sum_{i\in I}{}^{*}\partial f_i(x) \quad \text{for all } x \in X.$$

3. Strong Fenchel duality and its characterization. In this section, we provide some characterization of the strong Fenchel duality (in the sense that (1.2) holds for all $x^* \in X^*$). To do this, we need the following lemma.

LEMMA 3.1. *Let $\{f, f_i : i \in I\} \subseteq \Gamma(X)$ be such that*

$$f(x) = \sum_{i\in I} f_i(x) \text{ for all } x \in X. \tag{3.1}$$

Then the following inclusion holds:

$$\overline{\sum_{i\in I}{}^{*}\operatorname{epi} f_i^*}^{w^*} \subseteq \operatorname{epi} f^*. \tag{3.2}$$

Proof. Let $(x^*, \alpha) \in \sum_{i\in I}{}^{*}\operatorname{epi} f_i^*$, that is, for each $i \in I$, there exists $(x_i^*, \alpha_i) \in X^* \times \mathbb{R}$, with $f_i^*(x_i^*) \le \alpha_i$ such that

$$\sum_{i\in I}\alpha_i = \alpha \text{ and } \sum_{i\in I}\langle x_i^*, x\rangle = \langle x^*, x\rangle \text{ for all } x \in X. \tag{3.3}$$

Since $\operatorname{epi} f^*$ is weak* closed, to prove (3.2), it suffices to show that $f^*(x^*) \le \alpha$. Let $x \in \operatorname{dom} f$. Note that

$$\langle x_i^*, x\rangle - f_i(x) \le \sup_{z\in X}\{\langle x_i^*, z\rangle - f_i(z)\} = f_i^*(x_i^*) \le \alpha_i. \tag{3.4}$$

Applying Remark 2.1 and making use of (3.1), (3.3), and (3.4), we note that $\sum_{i\in I} f_i^*(x_i^*)$ exists, and

$$\langle x^*, x\rangle - f(x) = \sum_{i\in I}\big(\langle x_i^*, x\rangle - f_i(x)\big) \le \sum_{i\in I} f_i^*(x_i^*) \le \sum_{i\in I}\alpha_i = \alpha.$$

Taking supremum over all x in $\operatorname{dom} f$, this implies that

$$f^*(x^*) = \sup_{x\in\operatorname{dom} f}\big(\langle x^*, x\rangle - f(x)\big) \le \sum_{i\in I} f_i^*(x_i^*) \le \alpha, \tag{3.5}$$

as required to show. This completes the proof. □

The following result is known [9] (see also [11, Corollary 3.4]) for the special case when I is finite.

THEOREM 3.2. *Let $\{f, f_i : i \in I\}$ be as in Lemma 3.1. Then the following statements are equivalent:*

(i)
$$\partial_\epsilon f(x) \subseteq \bigcup\left\{\sum_{i\in I}{}^{*}\partial_{\epsilon_i} f_i(x) : \sum_{i\in I}\epsilon_i = \epsilon, \text{ each } \epsilon_i \ge 0\right\} \text{ for all } \epsilon \ge 0 \text{ and } x \in X. \tag{3.6}$$

(ii) $\partial_\epsilon f(x) = \bigcup\big\{\sum_{i\in I}{}^{*}\partial_{\epsilon_i} f_i(x) : \sum_{i\in I}\epsilon_i = \epsilon, \text{ each } \epsilon_i \ge 0\big\}$ *for all* $\epsilon \ge 0$ *and* $x \in \operatorname{dom} f$.

(iii) $\mathrm{epi} f^* = \sum_{i\in I}{}^* \mathrm{epi} f_i^*$.
(iv) *For any* $x^* \in X^*$,

$$\inf_{x\in X}\{f(x) - \langle x^*, x\rangle\} = \max\left\{-\sum_{i\in I} f_i^*(x_i^*) : \sum_{i\in I}{}^* x_i^* = x^*\right\},$$

that is, $f^*(x^*) = \min\left\{\sum_{i\in I} f_i^*(x_i^*) : \sum_{i\in I}{}^* x_i^* = x^*\right\}$.
Any of the statements (i)–(iv) *implies that*
(v) $\inf_{x\in X} f(x) = \max\left\{-\sum_{i\in I} f_i^*(x_i^*) : \sum_{i\in I}{}^* x_i^* = 0\right\}$.

Proof. First, (v) follows from (iv) by letting $x^* = 0$. Thus, we need only to show the equivalence of (i)–(iv).

[(i) $\Rightarrow$ (ii)] Let $x \in \mathrm{dom} f$, $\epsilon \geq 0$, and $\epsilon_i \geq 0$ be such that $\sum_{i\in I} \epsilon_i = \epsilon$. To prove (i) $\Rightarrow$ (ii), it suffices to show that

$$\sum_{i\in I}{}^* \partial_{\epsilon_i} f_i(x) \subseteq \partial_\epsilon f(x). \tag{3.7}$$

To do this, let $x^* = \sum_{i\in I}{}^* x_i^* \in X^*$, where each $x_i^* \in \partial_{\epsilon_i} f_i(x)$. Then, from Young's inequality and (2.5), we have

$$\langle x_i^*, x\rangle - \epsilon_i \leq f_i^*(x_i^*) + f_i(x) - \epsilon_i \leq \langle x_i^*, x\rangle.$$

Therefore, by Remark 2.1, $\sum_{i\in I}(f_i^*(x_i^*) + f_i(x) - \epsilon_i)$ exists in $\mathbb{R}$ and

$$\sum_{i\in I} f_i^*(x_i^*) + f(x) - \epsilon = \sum_{i\in I}\left(f_i^*(x_i^*) + f_i(x) - \epsilon_i\right) \leq \sum_{i\in I}\langle x_i^*, x\rangle = \langle x^*, x\rangle. \tag{3.8}$$

On the other hand, note that $f^*(x^*) \leq \sum_{i\in I} f_i^*(x_i^*)$ because, for each $z \in \mathrm{dom} f$, one has

$$\langle x^*, z\rangle - f(z) = \sum_{i\in I}\left(\langle x_i^*, z\rangle - f_i(z)\right) \leq \sum_{i\in I} f_i^*(x_i^*). \tag{3.9}$$

Thus, by (3.8),

$$f^*(x^*) + f(x) - \epsilon \leq \sum_{i\in I} f_i^*(x_i^*) + f(x) - \epsilon \leq \langle x^*, x\rangle.$$

Therefore, $x^* \in \partial_\epsilon f(x)$, and (3.7) holds.

[(ii) $\Rightarrow$ (iii)] In view of Lemma 3.1, it suffices to show that $\mathrm{epi} f^* \subseteq \sum_{i\in I}{}^* \mathrm{epi} f_i^*$. To do this, let $(x^*, \alpha) \in \mathrm{epi} f^*$. We have to show that $(x^*, \alpha) \in \sum_{i\in I}{}^* \mathrm{epi} f_i^*$. Take an arbitrary $x \in \mathrm{dom} f$; from (2.11), there exists $\epsilon \geq 0$ such that $x^* \in \partial_\epsilon f(x)$ and $\alpha = \epsilon + \langle x^*, x\rangle - f(x)$. It follows from (ii) that there exist $\epsilon_i \geq 0$ and $x_i^* \in \partial_{\epsilon_i} f_i(x)$ $\big($so $(x_i^*, \alpha_i) \in \mathrm{epi} f_i^*$, where $\alpha_i := \epsilon_i + \langle x_i^*, x\rangle - f_i(x)\big)$ such that $\epsilon = \sum_{i\in I} \epsilon_i$ and $x^* = \sum_{i\in I}{}^* x_i^*$. Thus

$$(x^*, \alpha) = \sum_{i\in I}{}^* (x_i^*, \alpha_i) \in \sum_{i\in I}{}^* \mathrm{epi} f_i^*,$$

as required to show.

[(iii) $\Rightarrow$ (iv)] Let $x^* \in X^*$. Note first that, by (3.9),

$$-f^*(x^*) = \inf_{z\in \mathrm{dom} f}\{f(z) - \langle x^*, z\rangle\} \geq -\sum_{i\in I} f_i^*(x_i^*) \tag{3.10}$$

whenever $\{x_i^* : i \in I\} \subseteq X^*$, with $x^* = \sum_{i\in I}{}^* x_i^*$. Thus, to prove (iv), it remains to show that there exists $x_i^* \in X^*$ $(i \in I)$ such that $x^* = \sum_{i\in I} x_i^*$ and

$$\inf_{z\in \mathrm{dom} f}\{f(z) - \langle x^*, z\rangle\} \le -\sum_{i\in I} f_i^*(x_i^*). \tag{3.11}$$

To do this, we can suppose that $\inf_{z\in\mathrm{dom}f}\{f(z) - \langle x^*, z\rangle\} > -\infty$, that is, $f^*(x^*) < +\infty$. Then $(x^*, f^*(x^*)) \in \mathrm{epi}f^*$. It follows from (iii) that $(x^*, f^*(x^*)) \in \sum_{i\in I}{}^* \mathrm{epi}f_i^*$, that is, there exist $(x_i^*, \alpha_i) \in \mathrm{epi}f_i^*$ $(i \in I)$ such that

$$\sum_{i\in I}{}^* x_i^* = x^* \ \text{ and } \ \sum_{i\in I}\alpha_i = f^*(x^*). \tag{3.12}$$

We claim that $\{x_i^* : i \in I\}$ satisfies (3.11). In fact, since $(x_i^*, \alpha_i) \in \mathrm{epi}f_i^*$ $(i \in I)$, Young's inequality implies that, for any $z \in X$,

$$\langle x_i^*, z\rangle - f_i(z) \le f_i^*(x_i^*) \le \alpha_i \ (i \in I). \tag{3.13}$$

Since $\sum_{i\in I} f_i(z) = f(z) \in \mathbb{R}$ if $z \in \mathrm{dom}f$, it follows from (3.12) and Remark 2.1 that $\sum_{i\in I} f_i^*(x_i^*)$ exists and, for any $z \in \mathrm{dom}f$,

$$\langle x^*, z\rangle - f(z) = \sum_{i\in I}\big(\langle x_i^*, z\rangle - f_i(z)\big) \le \sum_{i\in I} f_i^*(x_i^*) \le \sum_{i\in I}\alpha_i = f^*(x^*).$$

Taking supremum over all $z \in \mathrm{dom}f$, this implies that $f^*(x^*) \le \sum_{i\in I} f_i^*(x_i^*) \le \sum_{i\in I}\alpha_i = f^*(x^*)$. In view of (3.13), this forces that $f_i^*(x_i^*) = \alpha_i$ for all $i \in I$. Therefore, we obtain that

$$\inf_{z\in\mathrm{dom}f}\{f(z) - \langle x^*, z\rangle\} = -f^*(x^*) = -\sum_{i\in I}\alpha_i = -\sum_{i\in I} f_i^*(x_i^*).$$

Thus (3.11) holds as claimed.

[(iv) $\Rightarrow$ (i)] Let $\epsilon \ge 0$, $x \in X$, and $x^* \in \partial_\epsilon f(x)$. By the definition of $f^*(x^*)$, (iv) means that

$$f^*(x^*) = \min\left\{\sum_{i\in I} f_i^*(x_i^*) : \sum_{i\in I}{}^* x_i^* = x^*\right\}.$$

Thus, there exist $x_i^* \in X^*$, with $\sum_{i\in I}{}^* x_i^* = x^*$ such that $f^*(x^*) = \sum_{i\in I} f_i^*(x_i^*)$. Hence

$$f^*(x^*) + f(x) - \langle x^*, x\rangle = \sum_{i\in I}\big(f_i^*(x_i^*) + f_i(x) - \langle x_i^*, x\rangle\big),$$

where $0 \le f_i^*(x_i^*) + f_i(x) - \langle x_i^*, x\rangle$ for all $i \in I$ (by Young's inequality). Since $x^* \in \partial_\epsilon f(x)$ (that is $f^*(x^*) + f(x) - \langle x^*, x\rangle \le \epsilon$), it follows that there exist $\epsilon_i \ge 0$ $(i \in I)$ such that $\sum_{i\in I}\epsilon_i = \epsilon$ and

$$f_i^*(x_i^*) + f_i(x) - \langle x_i^*, x\rangle \le \epsilon_i \quad \text{for all } i \in I.$$

Then $x_i^* \in \partial_{\epsilon_i} f_i(x)$ $(i \in I)$ and $x^* \in \sum_{i\in I}{}^* \partial_{\epsilon_i} f_i(x)$ (as $x^* = \sum_{i\in I}{}^* x_i^*$). Therefore, x^* belongs to the set on the right-hand side of (i). This completes the proof. ∎

Note 3.1. The property (v), sometimes referred as the Fenchel duality, is strictly weaker (even when $|I| = 2$) than the properties (i)–(iv) listed in Theorem 3.2. Examples can be found in [5, pp. 2798–2799] and [26, Example 11.1 and Example 11.3].

COROLLARY 3.3 (an extension of the Fenchel duality). *Let* $\{f_i, h, f : i \in I \cup J\} \subseteq \Gamma(X)$, *with* $I \cap J = \emptyset$, $|J| < +\infty$, *and*

$$h(x) = \sum_{i\in I} f_i(x) \text{ and } f(x) = \sum_{i\in I} f_i(x) + \sum_{j\in J} f_j(x) \text{ for all } x \in X.$$

Suppose that

$$\operatorname{epi} h^* = \sum_{i\in I}^{*} \operatorname{epi} f_i^*, \tag{3.14}$$

and (at least) one of the following conditions holds:

$$\text{(i)}\ \operatorname{dom} h \cap \operatorname{int}\Big(\bigcap_{j\in J} \operatorname{dom} f_j\Big) \neq \emptyset. \tag{3.15}$$

(ii) *There exists* $j_0 \in J$ *such that*

$$\operatorname{int}(\operatorname{dom} h) \cap \operatorname{dom} f_{j_0} \cap \operatorname{int}\Big(\bigcap_{j\in J\setminus\{j_0\}} \operatorname{dom} f_j\Big) \neq \emptyset. \tag{3.16}$$

Then

$$\operatorname{epi} f^* = \sum_{i\in I}^{*} \operatorname{epi} f_i^* + \sum_{j\in J} \operatorname{epi} f_j^*, \tag{3.17}$$

and, in particular, one has

$$\inf_{x\in X} f(x) = \max\left\{ -\sum_{i\in I} f_i^*(x_i^*) - \sum_{j\in J} f_j^*(y_j^*) : \sum_{i\in I}^{*} x_i^* + \sum_{j\in J} y_j^* = 0 \right\}. \tag{3.18}$$

Proof. First, from the implication (iii) $\Rightarrow$ (v) in Theorem 3.2, we need only to show (3.17). Since $\{f_j, h, f\ : j \in J\} \subseteq \Gamma(X)$ and $f = h + \sum_{j\in J} f_j$, Lemma 2.1 implies that

$$\operatorname{epi} f^* = \operatorname{epi} h^* + \sum_{j\in J} \operatorname{epi} f_j^*,$$

provided that (i) or (ii) holds. Consequently, (3.17) holds by (3.14). □

4. Sufficient conditions. This section is devoted to providing sufficient conditions ensuring that, for $\{f_i, f : i \in I\} \subseteq \Gamma(X)$, $\operatorname{epi} f^* = \sum_{i\in I}^{*} \operatorname{epi} f_i^*$ (see Theorem 3.2 (iii)), where

$$f(x) = \sum_{i\in I} f_i(x) \text{ for all } x \in X. \tag{4.1}$$

4.1. Continuous type. Throughout this subsection, we assume that f and each f_i are continuous, that is,

$$\{f_i, f : i \in I\} \subseteq \Gamma_c(X). \tag{4.2}$$

THEOREM 4.1. *Assume* (4.1) *and* (4.2). *Then*

$$\mathrm{epi} f^* = \overline{\sum\nolimits_{i\in I}^{*} \mathrm{epi} f_i^*}^{w^*}. \tag{4.3}$$

Proof. Let $x \in X$. By continuity, each $\partial f_i(x) \neq \emptyset$; take $x_i^* \in \partial f_i(x)$. By Lemma 2.3, there exists $x^* \in \partial f(x)$ such that $x^* = \sum_{i\in I}{}^{*} x_i^*$. Denote $r := \langle x^*, x\rangle - f(x)$ and $r_i = \langle x_i^*, x\rangle - f_i(x)$. It follows from (4.1) that $r = \sum_{i\in I} r_i$. Moreover, by (2.5), each $(x_i^*, r_i) \in \mathrm{epi} f_i^*$, and so $(x^*, r) \in \sum_{i\in I}{}^{*} \mathrm{epi} f_i^*$. Therefore, by Lemma 3.1, $\emptyset \neq \sum_{i\in I}{}^{*} \mathrm{epi} f_i^* \subseteq \mathrm{epi} f^*$. Thus, since $\mathrm{epi} f^*$ is weak* closed, if (4.3) is not true, then there exists $(x^*, \alpha) \in \mathrm{epi} f^* \backslash \overline{\sum_{i\in I}{}^{*} \mathrm{epi} f_i^*}^{w^*}$. Recalling that a linear functional h on X^* is the form $h(x^*) = \langle a, x^*\rangle$ for some $a \in X$ if and only if h is continuous in the weak* topology of X^* (cf. [29, p. 112, Theorem 1], it follows from the separation theorem that there exists $(x_0, r_0) \in X \times \mathbb{R}$ such that

$$\sup\left\{\langle y^*, x_0\rangle + \beta r_0 : \ (y^*, \beta) \in \sum_{i\in I}^{*} \mathrm{epi} f_i^*\right\} < \langle x^*, x_0\rangle + \alpha r_0. \tag{4.4}$$

Considering $\beta > 0$ large, it follows that $r_0 \leq 0$. We claim that $r_0 < 0$. Indeed, if $r_0 = 0$, then (4.4) means $\sup\{\langle y^*, x_0\rangle : (y^*, \beta) \in \sum_{i\in I}{}^{*} \mathrm{epi} f_i^*\} < \langle x^*, x_0\rangle$. Since $x^* \in \mathrm{dom} f^*$ and $\mathrm{Im}\, \partial f$ is norm dense in $\mathrm{dom} f^*$ (cf. [24, Theorem 3.18]), there exist $a^* \in \mathrm{Im}\, \partial f$ (so $a^* \in \partial f(a)$ for some $a \in X$) such that $\sup\left\{\langle y^*, x_0\rangle : \ (y^*, \beta) \in \sum_{i\in I}{}^{*} \mathrm{epi} f_i^*\right\} < \langle a^*, x_0\rangle$. By Lemma 2.3, this implies that

$$\sup\left\{\langle y^*, x_0\rangle : \ (y^*, \beta) \in \sum_{i\in I}^{*} \mathrm{epi} f_i^*\right\} < \langle a_0^*, x_0\rangle \tag{4.5}$$

for some $a_0^* \in \sum_{i\in I}{}^{*} \partial f_i(a)$. Note that a_0^* can be expressed in the form $a_0^* = \sum_{i\in I}{}^{*} a_i^*$, with each $a_i^* \in \partial f_i(a)$. Since each $\langle a_i^*, a\rangle = f_i(a) + f_i^*(a_i^*)$ (Young's equality), it follows from (4.1) that $\langle a_0^*, a\rangle = f(a) + \sum_{i\in I} f_i^*(a_i^*)$, and hence that $(a_0^*, \beta_0) \in \sum_{i\in I}{}^{*} \mathrm{epi} f_i^*$, where $\beta_0 := \langle a_0^*, a\rangle - f(a) \in \mathbb{R}$. But then $\sup\left\{\langle y^*, x_0\rangle : \ (y^*, \beta) \in \sum_{i\in I}{}^{*} \mathrm{epi} f_i^*\right\} \geq \langle a_0^*, x_0\rangle$, contradicting (4.5). Henceforth, without loss of generality, we may assume that $r_0 = -1$. Then (4.4) becomes

$$\sup\left\{\langle y^*, x_0\rangle - \beta : \ (y^*, \beta) \in \sum_{i\in I}^{*} \mathrm{epi} f_i^*\right\} < \langle x^*, x_0\rangle - \alpha. \tag{4.6}$$

Note that $\langle x^*, x_0\rangle - \alpha \leq f(x_0)$ by Young's inequality and the fact that $(x^*, \alpha) \in \mathrm{epi} f^*$, and it follows from (4.6) that

$$\sup\left\{\langle y^*, x_0\rangle - \beta : \ (y^*, \beta) \in \sum_{i\in I}^{*} \mathrm{epi} f_i^*\right\} < f(x_0). \tag{4.7}$$

Moreover, for each $i \in I$, pick $x_i^* \in \partial f_i(x_0)$. Define $x_0^* := \sum_{i\in I}{}^{*} x_i^*$ (this is well-defined by Lemma 2.3). Let $\alpha_0 := \langle x_0^*, x_0\rangle - f(x_0)$. Note from Young's equality that $\langle x_i^*, x_0\rangle =$

$f_i(x_0) + f_i^*(x_0^*)$ for each $i \in I$, and it follows from (4.1) that $\alpha_0 = \sum_{i\in I} f_i^*(x_i^*)$, and hence that $(x_0^*, \alpha_0) = \sum_{i\in I}{}^*(x_i^*, f_i^*(x_i^*)) \in \sum_{i\in I}{}^* \mathrm{epi} f_i^*$. Consequently, by (4.7), $\langle x_0^*, x_0\rangle - \alpha_0 < f(x_0)$, contradicting the definition of α_0. □

If I is countable and if another assumption, namely,

$$\mathrm{dom}\, f^* = \mathrm{Im}\, \partial f \tag{4.8}$$

is added, the following result shows that the set $\sum_{i\in I}{}^* \mathrm{epi} f_i^*$ is weak* closed.

THEOREM 4.2. *Assume* (4.8) *in addition to* (4.1) *and* (4.2)*, and suppose that I is countable. Then* $\mathrm{epi} f^* = \sum_{i\in I}{}^* \mathrm{epi} f_i^*$.

Proof. Noting that $\mathrm{epi} f^* = \mathrm{gph} f^* + \{0\} \times [0, +\infty)$ and,

$$\sum_{i\in I}{}^* \mathrm{epi} f_i^* + \{0\} \times [0, \infty) \subseteq \sum_{i\in I}{}^* \mathrm{epi} f_i^* \tag{4.9}$$

(because $\mathrm{epi} f_i^* + \{0\} \times [0, \infty) = \mathrm{epi} f_i^*$ for each i) and making use of Theorem 4.1, we need only to show that

$$\mathrm{gph} f^* \subseteq \sum_{i\in I}{}^* \mathrm{epi} f_i^*, \tag{4.10}$$

where $\mathrm{gph} f^*$ denotes the graph of f^*. To see (4.10), let $(x^*, \alpha) \in \mathrm{gph} f^*$. Then $x^* \in \mathrm{dom} f^* = \mathrm{Im}\, \partial f$ thanks to (4.8). Hence there exists $x \in X$ such that $x^* \in \partial f(x)$. By Lemma 2.3, x^* can be expressed in the form

$$x^* = \sum_{i\in I}{}^* x_i^*,$$

where each $x_i^* \in \partial f_i(x)$. By Young's equality, $f^*(x^*) = \langle x^*, x\rangle - f(x)$ and each $f_i^*(x_i^*) = \langle x_i^*, x\rangle - f_i(x)$, and it follows from (4.1) that $\sum_{i\in I} f_i^*(x_i^*) = \langle x^*, x\rangle - f(x)$, that is, $\sum_{i\in I} f_i^*(x_i^*) = f^*(x^*) = \alpha$. Therefore, $(x^*, \alpha) = \sum_{i\in I}{}^*(x_i^*, f_i^*(x_i^*)) \in \sum_{i\in I}{}^* \mathrm{epi} f_i^*$. This completes the proof. □

4.2. Nonnegative type. Throughout this subsection, we assume that f and each f_i are nonnegative-valued, that is,

$$\{f_i, f : i \in I\} \subseteq \Gamma_+(X). \tag{4.11}$$

THEOREM 4.3. *Assume* (4.1) *and* (4.11)*. Then*

$$\mathrm{epi} f^* = \overline{\bigcup_{\substack{J\subseteq I,\\ |J|<\infty}} \sum_{i\in J} \mathrm{epi} f_i^*}^{w^*} = \overline{\sum_{i\in I}{}^* \mathrm{epi} f_i^*}^{w^*}. \tag{4.12}$$

Proof. Since each $\mathrm{epi} f_i^*$ is a convex set containing the origin (because $f_i \geq 0$), one has from (2.6) and Lemma 3.1 that

$$\bigcup_{\substack{J\subseteq I,\\ |J|<\infty}} \sum_{i\in J} \mathrm{epi} f_i^* \subseteq \sum_{i\in I}{}^* \mathrm{epi} f_i^* \subseteq \mathrm{epi} f^*,$$

and hence

$$\overline{\bigcup_{\substack{J\subseteq I,\\ |J|<\infty}} \sum_{i\in J} \mathrm{epi} f_i^*}^{w^*} \subseteq \overline{\sum_{i\in I}{}^* \mathrm{epi} f_i^*}^{w^*} \subseteq \mathrm{epi} f^*. \tag{4.13}$$

For each $J \subseteq I$, with $|J| < \infty$, let g_J denote the sum function of $\{f_i : i \in J\}$, namely, $g_J(x) = \sum_{i\in J} f_i(x)$ for all $x \in X$. Since each f_i is nonnegative-valued, we have that, by (4.1) and (2.9),

$$f = \sum_{i\in I} f_i = \sup_{\substack{J\subseteq I,\\ |J|<\infty}} g_J. \tag{4.14}$$

Hence, by Lemma 2.2 (applied to $\{g_J : J \subseteq I, |J| < +\infty\}$) and Lemma 2.1, we have that

$$\operatorname{epi} f^* = \overline{\bigcup_{\substack{J\subseteq I,\\ |J|<\infty}} \operatorname{epi} g_J^*}^{w^*} = \overline{\bigcup_{\substack{J\subseteq I,\\ |J|<\infty}} \overline{\sum_{i\in J} \operatorname{epi} f_i^*}^{w^*}}^{w^*} \tag{4.15}$$

(note that $\bigcup_{\substack{J\subseteq I,\\ |J|<\infty}} \operatorname{epi} g_J^*$ is a convex set since $\operatorname{epi} g_{J_1}^* \subseteq \operatorname{epi} g_{J_2}^*$ if $J_1 \subseteq J_2$). Combining this with (4.13) and (4.15), we see that (4.12) holds because the set on the right-hand side of (4.15) is equal to that on the left-hand side of (4.13) (to see the latter fact, note that, for any $J \subseteq I$, with $|J| < +\infty$, one has

$$\overline{\sum_{i\in J} \operatorname{epi} f_i^*}^{w^*} \subseteq \overline{\bigcup_{\substack{J\subseteq I,\\ |J|<\infty}} \sum_{i\in J} \operatorname{epi} f_i^*}^{w^*},$$

and so

$$\overline{\bigcup_{\substack{J\subseteq I,\\ |J|<\infty}} \overline{\sum_{i\in J} \operatorname{epi} f_i^*}^{w^*}}^{w^*} \subseteq \overline{\bigcup_{\substack{J\subseteq I,\\ |J|<\infty}} \sum_{i\in J} \operatorname{epi} f_i^*}^{w^*}.$$

This completes the proof. □

Next, we seek some sufficient conditions to ensure that the set $\sum_{i\in I}^* \operatorname{epi} f_i^*$ in Theorem 4.3 is weak* closed. It would be convenient for us to introduce some new notation first. Let Y be a Banach space, and let J be a finite set. Let $\{K_i\}_{i\in J}$ be closed convex cones of Y. Following [28], we define $\gamma(K_i; J)$ by

$$\gamma(K_i; J) = \inf\left\{\left\|\sum_{i\in J} y_i\right\| : \sum_{i\in J} \|y_i\| = 1, \text{ each } y_i \in K_i\right\}. \tag{4.16}$$

When $J = \{1, 2\}$ and Y is a Hilbert space, the corresponding value of $\cos^{-1}\gamma(K_i; J)$ is termed as the angle between the closed convex cones K_1 and K_2 (see [7] for a detailed discussion). Given $y^* \in Y^*$ and any subspace Z of Y, $y^*|_Z$ denotes the restriction of y^* to Z and $\|y^*\|_Z$ denotes the corresponding norm of $y^*|_Z$ in Z^*. Furthermore, let $D \subseteq Y^*$; we define $D|_Z := \{y^*|_Z : y^* \in D\}$. Let K be a subset of Y (resp. Y^*), the (negative) polar of K is denoted by K° and is defined by $K^\circ = \{y^* \in Y^* : \langle y^*, y\rangle \le 0$ for all $y \in K\}$) (resp. $K^\circ = \{y \in Y : \langle y^*, y\rangle \le 0$ for all $y^* \in K\}$). From the definition, it is clear that if K_1 and K_2 are two subsets of Y (resp. Y^*) and $K_1 \subseteq K_2$, then $K_2^\circ \subseteq K_1^\circ$.

When H is a subspace of X, $Y = X^* \times \mathbb{R}$, $Z = H \times \mathbb{R}$, and each K_i ($i \in J$) is a weak* closed convex cone of Y, $K_i|_Z$ ($i \in J$) and $\gamma(K_i|_Z; J)$ are, respectively, defined by

$$K_i|_Z = \{(x^*|_H, \alpha) : (x^*, \alpha) \in K_i\} \tag{4.17}$$

and

(4.18)
$$\gamma(K_i|_Z; J) = \inf\left\{\left\|\sum_{i\in J}(x_i^*|_H, \alpha_i)\right\| : \sum_{i\in J}\|(x_i^*|_H, \alpha_i)\| = 1, \text{ each } (x_i^*, \alpha_i) \in K_i\right\}$$

(see (4.16)). If H is finite-dimensional, the infimum in (4.18) is attained and hence can be replaced by minimum.

An important special case (that we shall consider in the next theorem) is as follows: Each f_i is given in the form

$$f_i(x) = \max\{\langle a_i^*, x\rangle + r_i, 0\} + \delta_{C_i}(x), \tag{4.19}$$

where C_i are closed convex subsets of X with $\bigcap_{i\in I} C_i \neq \emptyset$ and $a_i^* \in X^*$ and $r_i \in \mathbb{R}$. Let D_i denote the convex hull of the set $(a_i^*, -r_i) \cup (0,0)$, and let K_i denote the set $\mathrm{epi}\sigma_{C_i}$. Then D_i is a weak* compact set in $X^* \times \mathbb{R}$ containing the origin, and K_i is a weak* closed convex cone in $X^* \times \mathbb{R}$. We observe that

$$\mathrm{co}\{\big(\{a_i^*\} \times [-r_i, \infty)\big) \cup \big(\{0\} \times [0, \infty)\big)\} = \mathrm{co}\{(a_i^*, -r_i) \cup (0,0)\} + \{0\} \times [0, \infty). \tag{4.20}$$

Indeed, let $(x^*, r) \in \mathrm{co}\{\big(\{a_i^*\} \times [-r_i, \infty)\big) \cup \big(\{0\} \times [0, \infty)\big)\}$. There exist $t \in [0,1]$, $\epsilon, \delta \geq 0$ such that $(x^*, r) = t(a_i^*, -r_i + \epsilon) + (1-t)(0, \delta) = t(a_i^*, -r_i) + (0, t\epsilon + (1-t)\delta)$. Note that $t\epsilon + (1-t)\delta \geq 0$. It follows that $(x^*, r) \in \mathrm{co}\{(a_i^*, -r_i) \cup (0,0)\} + \{0\} \times [0, \infty)$, and hence $\mathrm{co}\{\big(\{a_i^*\} \times [-r_i, \infty)\big) \cup \big(\{0\} \times [0, \infty)\big)\} \subseteq \mathrm{co}\{(a_i^*, -r_i) \cup (0,0)\} + \{0\} \times [0, \infty)$. As the converse inclusion can be verified similarly, (4.20) is seen to hold. Consequently, we have that

$$\begin{aligned}
\mathrm{epi} f_i^* &= \mathrm{epi}(\max\{\langle a_i^*, \cdot\rangle + r_i, 0\})^* + \mathrm{epi}(\delta_{C_i})^* \\
&= \mathrm{co}\{\mathrm{epi}(\langle a_i^*, \cdot\rangle + r_i)^* \cup \big(\{0\} \times [0, \infty)\big)\} + \mathrm{epi}(\delta_{C_i})^* \\
&= \mathrm{co}\{\big(\{a_i^*\} \times [-r_i, \infty)\big) \cup \big(\{0\} \times [0, \infty)\big)\} + \mathrm{epi}(\delta_{C_i})^* \\
&= \mathrm{co}\{(a_i^*, -r_i) \cup (0,0)\} + \mathrm{epi}(\delta_{C_i})^* \\
&= D_i + K_i, \qquad (4.21)
\end{aligned}$$

where the first equality follows from (4.19) and Lemma 2.1, the second equality follows from Lemma 2.3, and the fourth equality holds by (4.20) and the fact $\mathrm{epi}(\delta_{C_i})^* = \mathrm{epi}(\delta_{C_i})^* + \{0\} \times [0, +\infty)$. Therefore, the condition (C1) in the following theorem is satisfied if the functions f_i are given in the form (4.19).

THEOREM 4.4. *Let I be a compact metric space. Assume* (4.1), (4.11), *and the following assumptions:*

(C1) *For each $i \in I$, there exist a weak* compact convex set D_i in $X^* \times \mathbb{R}$ containing the origin, and a weak* closed convex cone K_i in $X^* \times \mathbb{R}$ such that*

$$\mathrm{epi} f_i^* = D_i + K_i. \tag{4.22}$$

(C2) $\sum_{i\in I} \mathrm{diam}(D_i) < \infty$, *where* $\mathrm{diam}(D_i)$ *denotes the diameter of D_i ($i \in I$), i.e.,* $\mathrm{diam}(D_i) := \sup\{\|x - y\| : x, y \in D_i\}$.

(C3) *There exist $i_0 \in I$ and a finite-dimensional subspace H of X such that $K_{i_0}^\circ \subseteq Z := H \times \mathbb{R}$ (denote the corresponding dimension of Z by m).*

(C4) *For any $J \subseteq I$ with $|J| = m$, $\gamma(K_i|_Z; J) > 0$.*

(C5) *The set-valued mapping $i \mapsto K_i|_Z$ is upper semicontinuous, i.e., for any $\overline{i} \in I$,*

$$\limsup_{i\to\overline{i}}(K_i|_Z) \subseteq K_{\overline{i}}|_Z,$$

where $\limsup_{i\to\overline{i}}(K_i|_Z) := \{x^* \in Z^* : \exists\ x_i^* \in K_i|_Z$ *such that* $x^* = \lim_{i\to\overline{i}} x_i^*$ *(in the norm of* Z^**)*$\}$. *Then* $\sum_{i\in I}{}^*\mathrm{epi}f_i^*$ *is weak** *closed and*

$$\mathrm{epi}f^* = \sum_{i\in I}{}^*\mathrm{epi}f_i^*. \tag{4.23}$$

Proof. By Theorem 4.3, we need only prove the weak*-closedness assertion. Denote $Y := X \times \mathbb{R}$, and so Y^* is identified with $X^* \times \mathbb{R}$. Denote $A_i := \mathrm{epi}f_i^* \subseteq Y^*$ and $A := \sum_{i\in I}{}^* A_i$. Let $a^* \in \overline{A}^{w^*}$. We have to show that $a^* \in A$. To do this, we take a sequence $\{a_k^*\} \subseteq A$ such that $a_k^* \to a^*$ on $Z := H \times \mathbb{R}$ (thanks to the assumption that H is finite-dimensional and the weak* topology coincides with the norm topology on a finite-dimensional space). For each $k \in \mathbb{N}$, noting that $a_k^* \in A = \sum_{i\in I}{}^* A_i$, there exists a sequence in $\bigcup_{\substack{J\subseteq I,\\ |J|<\infty}} \sum_{i\in J} A_i$ weak* converging (and hence in norm $\|\cdot\|_Z$) to a^*. Thus, there exist a finite subset I_k of I and $a_{i,k}^* \in A_i$ $(i \in I_k)$ such that

$$\left\| a_k^* - \sum_{i\in I_k} a_{i,k}^* \right\|_Z \le \frac{1}{k}. \tag{4.24}$$

Hence

$$\lim_{k\to\infty} \|u_k^* - a^*\|_Z \to 0, \tag{4.25}$$

where $u_k^* := \sum_{i\in I_k} a_{i,k}^*$. Note that $u_k^* \in \sum_{i\in I_k} D_i + \sum_{i\in I_k} K_i$ $\big($by (4.22)$\big)$. Since Z is of dimension m and each K_i is a (convex) cone, it follows from the Carathédory theorem [27, Corollary 17.1.2] that, for each $k \in \mathbb{N}$, there exist $\{i_{1,k}, i_{2,k}, \ldots, i_{m,k}\} \subseteq I_k$, such that

$$u_k^* = \sum_{i\in I_k} \overline{y}_{i,k}^* + \sum_{j=1}^m \overline{z}_{i_{j,k}}^* \text{ on } Z \tag{4.26}$$

for some $\overline{y}_{i,k}^* \in D_i$ $(i \in I_k)$ and $\overline{z}_{i_{j,k}}^* \in K_{i_{j,k}}$ $(1 \le j \le m)$. Let $I' := \bigcup_{k\in\mathbb{N}} I_k$ and set $\overline{y}_{i,k}^* := 0$ for any $i \in I'\backslash I_k$. For each fixed $k \in \mathbb{N}$, it follows from (4.26) that

$$u_k^* = \sum_{i\in I'} \overline{y}_{i,k}^* + \sum_{j=1}^m \overline{z}_{i_{j,k}}^* \text{ on } Z, \tag{4.27}$$

where $\overline{y}_{i,k}^* \in D_i$ for all $i \in I'$ $\big($thanks to the assumption that each D_i contains the origin$\big)$. Next, we show that

$$\{\overline{z}_{i_{j,k}}^*|_Z\}_{k\in\mathbb{N}} \text{ are bounded sequences for all } 1 \le j \le m. \tag{4.28}$$

To prove this, we suppose on the contrary that $\{\overline{z}_{i_{j,k}}^*|_Z\}_{k\in\mathbb{N}}$ is an unbounded sequence for some $j \in \{1, 2, \ldots, m\}$. By passing to a subsequence if necessary, we may assume that

$$\lim_{k\to\infty} \sum_{j=1}^m \|\overline{z}_{i_{j,k}}^*\|_Z \to \infty. \tag{4.29}$$

Dividing by $\sum_{j=1}^m \|\overline{z}^*_{i_{j,k}}\|_Z$ on both sides of (4.27), we obtain

$$\frac{u^*_k}{\sum_{j=1}^m \|\overline{z}^*_{i_{j,k}}\|_Z} = \frac{\sum_{i\in I'} \overline{y}^*_{i,k}}{\sum_{j=1}^m \|\overline{z}^*_{i_{j,k}}\|_Z} + \sum_{j=1}^m \frac{\overline{z}^*_{i_{j,k}}}{\sum_{j=1}^m \|\overline{z}^*_{i_{j,k}}\|_Z} \text{ on } Z. \tag{4.30}$$

Note that $\{\|u^*_k\|_Z\}_{k\in\mathbb{N}}$ is a bounded numerical sequence (since $\|u^*_k - a^*\|_Z \to 0$), and

$$\left\|\sum_{i\in I'} \overline{y}^*_{i,k}\right\|_Z \le \sum_{i\in I'} \|\overline{y}^*_{i,k}\| \le \sum_{i\in I} \operatorname{diam}(D_i) < \infty. \tag{4.31}$$

Moreover, since I is compact, we may assume without loss of generality that $i_{j,k} \to \overline{i}_j$ for some $\overline{i}_j \in I$ $(1 \le j \le m)$ as $k \to \infty$. Considering subsequences if necessary, we may assume that there exists $z^*_j \in X^*$ such that the bounded sequence

$$\frac{\overline{z}^*_{i_{j,k}}}{\sum_{j=1}^m \|\overline{z}^*_{i_{j,k}}\|_Z} \to z^*_j \text{ on } Z \quad (1 \le j \le m) \tag{4.32}$$

(thanks to the fact that Z is finite-dimensional). By (4.32), it is clear that $\sum_{j=1}^m \|z^*_j\|_Z = 1$. Moreover, assumption (C5) entails that each $z^*_j|_Z \in K_{\overline{i}_j}|_Z$. Finally, by passing to the limits in (4.30) and making use of (4.31) and (4.29), we have $\sum_{j=1}^m z^*_j = 0$ on Z. Then $\gamma(K_i|_Z, \{\overline{i}_1, \overline{i}_2, \dots, \overline{i}_m\}) = 0$, contradicting assumption (C4). Therefore, (4.28) is proved.

By the compactness of I again and by passing to subsequences if necessary, we may assume that $i_{j,k} \to \hat{i}_j$ for some $\hat{i}_j \in I$ as $k \to \infty$ $(1 \le j \le m)$. For each j, since Z is finite-dimensional and by (4.28), we may assume that $\overline{z}^*_{i_{j,k}} \to \overline{z}^*_j$ on Z for some $\overline{z}^*_j \in X^*$. By (C5), $\overline{z}^*_j|_Z \in K_{\hat{i}_j}|_Z$, and so there exists $\omega^*_j \in K_{\hat{i}_j}$ such that $\overline{z}^*_j = \omega^*_j$ on Z. Hence, replacing $\overline{z}^*_j$ by ω^*_j if necessary, we may assume without loss of generality that

$$\overline{z}^*_j \in K_{\hat{i}_j} \quad (1 \le j \le m). \tag{4.33}$$

Since $u^*_k \to a^*$ on Z $\big($by (4.25)$\big)$, (4.27) implies that

$$\sum_{i\in I'} \overline{y}^*_{i,k} = u^*_k - \sum_{j=1}^m \overline{z}^*_{i_{j,k}} \to a^* - \sum_{j=1}^m \overline{z}^*_j \text{ on } Z \quad \text{as } k \to \infty. \tag{4.34}$$

Since I' is countable, we may represent I' in the form that $I' = \{i_1, \dots, i_n, \dots\}$, and hence

$$\sum_{n\in\mathbb{N}} \overline{y}^*_{i_n,k} \to a^* - \sum_{j=1}^m \overline{z}^*_j \text{ on } Z \quad \text{as } k \to \infty.$$

Since $\overline{y}^*_{i_1,k} \in D_{i_1}$ and $D_{i_1}|_Z$ is compact, there exists an infinite subset $N_1 \subseteq \mathbb{N}$ such that $\{\overline{y}^*_{i_1,k}\}_{k\in N_1}$ converges to $\overline{y}^*_{i_1}$ on Z for some $\overline{y}^*_{i_1} \in D_{i_1}$. Inductively, we can find a sequence of infinite subsets $N_n \subseteq \mathbb{N}$ such that $N_{n+1} \subseteq N_n$ and, for each $n \in \mathbb{N}$,

$$\{\overline{y}^*_{i_n,k}\}_{k\in N_n} \text{ converges to } \overline{y}^*_{i_n} \text{ on } Z \text{ for some } \overline{y}^*_{i_n} \in D_{i_n}. \tag{4.35}$$

Since $\sum_{n\in\mathbb{N}}\|\overline{y}^*_{i_n}\| \le \sum_{i\in I}\|\overline{y}^*_i\| \le \sum_{i\in I}\mathrm{diam}D_i < +\infty$ (by the assumption (C2)), $\sum_{n\in\mathbb{N}}\overline{y}^*_{i_n}$ exists as an element in X^*, and, in particular,

$$\sum_{n\in\mathbb{N}}\overline{y}^*_{i_n} \in \sum_{n\in\mathbb{N}}{}^* D_{i_n}. \tag{4.36}$$

Similarly, for all $k\in\mathbb{N}$,

$$\sum_{n\in\mathbb{N}}\|\overline{y}^*_{i_n,k}\| \le \sum_{i\in I}\mathrm{diam}D_i < +\infty,$$

and thus, for any $\epsilon>0$, there exists $n_0\in\mathbb{N}$ such that, for all $n'\ge n_0$,

$$\sum_{n>n'}\|\overline{y}^*_{i_n,k}\| \le \sum_{n>n'}\mathrm{diam}(D_{i_n}) \le \epsilon/2 \text{ for all } k\in\mathbb{N}. \tag{4.37}$$

Note that, for any fixed $n'\ge n_0$, (by (4.35) and (4.34)) there exists $k_0\in\mathbb{N}$ (depending on n') such that

$$\|\overline{y}^*_{i_n,k_0} - \overline{y}^*_{i_n}\|_Z \le \frac{\epsilon}{4n'} \text{ for all } n\in\{1,\dots,n'\} \quad \text{and}$$
$$\left\| a^* - \sum_{j=1}^m \overline{z}^*_j - \sum_{n\in\mathbb{N}}\overline{y}^*_{i_n,k_0}\right\|_Z \le \epsilon/4.$$

It follows from (4.37) that, for any $n'\ge n_0$,

$$\begin{aligned}\left\| a^* - \sum_{j=1}^m \overline{z}^*_j - \sum_{n=1}^{n'}\overline{y}^*_{i_n}\right\|_Z &\le \left\| a^* - \sum_{j=1}^m \overline{z}^*_j - \sum_{n\in\mathbb{N}}\overline{y}^*_{i_n,k_0}\right\|_Z \\ &\quad + \sum_{n=1}^{n'}\|\overline{y}^*_{i_n,k_0} - \overline{y}^*_{i_n}\|_Z + \sum_{n>n'}\|\overline{y}^*_{i_n,k_0}\|_Z \\ &\le \epsilon.\end{aligned}$$

Since ϵ is arbitrary, one has

$$a^* = \sum_{j=1}^m \overline{z}^*_j + \sum_{n\in\mathbb{N}}\overline{y}^*_{i_n} \text{ on } Z. \tag{4.38}$$

Then, one has

$$a^* - \sum_{j=1}^m \overline{z}^*_j - \sum_{n\in\mathbb{N}}\overline{y}^*_{i_n} \in Z^\perp. \tag{4.39}$$

From (4.33) and (4.35), we know that each $\overline{z}^*_j \in K_{\hat{i}_j}$ and $\overline{y}_{i_n}\in D_{i_n}$, and it follows from (4.36) that $\sum_{j=1}^m \overline{z}^*_j + \sum_{n\in\mathbb{N}}\overline{y}^*_{i_n} \in \sum_{j=1}^m K_{\hat{i}_j} + \sum_{n\in\mathbb{N}}{}^* D_{i_n}$. Therefore, (4.39) entails that

$$a^* \in \sum_{j=1}^m K_{\hat{i}_j} + \sum_{n\in\mathbb{N}}{}^* D_{i_n} + Z^\perp. \tag{4.40}$$

Note from the bipolar theorem (cf. [30, Theorem 1.1.9]) and assumption (C3) that $Z^\perp \subseteq K_{i_0}$. It follows that $a^* \in \sum_{j=1}^m K_{\hat{i}_j} + \sum_{n\in\mathbb{N}}{}^* D_{i_n} + K_{i_0}$. Since $0 \in D_i \cap K_i$, one has

$$a^* \in \sum_{j=1}^m K_{\hat{i}_j} + \sum_{n\in\mathbb{N}}{}^* D_{i_n} + K_{i_0} = \sum_{n\in\mathbb{N}}{}^* D_{i_n} + \left(\sum_{j=1}^m K_{\hat{i}_j} + K_{i_0}\right) \subseteq \sum_{i\in I}{}^* (D_i + K_i) = A,$$

as required to show. This completes the proof. □

COROLLARY 4.5. *Let I be a compact metric space, and let f_i be given by* (4.19)*. Assume* (4.1), (4.11)*, and the following assumptions:*

(B1) $\sum_{i\in I}(\|a_i^*\| + |r_i|) < \infty$.

(B2) *There exists $i_0 \in I$ such that $H := \operatorname{span}(C_{i_0})$ is of finite dimension. (Denote the corresponding dimension by n.)*

(B3) *For any $J \subseteq I\backslash\{i_0\}$ with $|J| \le n+1$, it holds that*

$$C_{i_0} \cap \left(\bigcap_{i\in J} \operatorname{int}_H C_i\right) \neq \emptyset,$$

where $\operatorname{int}_H C_i := \{x \in C_i :$ *there exists $\epsilon > 0$ s.t. $B(x,\epsilon) \cap H \subseteq C_i$*$\}$.

(B4) *The mapping $i \mapsto C_i \cap H$ is lower semicontinuous, i.e., for any $\overline{i} \in I$,*

$$C_{\overline{i}} \cap H \subseteq \liminf_{i\to\overline{i}} (C_i \cap H).$$

Then the conclusion of Theorem 4.4 *holds.*

Proof. Define $D_i := \operatorname{co}\{(a_i^*, -r_i) \cup (0,0)\}$, and $K_i := \operatorname{epi}\sigma_{C_i}$. Then, as we have mentioned before, assumption (C1) in Theorem 4.4 holds (see (4.21)). Therefore, to finish the proof, it suffices to show that the assumptions (C2)–(C5) in Theorem 4.4 are satisfied. First of all, from condition (B1), we see that $\sum_{i\in I} \operatorname{diam} D_i < \infty$, and hence assumption (C2) holds. To see (C3), noting that $Z^\perp = (\operatorname{span} C_{i_0})^\perp \times \{0\} \subseteq \operatorname{epi}\sigma_{C_{i_0}} = K_{i_0}$, one has $K_{i_0}^\circ \subseteq Z$. Therefore, assumption (C3) holds with $m := n+1$. To see that assumption (C4) holds with $m := n+1$, we proceed by contradiction and suppose that there exists $J_0 \subseteq I$, with $|J_0| = n+1$ such that $\gamma(\operatorname{epi}\sigma_{C_i}|_{H\times\mathbb{R}}; J_0) = 0$. Noting that $H \times \mathbb{R}$ is finite-dimensional, it follows from (4.18) that there exist $(x_i^*, \alpha_i) \in \operatorname{epi}\sigma_{C_i}$ $(i \in J_0)$ such that

$$\sum_{i\in J_0} \left(\|x_i^*\|_H + |\alpha_i|\right) = 1, \tag{4.41}$$

$$\sum_{i\in J_0} x_i^* = 0 \text{ on } H, \quad \text{and} \quad \sum_{i\in J_0} \alpha_i = 0. \tag{4.42}$$

We claim that

$$x_i^* = 0 \text{ on } H \text{ for all } i \in J_0. \tag{4.43}$$

Granting this, we have $0 = \sigma_{C_i\cap H}(x_i^*) \le \sigma_{C_i}(x_i^*) \le \alpha_i$ for all $i \in J_0$. This together with the second equality of (4.42) implies that $\alpha_i = 0$ for all $i \in J_0$. However, this and (4.43) contradict (4.41). To see (4.43), we divide our proof into two cases.

Case 1: $i_0 \notin J_0$,

Case 2: $i_0 \in J_0$.

For Case 1, we note that $J_0 \subseteq I\backslash\{i_0\}$, and it follows from (B3) that $\bigcap_{j\in J_0} \mathrm{int}_H C_i \neq \emptyset$. Thus there exist $x_0 \in X$ and $\epsilon > 0$ such that $x_0 \in \overline{\mathbb{B}}(x_0,\epsilon) \cap H \subseteq C_i$ for all $i \in J_0$. Recalling the fact that $(x_i^*, \alpha_i) \in \mathrm{epi}\sigma_{C_i}$ $(i \in J_0)$ and the definition of $\|\cdot\|_H$, it follows from (4.42) that

$$\begin{aligned}\epsilon \sum_{i\in J_0} \|x_i^*\|_H = \sum_{i\in J_0} \big(\langle x_i^*, x_0\rangle + \epsilon\|x_i^*\|_H\big) &\leq \sum_{i\in J_0} \sup_{x\in\overline{\mathbb{B}}(x_0,\epsilon)\cap H} \langle x_i^*, x\rangle \\ &\leq \sum_{i\in J_0} \sup_{x\in C_i} \langle x_i^*, x\rangle \\ &\leq \sum_{i\in J_0} \alpha_i = 0. \qquad (4.44)\end{aligned}$$

Thus (4.43) holds in this case. For Case 2, one applies (B3) again, and there exist $x_0 \in C_{i_0}$ and $\epsilon > 0$ such that $x_0 \in \overline{\mathbb{B}}(x_0,\epsilon) \cap H \subseteq C_i$ for all $i \in J_0\backslash\{i_0\}$. Hence $\sup_{x\in\overline{B}(x_0,\epsilon)\cap H}\langle x_i^*, x\rangle \leq \sup_{x\in C_i}\langle x_i^*, x\rangle$, that is,

$$\langle x_i^*, x_0\rangle + \epsilon\|x_i^*\|_H = \sup_{x\in\overline{\mathbb{B}}(x_0,\epsilon)\cap H} \langle x_i^*, x\rangle \leq \sigma_{C_i}(x_i^*) \text{ for all } i \in J_0\backslash\{i_0\}. \qquad (4.45)$$

Since $(x_i^*, \alpha_i) \in \mathrm{epi}\sigma_{C_i}$ for all $i \in J_0$ and $x_0 \in C_{i_0}$ (so $\langle x_{i_0}^*, x_0\rangle \leq \alpha_{i_0}$), it follows from (4.42) that

$$\epsilon \sum_{i\in J_0\backslash\{i_0\}} \|x_i^*\|_H = \sum_{i\in J_0} \langle x_i^*, x_0\rangle + \epsilon \sum_{i\in J_0\backslash\{i_0\}} \|x_i^*\|_H \leq \sum_{i\in J_0} \alpha_i = 0.$$

This together with the first equality in (4.42) gives that $x_i^* = 0$ on H for all $i \in J_0$. Thus (4.43) also holds in this case. Finally, for (C5), fix an $\overline{i} \in I$. Consider $i \to \overline{i}$ and $(x_i^*, \alpha_i) \in H^* \times \mathbb{R}$ $(i \in I)$ be such that $(x_i^*, \alpha_i) \in \mathrm{epi}\sigma_{C_i}|_{H\times\mathbb{R}}$, with $(x_i^*, \alpha_i) \to (x^*, \alpha)$ for some $(x^*, \alpha) \in H^* \times \mathbb{R}$. Let $x \in C_{\overline{i}} \cap H$. By (B4) and since H is of finite dimension, there exists a sequence $\{x_i\} \subseteq C_i \cap H$ such that $x_i \to x$. It follows that

$$\langle x^*, x\rangle = \lim_{i\to\overline{i}} \langle x_i^*, x_i\rangle \leq \limsup_{i\to\overline{i}} \sigma_{C_i\cap H}(x_i^*) \leq \lim_{i\to\overline{i}} \alpha_i = \alpha.$$

This implies that $(x^*, \alpha) \in \mathrm{epi}\sigma_{C_{\overline{i}}}|_{H\times\mathbb{R}}$, and hence (C5) in Theorem 4.4 holds. Therefore, the assumptions (C2)–(C5) in Theorem 4.4 hold. This finishes the proof. □

5. Application to the KKT theory. We first establish an ϵ-sum rule involving possibly infinitely many convex functions. In the special case where $I = \emptyset$, the following result has been presented in [18] (see also [30, Corollary 2.6.7]).

THEOREM 5.1. *Let I, J be two index sets with $I \cap J = \emptyset$ and $|J| < \infty$. Let $\{f_i\}_{i\in I} \subseteq \Gamma_+(X)$ and $\{f_j\}_{j\in J} \subseteq \Gamma(X)$. Let $f \in \Gamma(X)$ be such that $f(x) = \sum_{i\in I} f_i(x) + \sum_{j\in J} f_j(x)$ for each $x \in X$. Let $\epsilon \geq 0$ and $x \in X$. Then we have*

$$\partial_\epsilon f(x) \subseteq \bigcap_{\eta>0} \bigcup_{\substack{I'\subseteq I,\\ |I'|<\infty}} \overline{\left\{\sum_{i\in I'} \partial_{\epsilon_i} f_i(x) + \sum_{j\in J} \partial_{\epsilon_j} f_j(x) : \sum_{i\in I'} \epsilon_i + \sum_{j\in J} \epsilon_j \leq \epsilon + \eta\right\}}^{w^*}. \qquad (5.1)$$

Proof. To see the inclusion, let $x^* \in \partial_\epsilon f(x)$, $\eta > 0$, and let V be a weak* neighborhood of 0. It suffices to show that there exist I' with $|I'| < +\infty$, $\epsilon_i, \epsilon_j \geq 0$ $(i \in I', j \in J)$ such that

$$\sum_{i\in I'} \epsilon_i + \sum_{j\in J} \epsilon_j \leq \epsilon + \eta \qquad (5.2)$$

and

$$x^* \in \sum_{i\in I'} \partial_{\epsilon_i} f_i(x) + \sum_{j\in J} \partial_{\epsilon_j} f_j(x) + V. \tag{5.3}$$

Let $h := \sum_{i\in I} f_i$. Note that $h \in \Gamma_+(X)$ (since $f_i \in \Gamma_+(X)$ and $h(x_0) < +\infty$ for all $x_0 \in \mathrm{dom} f$). From Lemma 2.1 (applied to $\{h, f_j : j \in J\}$) and Theorem 4.3, we have

$$\begin{aligned}
\mathrm{epi} f^* = \mathrm{epi}\left(h + \sum_{j\in J} f_j\right)^* &= \overline{\mathrm{epi} h^* + \sum_{j\in J} \mathrm{epi} f_j^*}^{w^*} \\
&= \overline{\overline{\bigcup_{\substack{I'\subseteq I,\\ |I'|<\infty}} \sum_{i\in I'} \mathrm{epi} f_i^*}^{w^*} + \sum_{j\in J} \mathrm{epi} f_j^*}^{w^*} \\
&= \overline{\bigcup_{\substack{I'\subseteq I,\\ |I'|<\infty}} \sum_{i\in I'} \mathrm{epi} f_i^* + \sum_{j\in J} \mathrm{epi} f_j^*}^{w^*}.
\end{aligned}$$

Since $x^* \in \partial_\epsilon f(x)$, it follows from (2.5) and the above expression that

$$(x^*, \epsilon + \langle x^*, x\rangle - f(x)) \in \overline{\bigcup_{\substack{I'\subseteq I,\\ |I'|<\infty}} \sum_{i\in I'} \mathrm{epi} f_i^* + \sum_{j\in J} \mathrm{epi} f_j^*}^{w^*}, \tag{5.4}$$

and hence there exist $I' \subseteq I$ with $|I'| < \infty$, $(x_i^*, r_i) \in \mathrm{epi} f_i^*$ $(i \in I')$, and $(x_j^*, r_j) \in \mathrm{epi} f_j^*$ $(j \in J)$ such that

$$x^* \in \sum_{i\in I'} x_i^* + \sum_{j\in J} x_j^* + V \tag{5.5}$$

and

$$\sum_{i\in I'} r_i + \sum_{j\in J} r_j \le (\epsilon + \eta/2) + \langle x^*, x\rangle - f(x). \tag{5.6}$$

By shrinking V if necessary, we may assume without loss of generality that

$$\left|\left\langle \sum_{i\in I'} x_i^* + \sum_{j\in J} x_j^* - x^*,\ x \right\rangle\right| \le \eta/2. \tag{5.7}$$

For each $k \in I' \cup J$, let $\epsilon_k := f_k^*(x_k^*) + f_k(x) - \langle x_k^*, x\rangle$; then $\epsilon_k \ge 0$ (Young's inequality), and $x_k^* \in \partial_{\epsilon_k} f_k(x)$ (see (2.5)). Thus (5.3) holds by (5.5). It remains to show (5.2). To see this, note that $(x_k^*, r_k) \in \mathrm{epi} f_k^*$, and so $\epsilon_k \le r_k + f_k(x) - \langle x_k^*, x\rangle$ for each $k \in I' \cup J$. Further, $\sum_{i\in I'} f_i(x) + \sum_{j\in J} f_j(x) \le f(x)$ (since f_i are nonnegative for all $i \in I$). It follows from (5.6) that

$$\sum_{k\in I'\cup J} \epsilon_k \le \sum_{k\in I'\cup J} r_k + f(x) - \left\langle \sum_{i\in I'\cup J} x_k^*, x \right\rangle \le \left\langle x^* - \sum_{k\in I'\cup J} x_k^*, x \right\rangle + (\epsilon + \eta/2),$$

and so (5.2) holds by (5.7). This completes the proof. ∎

Let $|I| \leq +\infty$. Consider the following semi-infinite programming

$$\begin{aligned} &\min_{x\in X} \ f_0(x) \\ &\text{s.t.} \ \ f_i(x) \leq 0 \quad (i \in I), \end{aligned} \tag{5.8}$$

where $\{f_0, f_i : i \in I\} \subseteq \Gamma(X)$. We say that x is a feasible point of (5.8) if $f_0(x) < +\infty$ and $f_i(x) \leq 0$ for all $i \in I$. For any $\epsilon \geq 0$, a feasible point $\overline{x}$ of (5.8) is called an ϵ-solution if $f_0(\overline{x}) \leq f_0(x) + \epsilon$ for all feasible points x of (5.8). As an application of the preceding theorem, we have the following fuzzy KKT result for (5.8). In the special case when $\epsilon = 0$, X is reflexive and $\{f_0, f_i : i \in I\} \subseteq \Gamma_c(X)$. The fuzzy KKT condition was first derived by Jeyakumar, Lee, and Dinh in [19]. In the special case when $\epsilon = 0$ and each f_i $(i \in I)$ is epi-closed, this result was also established by Boţ, Csetnek, and Wanka in [3, 6] via the perturbation approach (indeed, [3, 6] gave the corresponding result for a more general problem: The cone constraint problem).

THEOREM 5.2. *Let $\epsilon \geq 0$, and let $\overline{x}$ be an ϵ-solution of* (5.8). *Let U be a weak*-neighborhood of* 0 *and $\eta > 0$. Then there exist a finite subset I' of I and $\{\epsilon_i : i \in \{0\} \cup I'\} \cup \{\lambda_i : i \in I'\} \subseteq [0, +\infty)$ such that*

$$0 \leq \sum_{i\in I'\cup\{0\}} \epsilon_i \leq \epsilon + \eta, \quad -(\epsilon + \eta) \leq \sum_{i\in I'} \lambda_i f_i(\overline{x}) \leq 0, \tag{5.9}$$

and

$$0 \in x_0^* + \sum_{i\in I'} x_i^* + U \tag{5.10}$$

for some

$$x_0^* \in \partial_{\epsilon_0} f_0(\overline{x}) \text{ and } x_i^* \in \partial_{\epsilon_i}(\lambda_i f_i)(\overline{x}) \ (i \in I'). \tag{5.11}$$

Proof. Define $f : X \to \mathbb{R} \cup \{+\infty\}$ by

$$f(x) = f_0(x) + \sum_{i\in I} g_i(x),$$

where $g_i = \delta_{A_i}$ and $A_i := \{x \in X : f_i(x) \leq 0\}$. Then $f(\overline{x}) = f_0(\overline{x}) < +\infty$ and $f \in \Gamma(X)$. Moreover, since $\overline{x}$ is an ϵ-solution of (5.8), one has $f(\overline{x}) \leq f(x) + \epsilon$ for all $x \in X$ so that $0 \in \partial_\epsilon f(\overline{x})$. We assume without loss of generality that the given weak* neighborhood U is convex. Note that each g_i is a nonnegative function $(i \in I)$. From Theorem 5.1 (applied to $\{I, \{0\}, \{g_i\}_{i\in I}, f_0, \epsilon\}$ in place of $\{I, J, \{f_i\}_{i\in I}, \{f_j\}_{j\in J}, \epsilon\}$), there exist a finite subset I' of I and $\{\overline{\epsilon}_i : i \in \{0\} \cup I'\} \subseteq [0, \infty)$ such that

$$\overline{\epsilon}_0 + \sum_{i\in I'} \overline{\epsilon}_i \leq \epsilon + \frac{\eta}{2} \text{ and } 0 \in x_0^* + \sum_{i\in I'} z_i^* + \frac{U}{2} \tag{5.12}$$

for some $x_0^* \in \partial_{\overline{\epsilon}_0} f_0(\overline{x})$ and $z_i^* \in \partial_{\overline{\epsilon}_i} g_i(\overline{x})$ $(i \in I')$. Since $g_i = \delta_{A_i}$, it follows from (2.4) that $g_i^*(z_i^*) = \sup_{a\in A_i} \langle z_i^*, a\rangle \leq \langle z_i^*, \overline{x}\rangle + \overline{\epsilon}_i$. By (2.12) (applied to f_i in place of f), it follows that

$$(z_i^*, \langle z_i^*, \overline{x}\rangle + \overline{\epsilon}_i) \in \overline{\bigcup_{\lambda>0} \operatorname{epi}(\lambda f_i)^*}^{w^*}.$$

Hence, for each $i \in I'$, there exist $\lambda_i > 0$ and $(x_i^*, s_i) \in \mathrm{epi}(\lambda_i f_i)^*$ such that

$$z_i^* \in x_i^* + \frac{U}{2|I'|}, \tag{5.13}$$

$$\langle z_i^* - x_i^*, \overline{x}\rangle < \eta/(4|I'|), \tag{5.14}$$

and

$$|\langle z_i^*, \overline{x}\rangle + \overline{\epsilon}_i - s_i| < \eta/(4|I'|). \tag{5.15}$$

By (2.11) $\big($applied to $\{\lambda_i f_i, \overline{x}\}$ in place of $\{f, x\}\big)$, for each $i \in I'$ there exists $\epsilon_i \geq 0$ such that

$$x_i^* \in \partial_{\epsilon_i}(\lambda_i f_i)(\overline{x}) \text{ and } s_i = \epsilon_i + \langle x_i^*, \overline{x}\rangle - \lambda_i f_i(\overline{x}). \tag{5.16}$$

Thus, letting $\epsilon_0 := \overline{\epsilon}_0$, (5.11) holds. By (5.12) and (5.13), (5.10) also holds. Since each $\epsilon_i \geq 0$ and $f_i(\overline{x}) \leq 0$ $(i \in I')$, we note that

$$\max\left\{\sum_{i\in\{0\}\cup I'} \epsilon_i \ , \ -\sum_{i\in I'} \lambda_i f_i(\overline{x})\right\} \leq \sum_{i\in\{0\}\cup I'} \epsilon_i - \sum_{i\in I'} \lambda_i f_i(\overline{x});$$

thus to prove (5.9), it suffices to show that

$$\sum_{i\in\{0\}\cup I'} \epsilon_i - \sum_{i\in I'} \lambda_i f_i(\overline{x}) \leq \epsilon + \eta. \tag{5.17}$$

To do this, note that, for each $i \in I'$,

$$\epsilon_i - \lambda_i f_i(\overline{x}) = s_i - \langle x_i^*, \overline{x}\rangle = (s_i - \langle z_i^*, \overline{x}\rangle) + \langle z_i^* - x_i^*, \overline{x}\rangle < \left(\overline{\epsilon}_i + \frac{\eta}{4|I'|}\right) + \frac{\eta}{4|I'|}$$

$\big($see (5.14), (5.15), and (5.16)$\big)$. Hence it follows from (5.12) that

$$\epsilon_0 + \sum_{i\in I'} \big(\epsilon_i - (\lambda_i f_i)(\overline{x})\big) < \overline{\epsilon}_0 + \sum_{i\in I} \overline{\epsilon}_i + \eta/2 \leq \epsilon + \eta.$$

Thus (5.17) is true, and the proof is completed. □

Acknowledgments. The authors would like to express their sincere thanks to the anonymous referees for many helpful comments and for pointing out the references [3, 4, 5, 6, 11, 26].

REFERENCES

[1] A. Auslender and M. Teboulle, *Asymptotic Cones and Functions in Optimization and Variational Inequalities*, Springer Monogr. Math. 12, Springer-Verlag, New York, 2003.

[2] V. Barbu and T. Precupanu, *Convexity and Optimization in Banach Spaces*, revised edition, translated from the Romanian. Editura Academiei, Bucharest; Sijthoff and Noordhoff, Gronigen, The Netherlands, 1978.

[3] R. I. Boţ, E. R. Csetnek, and G. Wanka, *Sequential optimality conditions in convex programming via perturbation approach*, J. Convex Anal., 15 (2008), pp. 149–164.

[4] R. I. Boţ, S. M. Grad, and G. Wanka, *Generalized Moreau-Rockafellar Results for Composed Convex Functions*, preprint 16/2007, Faculty of Mathematics, Chemnitz University of Technology, Chemnitz, Germany, 2007.

[5] R. I. Boţ and G. Wanka, *A weaker regularity condition for subdifferential calculus and Fenchel duality in infinite dimensional spaces*, Nonlinear Anal., 64 (2006), pp. 2787–2804.
[6] R. I. Boţ, E. R. Csetnek, and G. Wanka, *Sequential optimality conditions for composed convex optimization problems*, J. Math. Anal. Appl., 342 (2008), pp. 1015–1025.
[7] H. H. Bauschke and J. M. Borwein, *On projection algorithms for solving convex feasibility problems*, SIAM Rev., 38 (1996), pp. 367–426.
[8] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed., Athena Scientific, Belmont, MA, 1999.
[9] R. S. Burachik and V. Jeyakumar, *A new geometric condition for Fenchel's duality in infinite dimensional spaces*, Math. Program. Ser. B, 104 (2005), pp. 229–233.
[10] R. S. Burachik and V. Jeyakumar, *A dual condition for the convex subdifferential sum formula with applications*, J. Convex Anal., 12 (2005), pp. 279–290.
[11] R. S. Burachik, V. Jeyakumar, and Z. Y. Wu, *Necessary and sufficient conditions for stable conjugate duality*, Nonlinear Anal., 64 (2006), pp. 1998–2006.
[12] J. V. Burke and P. Tseng, *A unified analysis of Hoffman's bound via Fenchel duality*, SIAM J. Optim., 6 (1996), pp. 265–282.
[13] K. Deimling, *Nonlinear Functional Analysis*, Springer-Verlag, Berlin, 1985.
[14] F. R. Deutsch, *Best Approximation in Inner Product Spaces*, CMS Books Math., 7, Springer-Verlag, New York, 2001.
[15] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, Stud. Math. Appl. 1, North-Holland, Amsterdam-Oxford, 1976 (in French); American Elsevier, New York, 1976.
[16] W. Fenchel, *Convex Cones, Sets and Functions*, Mimeographed notes, Princeton University, Princeton, NJ, 1951.
[17] M. A. Goberna and M. A. López, *Linear Semi-infinite Optimization*, Wiley Ser. Math. Methods Pract. 2, John Wiley, Chichester, 1998.
[18] J. B. Hiriart-Urruty and R. R. Phelps, *Subdifferential calculus using ϵ-subdifferentials*, J. Funct. Anal., 118 (1993), pp. 154–166.
[19] V. Jeyakumar, G. M. Lee, and N. Dinh, *New sequential Lagrange multiplier conditions characterizing optimality without constraint qualification for convex programs*, SIAM J. Optim., 14 (2003), pp. 534–547.
[20] C. Li, K. F. Ng, and T. K. Pong, *The SECQ, linear regularity, and the strong CHIP for an infinite system of closed convex sets in normed linear spaces*, SIAM J. Optim., 18 (2007), pp. 643–665.
[21] M. López and G. Still, *Semi-infinite programming*, European J. Oper. Res., 180 (2007), pp. 491–518.
[22] J. J. Moreau, *Fonctions convexes duales et points proximaux dans un espace hilbertien*, C. R. Math. Acad. Sci. Paris, 255 (1962), pp. 2897–2899 (in French).
[23] K. F. Ng and W. Song, *Fenchel duality in infinite-dimensional setting and its applications*, Nonlinear Anal., 25 (2003), pp. 845–858.
[24] R. R. Phelps, *Convex Functions, Monotone Operators and Differentiability*, Lecture Notes in Math. 1364, Springer-Verlag, Berlin, 1993.
[25] S. Reich and S. Simons, *Fenchel duality, Fitzpatrick functions and the Kirszbraun-Valentine extension theorem*, Proc. Amer. Math. Soc., 133 (2005), pp. 2657–2660.
[26] S. Simon, *From Hahn-Banach to Monotonicity*, Springer-Verlag, Berlin, 2008.
[27] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[28] R. T. Rockafellar, *Extension of Fenchel's duality theorem for convex functions*, Duke Math. J., 33 (1966), pp. 81–89.
[29] K. Yosida, *Functional Analysis*, reprint of the 6th (1980) ed., Classics Math., Springer-Verlag, Berlin, 1995.
[30] C. Zalinescu, *Convex Analysis in General Vector Spaces*, World Scientific, River Edge, NJ, 2002.
[31] X. Y. Zheng and K. F. Ng, *Error bound moduli for conic convex systems on Banach spaces*, Math. Oper. Res., 29 (2004), pp. 213–228.
[32] X. Y. Zheng and K. F. Ng, *Metric regularity and constraint qualifications for convex inequalities on Banach spaces*, SIAM J. Optim., 14 (2004), pp. 757–772.

SIAM J. OPTIM.
Vol. 19, No. 4, pp. 1511–1533

MULTIVARIABLE UTILITY FUNCTIONS*

MARIA B. CHIAROLLA† AND ULRICH G. HAUSSMANN‡

Abstract. Utility functions of several variables are ubiquitous in economics. Their maximization requires inversion of the gradient map. Using convex analysis tools, we provide a representation of an extension of this inverse that accounts for possible constraints. To solve economic equilibrium problems, the utility functions of the agents are frequently aggregated into a representative (agent's) utility function. We establish regularity and inversion properties of such representative utility functions.

Key words. multivariable utility functions, representative agent utility function, convex analysis

AMS subject classification. 91B16

DOI. 10.1137/070702266

1. Introduction. Concave functions appear frequently in economics, e.g., as production functions (translating "input" into "output") and as utility functions (producing an ordering of "preferences"); cf. [1], [2], [3], [7], [8], [9], [10], [11]. In early, simple dynamic utility maximization problems, utility functions depended on one variable—consumption of a single good. But more recently models where preferences depend on the consumption of several goods, cf. [2], or on several variables such as leisure and money supply in addition to consumption, cf. [3], [10], require multivariable utility functions. And of course the variables are constrained—typically the variable x_i lies in an interval. The two main properties of utility functions used in these settings, inversion of the derivative and aggregation of the functions, are well understood in the scalar case, cf. [9], but less so in the multivariable case. It is the purpose of this paper to remedy this situation.

Utility functions are assumed to be strictly increasing, strictly concave, and usually smooth, with "derivative equal to zero at infinity," meaning in the scalar case that $\lim_{x\to\infty} u_x(x) = 0$, where u_x is the derivative of the function. If the feasible set in the utility maximization is bounded as in [11], then conditions at infinity can be dispensed with; however, that excludes the interesting cases. Dana and Pontier [7] present a nice perspective on Arrow–Debreu and Arrow–Radner equilibriums in a dynamic, stochastic economy with exogenous endowment without assumptions on the utility functions at zero or infinity, but in the single variable case. In [5] we study equilibrium in a dynamic, stochastic economy with endogenous endowment consisting in part from the return on labor provided to produce the commodity. And in [4] we study the simpler problem of a static, deterministic economy. Both of these papers work with multivariable utility functions and rely on the present work. Smooth utility

*Received by the editors September 7, 2007; accepted for publication (in revised form) July 20, 2008; published electronically December 31, 2008. This work was supported by the Natural Sciences and Engineering Research Council of Canada under grant 88051 and by the Program for Cultural and Scientific Cooperation between Università di Roma "La Sapienza" and the University of British Columbia.

http://www.siam.org/journals/siopt/19-4/70226.html

†Dipartimento di Matematica per le Dec. Econ. Finanz. e Assic., Facoltà di Economia, Università degli Studi di Roma "La Sapienza," via del Castro Laurenziano 9, 00161 Roma, Italy (maria.chiarolla@uniroma1.it).

‡Department of Mathematics, University of British Columbia, 1984 Mathematics Road, Vancouver, BC, V6T 1Z2, Canada (uhaus@math.ubc.ca).

functions of several variables are also considered by Bank and Riedel [2] for a portfolio optimization problem rather than an equilibrium problem, but they do not use a general representation of the extension of the inverse of the gradient of the utility function. Nonsmooth multivariable utility functions are considered in [8] for the problem of maximizing terminal utility derived from a portfolio of financial instruments under transaction costs.

We explain first the inversion result. Let $y^\top$ denote the transpose of y. Consider the problem

$$\max_{x\in A}\{u(x) - y^\top x\} \tag{1.1}$$

for a concave, strictly increasing function u and a convex set A. u may be a production function transforming a quantity x of resources into goods (or cash after the sale of the goods), and y is the price vector of the resources. Given a price y, the manager of the facility wants to determine the resources x required to maximize profit, so he wants to solve (1.1). The solution is $x = (\nabla u)^{-1}(y)$ if $y \in \nabla u(A)$. But what if y is not in this set? Convex analysis treats this problem by extending u to u_A, which is defined to be $-\infty$ off A. Then the solution is $x = (\partial u_A)^{-1}(y) = \partial u_A^*(y)$, where $\partial f(x)$ is the supergradient of a concave function f at x and f^* is the concave conjugate of f. The question then is to find the (effective) domain of ∂u_A^* since (1.1) has a solution when y lies in this domain. In section 3 we show that the mapping ∂u_A^* is single-valued; we call it I^{u_A}. Then we estimate its domain, and we show that $I^{u_A}(y) = (\nabla u)^{-1}(\mathcal{P}_A(y))$, where $\mathcal{P}_A$ is a suitable projection of y onto $\nabla u(A)$, so I^{u_A} is an extension of $(\nabla u)^{-1}$. We study its regularity; the tools are standard results from convex analysis; cf. [12].

In section 3 we allow somewhat general sets A, but in the economic applications A usually has more structure. In the above interpretation each component x_i must lie in the interval of available resources, $x_i \geq 0$ or possibly $x_i \in [0, b_i]$, which can be transformed into $[0, 1]$ by rescaling. In [5] where u is an agent's utility function and $x = (c, l)$, with c a consumption rate and l a leisure rate, $A = [0, \infty) \times [0, 1]$, i.e., we have "box" constraints. This is the setting for aggregation of several utility functions and preference sets treated in section 4.

We sketch now a simple equilibrium problem to indicate the purpose of aggregation of utility functions. We have J agents. Agent j is endowed with ρ^j units of a resource which he can sell at price y. The resource is transformed into consumables; z units of resource produces $v(z)$ units of consumable. The price of a consumable good is one. Agent j obtains utility $u^j(c^j, r^j)$ from consuming c^j units of good and holding r^j units of the resource. His personal set of preferences is described by a set in A^j; now we restrict A^j to be a right parallelepiped; i.e., his preferences are subject to "box" constraints. He wishes to maximize u^j over A^j, subject to a budget constraint $c^j \leq y(\rho^j - r^j)$ or equivalently

$$\max_{(c,r)\in A^j}\{u^j(c, r) - \eta[c + y\, r]\} \quad \text{with solution } (c^j, r^j) = I^{u^j_{A^j}}(\eta(1, y)),$$

where η is a Lagrange multiplier and must satisfy $(1, y)\cdot I^{u^j_{A^j}}(\eta(1, y)) = y\,\rho^j$. Observe that the budget constraint will reduce to an equality at the max. The total amount of resource in the economy is $\sum_{j=1}^J \rho^j := Z$, and the amount of good produced will be $v(z)$, where $z = Z - \sum_j r^j$. *Equilibrium* holds if all of the goods are consumed, i.e., $v(z) = \sum_j c^j$ (market clearing). The question is, Can a price y be found so that equilibrium is obtained?

If there is only one agent in the economy (with utility function u), then the equilibrium problem can be solved as follows: Since market clearing implies $c = v(z),\ r = Z - z$, take y such that $I^{u_A}(\eta(1,y)) = (v(z), Z-z)$, i.e., $\eta(1,y) = \nabla u(v(z), Z-z) = (u_c(v(z), Z-z), u_r(v(z), Z-z))$. z and y are unknown, but the choice of y implies that

$$y = \frac{u_r(v(z), Z-z)}{u_c(v(z), Z-z)}. \tag{1.2}$$

However, if z solves the auxiliary maximization problem

$$\max_{z \le Z}\ u(v(z), Z-z),$$

then the right side of (1.2) is $v'(z)$, and the solution of this maximization problem gives z independent of y; the latter can then be found as $v'(z)$. Then $(c,r) = I^{u_A}(\eta(1,y))$ solves the agent's problem. Market clearing follows from $I^{u_A}(\eta(1,y)) = (v(z), Z-z)$.

Motivated by this result when $J = 1$, we introduce a representative agent with a utility function $u(c, r; \Lambda)$, which is an aggregation of u^j and depends on a parameter $\Lambda = (\lambda_1, \ldots, \lambda_J)$, with $\lambda_j > 0$; cf. (4.1). The properties of this function are less transparent in the multivariable case, but we show that they are sufficient to apply the above process to solve the problem. This (in addition to the characterization of I^{u_A} above) is the main result of the paper and is applied to a static example as above in [4] and to a stochastic dynamic equilibrium model in [5].

In section 2 we summarize our notation. Section 3 contains the result on the extension of the inverse of ∇u. Here we keep A fairly general, but in section 4 we study the properties of the aggregated utility function assuming that each set A^j is a (possibly semi-infinite) right parallelepiped, i.e., a box.

2. Notation. We summarize some notation and results from convex analysis on $\Re^n$ (but in the context of *concave* functions); our reference is [12]. A set is *affine* if it is the translate of a subspace of $\Re^n$, including $\{0\}$ and $\Re^n$. $\operatorname{int}(A)$ denotes the interior of the set A, $\operatorname{cl}(A)$ denotes its closure, $\operatorname{bdy}(A)$ denotes its boundary, $\operatorname{aff}(A)$ denotes its affine hull (smallest affine set containing A), and $\operatorname{ri}(A)$ denotes the relative interior of A, i.e., the interior of A relative to $\operatorname{aff}(A)$. If u is a function $\Re^n \mapsto [-\infty, \infty)$, then the (effective) domain of u is $\operatorname{dom}(u) := \{x : u(x) > -\infty\}$ and $\operatorname{im}(u) := u(\operatorname{dom}(u))$. u is *nondecreasing* if $u(x) \le u(x')$ whenever $x, x' \in \operatorname{dom}(u)$ and $x \le x'$ and where the latter inequality in $\Re^n$ is taken componentwise. u is *strictly increasing* if $u(x) < u(x')$ for such x, x', with $x \ne x'$; i.e., u is strictly increasing in each of its n arguments. The function u is *(strictly) concave* if it is (strictly) concave on $\operatorname{dom}(u)$, which we assume to be nonempty. This makes the function a proper, concave function in the terminology of convex analysis.

The *conjugate* function of the concave function u is defined as

$$u^*(y) := \inf_{x \in \Re^n} \{x^\top y - u(x)\}.$$

Observe that $\bar{u} := -u$ is convex if u is concave and its (convex) conjugate function is

$$\bar{u}^*(y) := \sup_{x \in \Re^n} \{x^\top y - \bar{u}(x)\} = -u^*(-y).$$

This implies that $\operatorname{dom}(\bar{u}^*) := \{y : \bar{u}^*(y) < \infty\} = -\operatorname{dom}(u^*)$.

The *supergradients* of u at x are all $y \in \Re^n$ such that for all z

$$u(z) - u(x) \leq (z - x)^\top y.$$

The set of all supergradients (called the superdifferential) is denoted by $\partial u(x)$, and $\mathrm{dom}(\partial u) := \{x : \partial u(x) \neq \emptyset\}$ is the (effective) domain of ∂u. If we write $\partial_o \bar{u}$ for the subgradient of the convex function $\bar{u}$ (both ∂ and ∂_o are generalized gradients in the sense of Clarke [6]), then $\partial u = -\partial_o(\bar{u})$, so the results for convex functions can be translated to concave functions. We observe that if u is concave, then

$$(x - x')^\top y \leq u(x) - u(x') \leq (x - x')^\top y' \tag{2.1}$$

for all $x \neq x'$ and all $y \in \partial u(x),\ y' \in \partial u(x')$, with strict inequality in the case u is strictly concave. If in addition u is differentiable on $\mathrm{int}(\mathrm{dom}(u))$, then $\partial u = \{\nabla u\}$, the gradient of u, on this set. Note that ∂u is monotone in the terminology of convex analysis and strictly monotone in the case of strict concavity. Note bene (N.b.) for $A \subset \Re^n$ a multivalued function $g : A \mapsto 2^{\Re^n}$ is *monotone* if $(x - x')^\top (y - y') \leq 0$ for all $y \in g(x), y' \in g(x')$, and $x, x' \in A$. It is strictly monotone if the inequality is strict for $x \neq x', y \neq y'$. We have translated monotonicity from the convex setting to the concave by reversing the inequality; hence if $n = 1$ the graph of g is decreasing rather than increasing. Finally when $\partial u = \{\nabla u\}$, we will write $\partial u = \nabla u$.

The *recession cone* of a convex set $A \in \Re^n$ is, cf. Theorem 8.1 of [12]

$$0^+ A := \{z \in \Re^n : A + z \subset A\}.$$

If u is a closed, proper concave function, then $\bar{u}$ is a proper convex function and its *recession function* is, cf. Theorem 8.5 of [12] (n.b. $\mathrm{dom}(\bar{u}) = \mathrm{dom}(u)$),

$$(\bar{u}0^+)(x) := \sup_{z \in \mathrm{dom}(\bar{u})} [\bar{u}(z + x) - \bar{u}(z)] = - \inf_{z \in \mathrm{dom}(u)} [u(z + x) - u(z)].$$

Then Theorem 13.3 of [12] implies that

$$\inf_{y \in \mathrm{dom}(u^*)} y^\top x = - \sup_{y \in \mathrm{dom}(\bar{u}^*)} y^\top x = -(\bar{u}0^+)(x). \tag{2.2}$$

The *normal cone* to A at $x \in A$, denoted by $\mathcal{N}_A(x)$, consists of the outward normals to A at x. It is empty for $x \notin A$ and is $\{0_n\}$ for $x \in \mathrm{int}(A)$. Here 0_n denotes the zero vector in $\Re^n$. Note that $\mathrm{aff}(\mathcal{N}_A(x)) = \mathcal{N}_A(x) - \mathcal{N}_A(x)$ is a subspace. Define

$$\chi_A(x) := \begin{cases} 0 & \text{if } x \in A, \\ \infty & \text{if } x \notin A; \end{cases}$$

it is the indicator function of A. Then $\partial_o \chi_A(x) = \mathcal{N}_A(x)$ for $x \in A$ and is empty for $x \notin A$. Define $u_A(x) := u(x) - \chi_A(x)$, $\Re^n_+ := \{x \in \Re^n : x_i \geq 0,\ i = 1, \ldots, n\}$, and $\Re^n_{++} := \{x \in \Re^n : x_i > 0,\ i = 1, \ldots, n\}$.

We shall have occasion to decompose $\Re^n$ as a direct sum of orthogonal subspaces, $\Re^n = Z \oplus Z^\perp$. Then $x \in \Re^n$ decomposes as $x = z \oplus z^\perp$. We denote by $\partial_z u(x)$ the superdifferential with respect to z of u at x, similarly for $\nabla_z u(x)$. In section 4 the subspaces Z will always be generated by a subset of the standard basis vectors $\vec{e}_1, \ldots, \vec{e}_n$, but in section 3 a rotation may intervene.

3. The inverse. For a positive integer n, consider a function $u : \Re^n \mapsto [-\infty, \infty)$. Henceforth we work under the following assumptions. The smoothness assumptions are very strong relative to the current work in convex analysis, but they are standard in a large part of the mathematical finance and economics literature.

Assumption 1.

(i) u is upper semicontinuous, concave, and nondecreasing, and $\text{int}(\text{dom}(u)) \neq \emptyset$;
(ii) u is continuous on $\text{dom}(u)$;
(iii) u is continuously differentiable on $\text{int}(\text{dom}(u))$.

From (i) it follows that u is a closed, proper concave function. (ii) removes some possible pathologies at any boundary of $\text{dom}(u)$, and (i) and (iii) imply that $\partial u = \nabla u$ on $\text{int}(\text{dom}(u)) \subset \text{dom}(\partial u)$. Typically for us $\text{int}(\text{dom}(u)) = \Re^n$ or $x_o + \Re^n_{++}$ for some $x_o \in \Re^n$.

Assumption 2.

(i) $A \neq \emptyset$, convex, $A \subset \Re^n_+ \cap \text{dom}(u)$, and A is closed relative to $\text{dom}(u)$;
(ii) $\text{dom}(\partial u_A) \subset \text{int}(\text{dom}(u)) \cap A$;
(iii) u is strictly increasing, strictly concave on $\text{dom}(\partial u_A)$.

From Assumption 2(i) it follows that u_A is a closed, proper concave function with $\text{dom}(u_A) = A$. Assumption 2(ii) implies that $\text{ri}(A) = \text{ri}(\text{dom}(u_A)) \subset \text{dom}(\partial u_A) \subset \text{int}(\text{dom}(u))$ and hence Theorem 23.8 of [12] implies $\partial u_A(x) = \partial u(x) - \mathcal{N}_A(x)$ since $\partial_o \chi_A(x) = \mathcal{N}_A(x)$. Furthermore (the last inclusion follows from Assumption 2(ii) again)

$$\text{int}(\text{dom}(u)) \cap A \subset \text{dom}(\partial u) \cap A = \text{dom}(\partial u_A) \subset \text{int}(\text{dom}(u)) \cap A, \tag{3.1}$$

so $\text{dom}(\partial u_A) = \text{int}(\text{dom}(u)) \cap A$ is convex and $\partial u_A(x) = \{\nabla u(x)\} - \mathcal{N}_A(x)$, and hence $\nabla u(x) \in \partial u_A(x)$ on $\text{dom}(\partial u_A)$ with equality on $\text{int}(A)$.

Note that Assumption 2(ii) can often be satisfied by extending u. For example, if $u(x) = \sqrt{x+1}$ on $\text{dom}(u) := [0, \infty) = A$, then (ii) fails, but by extending the effective domain to $\text{dom}(u) := (-1, \infty)$ we assure Assumption 2(ii) without changing the problem (1.1).

Assumption 3. There exists $\mu \in 0^+A$ such that for all $x \in 0^+A$

$$\inf_{z \in \text{dom}(\partial u_A)} \nabla u(z)^\top x \leq \mu^\top x. \tag{3.2}$$

Because ∂u is monotone, Assumption 3 is a growth condition at infinity.

Example 3.1.

(i) $n = 1$ and $u(x) = \ln(x - \varepsilon)$, so $\text{dom}(u) = \text{dom}(\partial u) = (\varepsilon, \infty)$. Take $A = (0, \infty)$ if $\varepsilon = 0$ or $[0, \infty)$; if $\varepsilon < 0$, then Assumptions 1, 2, and 3 are satisfied with $\mu = 0$.
(ii) $n = 2$ and

$$u(x_1, x_2) := \begin{cases} (x_1 - \varepsilon_1)^{\gamma_1}(x_2 - \varepsilon_2)^{\gamma_2} & \text{if } x_i \geq \varepsilon_i,\ i = 1, 2, \\ -\infty & \text{if } x_i < \varepsilon_i, \end{cases}$$

with $\gamma_i > 0$, $\gamma_1 + \gamma_2 < 1$; then $\text{dom}(\partial u) = (\varepsilon_1, \varepsilon_2) + \Re^2_{++} = \text{int}(\text{dom}(u))$, and u satisfies Assumption 1. If

(a) $A = [0, \infty) \times [0, 1]$, then Assumption 2 holds for $\varepsilon_i \leq 0$. Note that $\text{dom}(\partial u_A) = A$ if both $\varepsilon_i < 0$, but $\text{dom}(\partial u_A) = A \cap \Re^2_{++}$ if both $\varepsilon_i = 0$;
(b) $A = [0, \infty) \times \{0\}$ and we require $\varepsilon_2 < 0$, then $\text{dom}(\partial u_A) = A$ if $\varepsilon_1 < 0$, but $\text{dom}(\partial u_A) = A \cap \{x_1 > 0\}$ if $\varepsilon_1 = 0$. Assumption 2 holds. Observe that $\mathcal{N}_A(x) = \{0\} \times \Re$ for $x \in \text{ri}(A)$ and $\mathcal{N}_A(x) = (-\infty, 0] \times \Re$ for $x = (0, 0)$ and is empty otherwise.

In both cases $0^+A = [0,\infty)\times\{0\}$ and Assumption 3 holds with $\mu = 0_2$ since $\gamma_1 < 1, \gamma_2 > 0$.

(iii) For $A = [0,\infty)\times\{0\}$, it is also interesting to consider the case $\gamma_2 = 0$, i.e.,

$$u(x_1,x_2) := \begin{cases} (x_1-\varepsilon_1)^{\gamma_1} & \text{if } x_1 \geq \varepsilon_1, \\ -\infty & \text{if } x_1 < \varepsilon_1, \end{cases}$$

with $0 < \gamma_1 < 1$. Assumption 1 holds with $\mathrm{dom}(\partial u) = (\varepsilon_1,\infty)\times\Re = \mathrm{int}(\mathrm{dom}(u))$. Since $\mathrm{dom}(\partial u_A) = A$ for $\varepsilon_1 < 0$, $\mathrm{dom}(\partial u_A) = (0,\infty)\times\{0\}$ for $\varepsilon_1 = 0$; then Assumption 2 holds. Assumption 3 also holds with $\mu = 0_2$.

(iv) For $\mu_o \geq 0$, $u(x_1,x_2) := \mu_o x_1 - x_1^{-1} + \sqrt{x_2}$ and $A := (0,\infty)\times[0,1)$ satisfy Assumptions 1, 2, and 3 with $\mu = (\mu_o, 0)$.

Later we shall be considering several utility functions which we will want to define on the same space $\Re^n$, but not all of the functions will depend on all n arguments. Example 3.1(iii) above gives such a function.

Recall (3.1) and define

$$\mathcal{R}_{u_A} := \nabla u(A \cap \mathrm{int}(\mathrm{dom}(u))) = \nabla u(\mathrm{dom}(\partial u_A)). \tag{3.3}$$

Notice that as u is strictly increasing on $\mathrm{dom}(\partial u_A)$, then $\mathcal{R}_{u_A} \subset \Re^n_{++}$. Recall that $\nabla u \in \partial u_A$ on $\mathrm{dom}(\partial u_A)$ with equality on $\mathrm{int}(A)$. Although ∇u_A is only defined on $\mathrm{int}(A)$, we extend it as $\mathrm{gr}_A u(x) := \nabla u(x)$ for $x \in \mathrm{dom}(\partial u_A)$. We explain in Remark 3.11 below why this works for us even when $\mathrm{int}(A) = \emptyset$. Since ∇u is strictly monotone on $A \cap \mathrm{int}(\mathrm{dom}(u))$, we should be able to invert $\mathrm{gr}_A u$ on $\mathcal{R}_{u_A}$, but we want to extend this inverse function as far as possible so that it solves (1.1) for as large a selection of y as possible. Note that $(\nabla u)^{-1}$ will usually not be the correct extension; for example, if $n = 1$ additional conditions (the Inada conditions) must be imposed for this to be true.

As u_A is a closed, proper concave function, the conjugate concave function of u_A is again closed, proper, and concave, and is

$$u_A^*(y) = \inf_{x\in\Re^n}\{x^\top y - u_A(x)\} = \inf_{x\in A}\{x^\top y - u(x)\}.$$

PROPOSITION 3.2. *Under Assumptions* 1 *and* 2*, there exists a continuous, monotone function* $I^{u_A} : \mathrm{int}(\mathrm{dom}(u_A^*)) \to \mathrm{dom}(\partial u_A) = \mathrm{dom}(\partial u) \cap A$ *that extends* $(\mathrm{gr}_A u)^{-1}$ *beyond* $\mathcal{R}_{u_A}$ *and solves* (1.1)*.* I^{u_A} *is strictly monotone on* $\mathcal{R}_{u_A}$*. If* ∇u *is* p *times continuously differentiable, then so is* I^{u_A} *on* $\mathrm{int}(\mathcal{R}_{u_A})$*.*

Proof. The solution of (1.1) is

$$I^{u_A}(y) := \arg\min_{x\in\Re^n}\{x^\top y - u_A(x)\} = \{x : y \in \partial u_A(x)\} = \partial u_A^*(y) \tag{3.4}$$

according to Theorem 23.5 and Corollary 23.5.1 of [12]. The strict concavity of u_A on $\mathrm{dom}(\partial u_A)$ and (2.1) implies that I^{u_A} is single-valued, so $\partial u_A^*(y)$ is a singleton and hence a bounded set, and Theorem 23.4 of [12] implies that $\mathrm{dom}(I^{u_A}) = \mathrm{dom}(\partial u_A^*) = \mathrm{int}(\mathrm{dom}(u_A^*))$. Moreover, $y \in \mathcal{R}_{u_A}$ implies that $y = \nabla u(x)$ for some $x \in \mathrm{dom}(\partial u_A)$, i.e., $y \in \partial u_A(x)$ and hence $x \in \partial u_A^*(y) = I^{u_A}(y)$, cf. Corollary 23.5.1 of [12] and we conclude that $I^{u_A} = (\mathrm{gr}_A u)^{-1}$ on $\mathcal{R}_{u_A}$; i.e., I^{u_A} extends the inverse of $\mathrm{gr}_A u$ beyond $\mathcal{R}_{u_A}$, and $\mathrm{im}(I^{u_A}) = \mathrm{dom}(\partial u_A^{**}) = \mathrm{dom}(\partial u_A)$.

Since I^{u_A} is single-valued, then $I^{u_A} = \nabla u_A^*$, and so u_A^* is differentiable on $\mathrm{dom}(I^{u_A})$. Since I^{u_A} is proper and concave, then it is continuous on this set; cf.

Theorem 25.5 of [12]. The concavity of u_A^* implies that $I^{u_A} = \nabla u_A^*$ is monotone. Strict concavity of u on $\mathrm{dom}(\partial u_A)$ implies

$$(x^1 - x^2)^\top (\nabla u(x^1) - \nabla u(x^2)) < 0 \quad \text{for } x^1 \neq x^2 \in \mathrm{dom}(\partial u_A), \quad \text{so}$$
$$(I^{u_A}(y^1) - I^{u_A}(y^2))^\top (y^1 - y^2) < 0 \quad \text{for } y^1 \neq y^2 \in \mathcal{R}_{u_A}. \tag{3.5}$$

It follows that I^{u_A} is strictly monotone on $\mathcal{R}_{u_A}$.

Let us now look at the differentiability of I^{u_A}. If ∇u is differentiable, then the Hessian of u at x, i.e., $H_u(x)$, exists and is negative definite, and hence so is its inverse $(H_u(x))^{-1}$. The inverse function theorem now implies that I^{u_A} is p times continuously differentiable on $\mathrm{int}(\mathcal{R}_{u_A})$ if ∇u is. □

We now find inner and outer estimates for $\mathrm{dom}(I^{u_A})$. Define the polar of a nonempty convex cone K as $K^\circ := \{y : y^\top x \leq 0 \quad \forall x \in K\}$, and observe that $(-K)^\circ = -K^\circ$.

PROPOSITION 3.3. *Under Assumptions* 1, 2, *and* 3,

$$\mu - \mathrm{int}((0^+A)^\circ) \subset \mathrm{dom}(I^{u_A}) \subset -\mathrm{int}((0^+A)^\circ).$$

Proof. For the inner estimate it suffices to show that $\inf_{y \in \mathrm{dom}(u_A^*)} y^\top x \leq \inf_{y \in \mu - (0^+A)^\circ} y^\top x$ since this implies $\mathrm{cl}(\mu - (0^+A)^\circ) \subset \mathrm{cl}(\mathrm{dom}(u_A^*))$, and hence $\mu - \mathrm{int}((0^+A)^\circ) = \mathrm{int}(\mathrm{cl}(\mu - (0^+A)^\circ)) \subset \mathrm{int}(\mathrm{cl}(\mathrm{dom}(u_A^*))) = \mathrm{ri}(\mathrm{dom}(u_A^*)) \subset \mathrm{dom}(\partial u_A^*) = \mathrm{dom}(I^{u_A})$. As $A \subset \Re^n_+$, then 0^+A cannot contain any nontrivial subspace, so Corollary 14.6.1 of [12] implies that $\mathrm{int}((0^+A)^\circ) \neq \emptyset$, and the same must be true for $\mathrm{dom}(u_A^*)$; hence $\mathrm{int}(\mathrm{dom}(u_A^*)) = \mathrm{ri}(\mathrm{dom}(u_A^*))$.

From (2.2) and concavity

$$\inf_{y \in \mathrm{dom}(u_A^*)} y^\top x = \inf_{z \in \mathrm{dom}(u_A)} [u_A(z + x) - u_A(z)] \leq \inf_{\substack{z \in \mathrm{dom}(u_A) \\ y \in \partial u_A(z)}} y^\top x. \tag{3.6}$$

If $x \notin 0^+A = 0^+(\mathrm{ri}(A))$, then the second infimum is $-\infty$ since for some $z \in \mathrm{ri}(A) \subset \mathrm{dom}(u_A)$ we must have $z + x \notin \mathrm{ri}(A)$ and a slight perturbation of z produces $z' \in \mathrm{ri}(A)$ (so $u_A(z') > -\infty$) with $z' + x \notin \mathrm{cl}(A) \supset \mathrm{dom}(u_A)$; hence $u_A(z' + x) = -\infty$.

For $x \in 0^+A$, (3.6) and (3.2) imply

$$\inf_{y \in \mathrm{dom}(u_A^*)} y^\top x \leq \inf_{\substack{z \in \mathrm{dom}(\partial u_A) \\ y \in \partial u_A(z)}} y^\top x \leq \inf_{z \in \mathrm{dom}(\partial u_A)} \nabla u(z)^\top x \leq \mu^\top x.$$

On the other hand, if $x \in 0^+A$, then

$$\inf_{y \in \mu - (0^+A)^\circ} y^\top x = \mu^\top x + \inf_{y \in -(0^+A)^\circ} y^\top x.$$

But $y \in -(0^+A)^\circ \Leftrightarrow y^\top z \leq 0 \quad \forall z \in -(0^+A) \Rightarrow y^\top x \geq 0$. Taking $\lim_{\|y\| \downarrow 0}$ yields $\inf_{y \in -(0^+A)^\circ} y^\top x = 0$, so for all x

$$\inf_{y \in \mathrm{dom}(u_A^*)} y^\top x \leq \inf_{y \in \mu - (0^+A)^\circ} y^\top x$$

and the inner approximation follows.

For the outer approximation, observe that for $z \in 0^+A$ and $y \in \Re^n$

$$\begin{aligned} \sup_{x \in A} \{u(x) - y^\top x\} &\geq \sup_{x \in A + z} \{u(x) - y^\top x\} = \sup_{x \in A} \{u(z + x) - y^\top (z + x)\} \\ &= \sup_{x \in A} \{u(z + x) - y^\top x\} - y^\top z \geq \sup_{x \in A} \{u(x) - y^\top x\} - y^\top z \end{aligned}$$

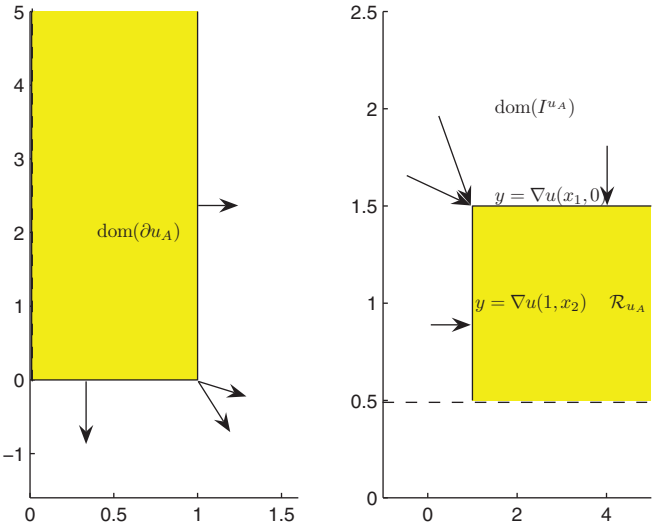


FIG. 1. *A*, $\mathrm{dom}(\partial u_A)$, $\mathrm{dom}(I^{u_A})$, *and* $\mathcal{R}_{u_A}$.

since u is strictly increasing and $0^+A \subset \Re^n_+$. Hence $u^*_A(y) \le u^*_A(y) + y^\top z$, i.e., $y^\top z \ge 0$ for all $z \in 0^+A$ and $y \in \mathrm{dom}(u^*_A)$. But $y \in -(0^+A)^\circ \Leftrightarrow y^\top z \ge 0 \ \ \forall z \in 0^+A$, so $\mathrm{dom}(I^{u_A}) \subset \mathrm{dom}(u^*_A) \subset -(0^+A)^\circ$. The result follows since $\mathrm{dom}(I^{u_A})$ is open; cf. Proposition 3.2. □

Remark 3.4. We observe that if $\mu = 0$, then our estimates are exact, i.e., $\mathrm{dom}(I^{u_A}) = -(0^+A)^\circ$.

Example 3.5. Let $u(x_1, x_2) := 2\sqrt{x_1} + \mu_o x_2 + \frac{x_2}{1+x_2}$ and $A := [0,1] \times [0,\infty)$, so $\mathrm{dom}(\partial u) = (0,\infty) \times (-1,\infty)$, $\mathcal{R}_{u_A} = [1,\infty) \times (\mu_o, \mu_o + 1]$, and $\mathrm{dom}(I^{u_A}) = \{y \in \Re^2 : y_2 > \mu_o\}$. In Figure 1 we have used $\mu_o = 0.5$. In the left panel the shaded region is A, and with the left boundary deleted it is $\mathrm{dom}(\partial u_A)$. The arrows denote outward normals. In the right panel $\mathcal{R}_{u_A}$ is the shaded region, $\mathrm{dom}(I^{u_A})$ is the region above the dashed line, and the arrows represent the projection $\mathcal{P}_{u_A}$; cf. Corollary 3.9.

Remark 3.6. Write $H_u(x)$ for the Hessian of u at x. If u is twice continuously differentiable on $\mathrm{int}(\mathrm{dom}(u))$, then ∇I^{u_A} is a continuous function on $\mathcal{R}_{u_A}$ since $\nabla I^{u_A}(y) = (H_u(I^{u_A}(y)))^{-1}$.

Remark 3.7. Sometimes u will also depend on a parameter, i.e., $u(x,t)$; nevertheless, $\nabla u(x,t)$ will still denote the gradient of u *with respect to* x only. Assumptions 1, 2, and 3 are assumed to hold for each t. For such u with t in an interval $\mathcal{T}$, I^{u_A} is a function of (y,t) and $\mathcal{R}_{u_A}$ also depends on t. If $\nabla u(x,t)$ is continuously differentiable in (x,t), then I^{u_A} is continuously differentiable in $\{(y,t) : y \in \mathrm{int}(\mathcal{R}_{u_A}(t)), t \in \mathrm{int}(\mathcal{T})\}$. This follows again from the implicit function theorem.

Let us give a representation of $I^{u_A}(y)$ as a projection of $\mathrm{dom}(I^{u_A})$ onto $\mathcal{R}_{u_A}$ parallel to a vector in $\mathcal{N}_A(x)$, the set of outward normals to A at x, for some $x \in A$ such that $y = \nabla u(x)$. If A is a polyhedral set, then $\mathcal{N}_A(x)$ is constant on each face, so the projection is particularly simple to identify; cf. Examples 3.5 and 4.8 and Figures 1 and 2.

LEMMA 3.8. *For every* $y \in \mathrm{dom}(I^{u_A})$ *there exists a unique* $x \in \mathrm{dom}(\partial u_A)$ *such that* $y + \vec{n} = \nabla u(x)$ *for some* $\vec{n} \in \mathcal{N}_A(x)$.

Proof. Define the convex function $x \mapsto h^y(x) := x^\top y - u(x)$ so $\partial_o h^y(x) = y - \partial u(x)$. Recall ∂_o denotes the subgradient. Then $\mathrm{dom}(h^y) = \mathrm{dom}(u)$ and Assumptions

1(iii) and 2(ii) and Theorem 23.4 of [12] imply

$$\mathrm{ri}(\mathrm{dom}(h^y)) = \mathrm{int}(\mathrm{dom}(u)) \supset \mathrm{dom}(\partial u_A) \supset \mathrm{ri}(\mathrm{dom}(u_A)) = \mathrm{ri}(A) \neq \emptyset.$$

Hence Theorem 27.4 of [12] yields that $x = I^{u_A}(y)$, cf. (3.4), if and only if there exists $x^* \in \partial_o h^y(x)$ such that $-x^* \in \mathcal{N}_A(x)$. Since $x^* \in \partial_o h^y(x)$, then $y - x^* \in y - \partial_o h^y(x) = \partial u(x)$. But $x = I^{u_A}(y) \in \mathrm{dom}(\partial u_A) \subset \mathrm{int}(\mathrm{dom}(u))$, so $y - x^* = \nabla u(x)$. Hence $y \in \nabla u(x) - \mathcal{N}_A(x)$ for $x = I^{u_A}(y)$.

$I^{u_A}(y)$ is the only such x since $\nabla u(x) - \vec{n} = \nabla u(x') - \vec{n}' = y$ implies

$$[\nabla u(x) - \nabla u(x')]^\top [x - x'] = [\vec{n} - \vec{n}']^\top [x - x'] \geq 0$$

because $x \mapsto \mathcal{N}_A(x) = \partial_o \chi(x)$ is maximal monotone, being the convex subdifferential of a convex function. This contradicts the strict concavity of u on $\mathrm{dom}(\partial u_A)$. □

COROLLARY 3.9. *There exists a projection* $\mathcal{P}_{u_A} : \mathrm{dom}(I^{u_A}) \mapsto \mathcal{R}_{u_A}$ *such that*

$$I^{u_A}(y) = (\nabla u)^{-1}(\mathcal{P}_{u_A}(y)).$$

Proof. The required projection is $\mathcal{P}_{u_A}(y) := \nabla u(x)$ with x given by Lemma 3.8. □

Remark 3.10. Corollary 3.9 implies that $\mathrm{dom}(I^{u_A}) = \bigcup_{x \in A \cap\, Int(\mathrm{dom}(u))} [\nabla u(x) - \mathcal{N}_A(x)]$. If $x \in \mathrm{bdy}(A) \cap \mathrm{dom}(\partial u)$, then $\nabla u(x) \in \mathcal{R}_{u_A} \cap \mathrm{bdy}(\mathcal{R}_{u_A})$ and I^{u_A} extends $(\mathrm{gr}_A u)^{-1}$ across these boundary points of $\mathcal{R}_{u_A}$; cf. Example 3.5 and Figure 1. The other boundary points of $\mathcal{R}_{u_A}$ are not in $\mathcal{R}_{u_A}$ and in fact are on $\mathrm{bdy}(\mathrm{dom}(I^{u_A}))$.

Remark 3.11. Let us explain the situation when $\mathrm{int}(A) = \emptyset$. Consider $x \in \mathrm{ri}(A)$. Then $\Re^n$ decomposes into $\mathcal{N}_A(x)^\perp \oplus \mathcal{N}_A(x)$, with $\mathrm{aff}(A)$ a translation of $\mathcal{N}_A(x)^\perp$. For convenience take $n = 2$, $\mu = 0_2$, $A = [0, \infty) \times \{b\}$, with $b \geq 0$, so $\mathcal{N}_A(x)^\top = \{x : x_2 = 0\}$. For $\max_{x \in A}[u(x) - y^\top x]$ the map of interest is the partial derivative u_{x_1} or, to embed it in $\Re^2$, $(u_{x_1}, 0)$. For any $y \in \Re^2$ with $0 < y_1 < u_{x_1}(0, b)$, the appropriate inverse image projects y into $(y_1, 0)$ and then takes the inverse under $(u_{x_1}, 0)$ to produce $x = ((u_{x_1}(\cdot, b))^{-1}(y_1), b) \in \mathrm{ri}(A)$. On the other hand, $(\nabla u)^{-1}(\mathcal{P}_{u_A}(y))$ projects y into $(y_1, y_2^o) \in \nabla u(A)$ for some y_2^o. Then $(\nabla u)^{-1}(y_1, y_2^o)$ is the same x as above. A similar situation occurs for $y_1 \geq u_{x_1}(0, b)$; either approach maps y into $(0, b)$. It is much more convenient to work with ∇u rather than $(u_{x_1}, 0)$; that is why we defined $\mathrm{gr}_A u$ as the restriction of ∇u to $A \cap \mathrm{int}(\mathrm{dom}(u))$.

DEFINITION 3.12. *If* $\mathrm{cl}(A)$ *is polyhedral, then it has a finite number of nonempty faces of dimension less than* n*, call them* $\bar{\mathcal{C}}_k, k = 1, \ldots, K$*; cf. Theorem* 19.1 *of* [12]*. Define* $\mathcal{C}_k := \mathrm{ri}(\bar{\mathcal{C}}_k), k = 1, \ldots, K$*, and* $\mathcal{C}_0 := \mathrm{int}(A)$*, possibly empty. Then* $\mathcal{C}_0, \ldots, \mathcal{C}_K$ *is a partition of* $\mathrm{cl}(A)$*, cf. Theorem* 18.2 *of* [12]*, and* $\mathcal{C}_1, \ldots, \mathcal{C}_K$ *is a partition of* $\mathrm{bdy}(A)$*. Set* $\mathcal{S}_k := (I^{u_A})^{-1}(\mathcal{C}_k), k = 0, \ldots, K$*. Then* $\mathcal{S}_0 = (I^{u_A})^{-1}(\mathrm{int}(A)) = \nabla u(\mathrm{int}(A)) = \mathrm{int}(\mathcal{R}_{u_A})$*. Write* $\mathrm{bdy}(\mathcal{R}_{u_A}, \mathcal{S}_k)$ *for the boundary between* $\mathcal{R}_{u_A}$ *and* $\mathcal{S}_k$*.*

We now discuss differentiability of I^{u_A} off $\mathcal{R}_{u_A}$.

COROLLARY 3.13. *Assume that* A *is a polyhedron and* ∇u *is continuously differentiable. Then* I^{u_A} *and* $\mathcal{P}_{u_A}$ *are continuously differentiable on* $\mathrm{int}(\mathcal{S}_k)$ *for each* k*. The directional derivative of* I^{u_A} *parallel to the boundary is continuous, but normal to the boundary it is not.*

Proof. $\mathcal{N}_A(x)$ is independent of x for $x \in \mathcal{C}_k$, i.e., $\mathcal{N}_A(x) = \mathcal{N}_A(\mathcal{C}_k)$. Now for $y \in \mathcal{S}_k$ we have an orthogonal decomposition $y = y^\perp \oplus y^o$, with $y^o \in \mathrm{aff}(\mathcal{N}_A(\mathcal{C}_k))$, and similarly for $x \in \mathcal{C}_k$ we have $x = x^\perp \oplus x^o$, with $x^o \in \mathrm{aff}(\mathcal{N}_A(\mathcal{C}_k))$, *same* for all $x \in \mathcal{C}_k$, and $x^\perp \in \mathrm{aff}(\mathcal{N}_A(\mathcal{C}_k))^\perp$, the subspace parallel to $\mathrm{aff}(\mathcal{C}_k)$. This is equivalent

to rotating and reordering the axes so that $\mathrm{aff}(\mathcal{C}_k)$ is parallel to the span of the standard basis vectors $\{e_1, \dots, e_\ell\}$; of course the transformation changes with k. Then $\mathcal{P}_{u_A}(y) = \nabla u(x^\perp \oplus x^o) = \nabla_{x^\perp} u(x^\perp \oplus x^o) \oplus \nabla_{x^o} u(x^\perp \oplus x^o)$ where, $\nabla_{x^\perp}$ denotes the gradient with respect to $x^\perp$ and similarly for ∇_{x^o}. Moreover $\mathcal{P}_{u_A}(y) = y^\perp \oplus \phi(y^\perp)$, where $\phi(y^\perp) = \nabla_{x^o} u(x^\perp(y^\perp) \oplus x^o)$ and $x^\perp(y^\perp)$ is the solution of $y^\perp = \nabla_{x^\perp} u(x^\perp \oplus x^o)$. The Hessian with respect to $x^\perp$ of $u(x^\perp \oplus x^o)$ is negative definite since H_u is. The implicit function theorem now implies the continuous differentiability of $x^\perp$. The continuous differentiability on $\mathrm{int}(\mathcal{S}_k)$ follows since $I^{u_A}(y) = x^\perp(y^\perp) \oplus x^o$.

Since $\nabla I^{u_A}(y) = (H_u(I^{u_A}(y)))^{-1}$, then the directional derivative of $I^{u_A}(\cdot)$ in the direction v at $y \in \mathcal{R}_{u_A}$ (where $I^{u_A} = (\nabla u)^{-1}$) is

$$(I^{u_A})'(y; v) = \Big(H_u(I^{u_A}(y))\Big)^{-1} v.$$

For $y = y^\perp \oplus y^o \in \mathcal{S}_k$ the Jacobian of $\mathcal{P}_{u_A}$ intervenes ($\mathcal{P}_{u_A}$ is just $\mathbf{I}$ on $\mathcal{R}_{u_A}$), so we obtain

$$\begin{aligned}
(I^{u_A})'(y; v) &= \Big(H_u(I^{u_A}(y))\Big)^{-1} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \nabla\phi(y^\perp) & \mathbf{0} \end{pmatrix} v \\
&= \Big(H_u(I^{u_A}(y))\Big)^{-1} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ (H_u(I^{u_A}(y)))_{x^o x^\perp}((H_u(I^{u_A}(y)))_{x^\perp x^\perp})^{-1} & \mathbf{0} \end{pmatrix} v \qquad (3.7) \\
&= \begin{pmatrix} ((H_u(I^{u_A}(y)))_{x^\perp x^\perp})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} v,
\end{aligned}$$

where $(H_u)_{x^\perp x^\perp} := (\frac{\partial}{\partial x^\perp})^2 u$ has dimension $\dim(y^\perp) \times \dim(y^\perp) = \ell \times \ell$, $(H_u)_{x^o x^\perp} := \frac{\partial^2}{\partial x^o \partial x^\perp} u$ has dimension $\dim(y^o) \times \dim(y^\perp) = (n - \ell) \times \ell$, and $\mathbf{I}$, $\mathbf{0}$ are the identity and zero matrices, respectively, of appropriate dimensions. We use the obvious matrix notation for linear transformations mapping $\mathrm{aff}(\mathcal{N}(\mathcal{C}_k))^\perp \oplus \mathrm{aff}(\mathcal{N}_A(\mathcal{C}_k)) \mapsto \mathrm{aff}(\mathcal{N}_A(\mathcal{C}_k))^\perp \oplus \mathrm{aff}(\mathcal{N}(\mathcal{C}_k))$.

Note that $\mathrm{bdy}(\mathcal{R}_{u_A}, \mathcal{S}_k) \subset \mathcal{S}_k \cap \mathcal{R}_{u_A}$, and for $y \in \mathrm{bdy}(\mathcal{R}_{u_A}, \mathcal{S}_k)$, $\mathcal{P}_{u_A}(y) = y$. Then the two versions of $(I^{u_A})'$ agree only if v is orthogonal to $\mathrm{aff}(\mathcal{N}_A(\mathcal{C}_k))$. We conclude that the directional derivatives of I^{u_A} in the directions parallel to $\mathcal{N}_A(\mathcal{C}_k)$ are discontinuous across the boundary.

A similar result holds for the boundaries between $\mathcal{S}_k$ and $\mathcal{S}_j$ since the dimension of $(H_u)_{x^\perp x^\perp}$ decreases when the dimension of $\mathcal{C}_k$ decreases. ∎

This work can be extended to the time-dependent case. Consider a function $u(x, t)$, where $t \in \mathcal{T}$ denotes time. We maintain Assumptions 1, 2, and 3 for each t, and we assume that $A = A(t)$ is polyhedral. We say that it is p times continuously differentiable if the maps $t \mapsto a_k(t)$ are p times continuously differentiable, where the a_k, $k \in \mathcal{K}$, are the generators of $A(t)$ (terminology of section 19 of [12]). The following proposition extends Proposition 3.2, Remark 3.7, and Corollary 3.13. Recall that ∇ refers to the gradient with respect to x only. We write $I^{u_A}(y, t)$ for $I^{u_A(\cdot,t)}(y)$ and similarly for $\mathcal{P}_{u_A}$.

PROPOSITION 3.14. *If $t \mapsto A(t)$ and $(x, t) \mapsto \nabla u(x, t)$ are continuous on $\mathcal{T}$, $\mathrm{int}(\mathrm{dom}(u)) \times \mathcal{T}$, respectively, then so are $I^{u_A}(y, t)$ and $\mathcal{P}_{u_A}(y, t)$ on $\mathrm{dom}(I^{u_A}) := \{(y, t) : y \in \mathrm{dom}(I^{u_A(\cdot,t)}),\ t \in \mathcal{T}\}$. If A and ∇u are continuously differentiable in (x, t), then $I^{u_A}(y, t)$ and $\mathcal{P}_{u_A}(y, t)$ are similarly continuously differentiable on $\{(y, t) : y \in \mathcal{S}_k(t), t \in \mathrm{int}(\mathcal{T})\}$ for each k. The directional derivative with respect to y in the normal direction is discontinuous across the boundary of $\mathcal{S}_k(t)$.*

Proof. From Proposition 3.2 we know that I^{u_A} is continuous in y on each t-section of $\mathrm{dom}(I^{u_A})$. On the other hand, the continuity of ∇u gives that of $t \mapsto \mathcal{R}_{u_A}(t)$, cf.

(3.3), and hence of $(y,t) \mapsto \mathcal{P}_{u_A}(y,t)$, cf. the proof of Corollary 3.13. Moreover $I^{u_A}(y,t)$ is the solution x of $\nabla u(x,t) = \mathcal{P}_{u_A}(y,t)$, so the joint continuity follows from the implicit function theorem.

The piecewise continuous differentiability follows similarly. □

Remark 3.15. We can also compute the time derivative of $I^{u_A(\cdot,t)}(y)$ when A is constant. For $y \in \mathcal{R}_{u_A}(t)$

$$\frac{\partial}{\partial t} I^{u_A(\cdot,t)}(y) = -\Big(H_{u(\cdot,t)}((\nabla u(\cdot,t))^{-1}(y))\Big)^{-1} \frac{\partial}{\partial t} \nabla u(x,t)\Big|_{x=(\nabla u(\cdot,t))^{-1}(y)}. \tag{3.8}$$

For $y \in \mathcal{S}_k(t)$, $y^\perp$ and x^o are constant since A is; moreover $I^{u_A(\cdot,t)}(y) = x = x^\perp \oplus x^o$, with $x^\perp$ solution of $y^\perp = \nabla_{x^\perp} u(x^\perp \oplus x^o, t)$. Then $0 = \nabla_{x^\perp x^\perp} u(x^\perp \oplus x^o, t) \frac{\partial}{\partial t} x^\perp + \frac{\partial}{\partial t} \nabla_{x^\perp} u(x^\perp \oplus x^o, t)$, and hence

$$\frac{\partial}{\partial t} x^\perp = -\Big[\nabla_{x^\perp x^\perp} u(x^\perp \oplus x^o, t)\Big]^{-1} \frac{\partial}{\partial t} \nabla_{x^\perp} u(x^\perp \oplus x^o, t).$$

It follows that

$$\frac{\partial}{\partial t} I^{u_A(\cdot,t)}(y) = -\begin{pmatrix} ((H_{u(\cdot,t)}(I^{u_A(\cdot,t)}(y)))_{x^\perp x^\perp})^{-1} \\ \mathbf{0} \end{pmatrix} \frac{\partial}{\partial t} \nabla_{x^\perp} u(x,t)\Big|_{x=I^{u_A(\cdot,t)}(y)}. \tag{3.9}$$

We conclude that $\frac{\partial}{\partial t} I^{u_A(\cdot,t)}(y)$ is continuous except at the boundaries of $\mathcal{R}_{u_A}(t)$ and each $\mathcal{S}_k(t)$. Moreover $\frac{\partial}{\partial t} I^{u_A(\cdot,t)}(y)$ is orthogonal to $\mathcal{N}_A(I^{u_A(\cdot,t)}(y))$ since, cf. (3.9),

$$(\,\mathbf{0} \quad \mathbf{I}\,) \begin{pmatrix} ((H_{u(\cdot,t)}(I^{u_A(\cdot,t)}(y)))_{x^\perp x^\perp})^{-1} \frac{\partial}{\partial t} \nabla_{x^\perp} u(x,t)\Big|_{x=I^{u_A(\cdot,t)}(y)} \\ 0_{\dim(y^o)} \end{pmatrix} = 0,$$

and this holds *for all* $y \in \Re^+_n$ since $\mathcal{N}_A(I^{u_A(\cdot,t)}(y)) = \{0_n\}$ for $y \in \mathcal{R}_{u_A}(t)$.

4. Aggregate utility function. We now aggregate the actions of several agents into the action of a single representative agent. His utility function must opportunely weight the utility functions of the individual agents in the economy; the factor Λ below will accomplish this. We strengthen Assumption 2 so that A corresponds to "box" constraints.

Assumption 2′.

(i) $A = \prod_{i=1}^n D_i \subset \operatorname{dom}(u)$, $\quad D_i = \begin{cases} [0,\infty) \text{ or } (0,\infty) & \text{if } i \in \mathcal{M}, \\ [0,1] \text{ or } (0,1] & \text{if } i \in \tilde{\mathcal{M}}, \\ \{0\} & \text{if } i \in \mathcal{M}_o, \end{cases}$

$\mathcal{M} \cup \tilde{\mathcal{M}} \cup \mathcal{M}_o = \{1,\dots,n\}$, A closed relative to $\operatorname{dom}(u)$;

(ii) $\operatorname{dom}(\partial u_A) \subset \operatorname{int}(\operatorname{dom}(u)) \cap A$;

(iii) u is strictly increasing, strictly concave on $\operatorname{dom}(\partial u_A)$.

In (i) above the inclusion or exclusion of 0 in D_i is determined by the requirement that A be closed in the relative topology of $\operatorname{dom}(u)$. The more general case where the left end point of D_i is a_i and the right is $b_i > a_i$ is reduced to the above by shifting the origin to a and changing the scale. This will not alter the concavity or regularity of u. $\mathcal{M}$ gives the variables which are unbounded so $0^+A = \prod_{i\in\mathcal{M}} D_i$, $\tilde{\mathcal{M}}$ gives the variables which are bounded, and $\mathcal{M}_o$ gives those variables that do not affect u. Write m for $\operatorname{card}(\mathcal{M}) = \dim(0^+A)$, so if we think of 0^+A as $\Re^m_+ \oplus \{0\}$, then $-(0^+A)^\circ = \Re^m_+ \oplus \Re^{n-m}$.

We will aggregate J utility functions, $u^1, \dots, u^J$; we assume that for each j, u^j and A^j satisfy Assumptions 1, 2′, and 3 with corresponding D_i^j, $\mathcal{M}^j$, $\tilde{\mathcal{M}}^j$, $\mathcal{M}_o^j$, m^j, and μ^j. We may assume that $\bigcap_j \mathcal{M}_o^j = \emptyset$.

For $\Lambda = (\lambda_1, \dots, \lambda_J) \in \Re^J_{++}$ we will define the function $u(x;\Lambda)$ as a supremal convolution on $\Re^n$:

$$u(x;\Lambda) := \sup_{\sum_j x^j = x} \sum_{j=1}^J \lambda_j u^j_{A^j}(x^j), \tag{4.1}$$

where $x^j = (x_1^j, \dots, x_n^j)^\top \in \Re^n$.

We require more notation. Define

$$I^u(y;\Lambda) := \sum_{j=1}^J I^{u^j_{A^j}}\Big(\frac{y}{\lambda_j}\Big), \tag{4.2}$$

and set $\mathcal{D}(\Lambda) := \operatorname{dom}(I^u(\cdot\,;\Lambda)) = \bigcap_j \lambda_j \operatorname{dom}(I^{u^j_{A^j}})$, an open set since $\operatorname{dom}(I^{u^j_{A^j}}) = \operatorname{int}(\operatorname{dom}((u^j_{A^j}))^*)$. Then Proposition 3.3 implies

$$\emptyset \neq \bigcap_j \Big(\lambda_j \mu^j + \Re^{m^j}_{++} \oplus \Re^{n-m^j}\Big) \subset \mathcal{D}(\Lambda) \subset \bigcap_j \Big(\Re^{m^j}_{++} \oplus \Re^{n-m^j}\Big). \tag{4.3}$$

Observe that if $\mu^j = 0$ for all j, then $\mathcal{D}(\Lambda) = \bigcap_j \Big(\Re^{m^j}_{++} \oplus \Re^{n-m^j}\Big)$. Define

$$\mathcal{R}_o(\Lambda) := \Big\{ y \in \mathcal{D}(\Lambda) : \bigcap_{j=1}^J \operatorname{aff}\Big(\mathcal{N}_{A^j}\Big(I^{u^j_{A^j}}\Big(\frac{y}{\lambda_j}\Big)\Big)\Big) = \{0_n\}\Big\}. \tag{4.4}$$

Note that $\mathcal{R}_o(\Lambda) \supset \mathcal{D}(\Lambda) \cap \bigcup_j \lambda_j \operatorname{int}(\mathcal{R}_{u^j_{A^j}})$ (since $\mathcal{N}_{A^j}(x) = \{0_n\}$ for $x \in \operatorname{int}(A^j)$), but the inclusion may be strict. If $\mathcal{M}_o^j \neq \emptyset$, then $\operatorname{int}(\mathcal{R}_{u^j_{A^j}}) = \emptyset$, and if $\mathcal{C}_k^j$ is a vertex of A^j, then $\operatorname{aff}(\mathcal{N}_{A^j}(\mathcal{C}_k^j)) = \Re^n$.

Definition 4.1. *$\mathcal{R}_o$ is* set-continuous *if for $y \in \mathcal{R}_o(\Lambda)$ there exists $\varepsilon > 0$ such that whenever y', Λ' satisfy $\max_j \|\frac{y'}{\lambda'_j} - \frac{y}{\lambda_j}\| < \varepsilon$, then $y' \in \mathcal{R}_o(\Lambda')$.*

Lemma 4.2. *$\mathcal{R}_o(\Lambda) \subset \Re^n_{++}$ is open, and $\mathcal{R}_o$ is set-continuous.*

Proof. Suppose $y \in \mathcal{D}(\Lambda)$ but $y \notin \Re^n_{++}$, so $y_i \leq 0$ for some $i \notin \bigcup_j \mathcal{M}^j$; cf. (4.3). Since u^j is strictly increasing on $\operatorname{dom}(\partial u^j_{A^j})$, then $\mathcal{R}_{u^j_{A^j}} = \nabla u^j(\operatorname{dom}(\partial u^j_{A^j})) \subset \Re^n_{++}$, so $\mathcal{P}_{u^j_{A^j}}$ projects $\frac{y}{\lambda_j}$ into $\Re^n_{++}$. Then $\vec{n}^j$ of Lemma 3.8 can be written as $\vec{n}^j = \sum_{\ell \in \mathcal{L}} \alpha_\ell^j \tilde{e}_\ell^j$, where $\alpha_\ell^j > 0$, $\tilde{e}_\ell^j$ is plus or minus a standard basis vector, and the convex hull of $\{\tilde{e}_\ell : \ell \in \mathcal{L}\}$ generates the convex cone $\mathcal{N}_{A^j}(I^{u^j_{A^j}}(\frac{y}{\lambda_j}))$. Because $\mathcal{P}_{u^j_{A^j}}$ projects $\frac{y}{\lambda_j}$ into $\Re^n_{++}$ and $y_i \leq 0$, then $\vec{e}_i$ is one of the $\tilde{e}_\ell^j$ for all j, and hence $\vec{e}_i \in \mathcal{N}_{A^j}(I^{u^j_{A^j}}(\frac{y}{\lambda_j}))$ for all j. From the definition of $\mathcal{R}_o$ it follows that $y \notin \mathcal{R}_o(\Lambda)$, and hence $\mathcal{R}_o(\Lambda) \subset \Re^n_{++}$.

Next we show that $\mathcal{R}_o(\Lambda)$ is open. $\mathcal{D}(\Lambda)$ is open; take $y \in \mathcal{R}_o(\Lambda)$, and write $\hat{x}^j(\frac{y}{\lambda_j})$ for $I^{u^j_{A^j}}(\frac{y}{\lambda_j})$. Recall that we set $\mathcal{C}_0^j = \operatorname{int}(A^j)$; cf. Definition 3.12. If $\hat{x}^j(\frac{y}{\lambda_j}) \in \operatorname{int}(A^j)$, then so is $\hat{x}^j(\frac{y'}{\lambda_j})$ for y' a sufficiently small pertubation of y; cf. Proposition 3.2. If $\hat{x}^j(\frac{y}{\lambda_j}) \in \mathcal{C}_k^j$, then $\hat{x}^j(\frac{y'}{\lambda_j}) \in \mathcal{C}_k^j$ or $\hat{x}^j(\frac{y'}{\lambda_j}) \in \mathcal{C}_{k'}^j$ with $\operatorname{cl}(\mathcal{C}_{k'}^j) \supset \mathcal{C}_k^j$ since $\mathcal{C}_k^j$ is relatively open (think of $\mathcal{C}_k^j$ as an edge with the end point removed and $\mathcal{C}_{k'}^j$ as the relative interior

of an adjoining face). In the latter case it follows that $\mathcal{N}_{A^j}(\hat{x}^j(\frac{y'}{\lambda_j})) \subset \mathcal{N}_{A^j}(\hat{x}^j(\frac{y}{\lambda_j}))$. In any case $\bigcap_{j=1}^J \operatorname{aff}(\mathcal{N}_{A^j}(I^{u^j_{A^j}}(\frac{y'}{\lambda_j}))) \subset \bigcap_{j=1}^J \operatorname{aff}(\mathcal{N}_{A^j}(I^{u^j_{A^j}}(\frac{y}{\lambda_j}))) = \{0_n\}$, and hence $y' \in \mathcal{R}_o(\Lambda)$.

The set continuity of $\mathcal{R}_o$ is established as in the previous paragraph but with $\frac{y'}{\lambda_j}$ replaced by $\frac{y'}{\lambda'_j}$ since $y \in \mathcal{D}(\Lambda)$ if and only if $\frac{y}{\lambda_j} \in \operatorname{dom}(I^{u^j_{A^j}})$ for all j, and $\operatorname{dom}(I^{u^j_{A^j}})$ is open. □

Recall that $\operatorname{bdy}(A^j)$ decomposes into the relatively open pieces $\mathcal{C}^j_k$, $k > 0$. We will demand a bit more regularity of u^j:

$$\text{if } \mathcal{C}^j_k \cap \operatorname{dom}(\partial u^j) \neq \emptyset, \text{ then } \mathcal{C}^j_k \subset \operatorname{dom}(\partial u^j). \tag{4.5}$$

THEOREM 4.3. *Assume that u^j, $j = 1, \dots, J$, are utility functions satisfying Assumptions 1, 2′, and 3. If (4.5) holds for each j, then the following hold:*

(i) *For each $\Lambda \in \Re^J_{++}$, $u(\cdot\,;\Lambda) : \Re^n \mapsto [-\infty, \infty)$ is a closed, proper, concave, nondecreasing function on $\Re^n$ with $\operatorname{dom}(u(\cdot;\Lambda)) = A := \sum_{j=1}^J A^j$. For each $x \in A$ there exist $\hat{x}^j \in A^j$ such that*

$$x = \sum_j \hat{x}^j, \quad u(x;\Lambda) = \sum_j \lambda_j u^j_{A^j}(\hat{x}^j). \tag{4.6}$$

Moreover $I^u(\cdot\,;\Lambda)$ is the inverse of $\partial u(\cdot\,;\Lambda)$ and is monotone. $u(\cdot\,;\Lambda)$ is strictly concave on $\operatorname{dom}(\partial u(\cdot\,;\Lambda)) = \operatorname{im}(I^u(\cdot\,;\Lambda))$.

(ii) *$\operatorname{im}(I^u(\cdot\,;\Lambda)) = \tilde{A} := \sum_j \operatorname{dom}(\partial u^j_{A^j})$ is convex. For $x \in \tilde{A}$ there exists $y \in (I^u(\cdot\,;\Lambda))^{-1}(x)$ such that*

$$\hat{x}^j = I^{u^j_{A^j}}\left(\frac{y}{\lambda_j}\right). \tag{4.7}$$

(iii) *$u(\cdot;\Lambda)$ is continuously differentiable on $A^u(\Lambda) := I^u(\mathcal{R}_o(\Lambda);\Lambda)$. This set is dense in A. Moreover $(\nabla u(\cdot;\Lambda))^{-1} = I^u(\cdot;\Lambda)$ on $\nabla u(A^u(\Lambda);\Lambda) = \mathcal{R}_o(\Lambda)$, so $I^u(y;\Lambda)$ is a continuous, monotone extension of $(\nabla u(\cdot;\Lambda))^{-1}$, strictly monotone on $\operatorname{cl}(\mathcal{R}_o(\Lambda)) \cap \mathcal{D}(\Lambda)$. $u(\cdot\,;\Lambda)$ is strictly increasing on $A^u(\Lambda)$. For $y \in \mathcal{R}_o(\Lambda)$ we have*

$$\nabla u(I^u(y;\Lambda);\Lambda) = y. \tag{4.8}$$

(iv) *For each $x \in A^u(\Lambda)$,*

$$\nabla u(x;\Lambda) = \lambda_j \nabla u^j(\hat{x}^j) - \vec{n}^j(\hat{x}^j), \quad j = 1, \dots, J, \tag{4.9}$$

where $\vec{n}^j(\hat{x}^j) \in \mathcal{N}_{A^j}(\hat{x}^j)$ and for each $i = 1, \dots, n$ there exists $j(i)$ such that $[\vec{n}^{j(i)}(\hat{x}^{j(i)})]_i = 0$.

(v) *If for each j, $\nabla u^j(x)$ is continuously differentiable, then $\nabla u(x;\Lambda)$ is piecewise continuously differentiable.*

Proof. Define the convex function $f^j := -\lambda_j u^j_{A^j}$. Recall that for a concave function u, $(-u)^*(y) = -u^*(-y)$. From this and (4.3) it follows that

$$\bigcap_j \operatorname{int}\Big(\operatorname{dom}((f^j)^*)\Big) = -\bigcap_j \operatorname{int}\Big(\operatorname{dom}((\lambda_j u^j_{A^j})^*)\Big) \neq \emptyset.$$

It follows from Theorem 16.4 of [12] that u is closed, concave, and proper and (4.6) holds with $\hat{x}^j \in A^j$. Evidently $\operatorname{dom}(u(\cdot\,;\Lambda)) = A$. Since the constraint in (4.1) can be relaxed to $\sum x^j \leq x$, then it follows that u is nondecreasing.

The same theorem also implies

$$u^*(y;\Lambda) = \sum_j (\lambda_j u^j_{A^j})^*(y) = \sum_j \lambda_j u^{j*}_{A^j}\Big(\frac{y}{\lambda_j}\Big).$$

Now Theorem 23.8 of [12] implies that the superdifferential of $u^*(\cdot;\Lambda)$ is

$$\partial u^*(y;\Lambda) = \sum_j \partial(\lambda_j u^j_{A^j})^*(y) = \sum_j \partial u^{j*}_{A^j}\Big(\frac{y}{\lambda_j}\Big) = \sum_j I^{u^j_{A^j}}\Big(\frac{y}{\lambda_j}\Big) = I^u(y;\Lambda); \tag{4.10}$$

cf. (4.2). Thus $\partial u^*(y;\Lambda)$ is single-valued on $\mathcal{D}(\Lambda)$. We can now conclude that the concave function u^* is continuously differentiable, cf. Theorems 25.1 and 25.5 of [12], on $\mathcal{D}(\Lambda) = \operatorname{dom}(I^u) = \operatorname{dom}(\partial u^*)$, and as usual $\partial u^* = (\partial u)^{-1}$; i.e., I^u is the inverse of ∂u. Furthermore as the supergradient of a concave function, I^u is monotone.

Since ∂u^* is single-valued, then $\partial u(x) \cap \partial u(x') = \emptyset$ for $x \neq x'$. As in the proof of Theorem 26.3 of [12], this implies that u is strictly concave on $\operatorname{dom}(\partial u) = \operatorname{im}(I^u)$. That this set is convex will follow from (ii). This establishes (i) of the theorem.

We turn to (ii). Notice that $\tilde{A}$ is convex since, cf. (3.1),

$$\tilde{A} = \sum_j [\operatorname{dom}(\partial u^j) \cap A^j] = \sum_j [A^j \cap \operatorname{int}(\operatorname{dom}(u^j))].$$

As $\operatorname{im}(I^u) \subset \tilde{A}$, cf. (4.2) and Proposition 3.2, it suffices to show the reverse inclusion. In the process we shall establish (4.7). For $x \in \tilde{A}$ we consider the maximization problem whose value is $u(x;\Lambda)$:

$$\begin{cases} \sup\Big\{f(x^1,\dots,x^J) : (x^1,\dots,x^J) \in C, \quad \sum_j x^j = x\Big\}, \\ f(x^1,\dots,x^J) := \sum_{j=1}^J \lambda_j u^j_{A^j}(x^j), \quad C := \prod_{j=1}^J A^j. \end{cases} \tag{4.11}$$

Note that the unique solution is $(\hat{x}^1,\dots,\hat{x}^J)$. Let us first modify the problem to ensure that there is always a feasible solution in $\operatorname{ri}(C)$ so that we can apply some Lagrange multiplier results.

Consider $x \in \tilde{A} \subset A = \sum_j A^j$. If some components of x have certain values, to be made precise below, the constraints in (4.11) will fix the corresponding components of x^j in any decomposition $x = \sum_j x^j$, with $x^j \in A^j$, and hence these components may be removed from (4.11), reducing x to $\bar{x}$ and x^j to $\bar{x}^j$. We will show that for any decomposition $\bar{x} = \sum_j \bar{x}^j$, with $\bar{x}^j \in \bar{A}^j$, there is a decomposition $\bar{x} = \sum_j \tilde{x}^j$, with $\tilde{x}^j \in \operatorname{ri}(\bar{A}^j)$, where $\bar{A}^j$ is the set corresponding to A^j after dropping the fixed components.

We now find the components which will be dropped. Suppose that $x_i = 0$; then for any decomposition $x = \sum_j x^j$, with $x^j \in A^j$, i.e., $x^j_i \in D^j_i$ so $x^j_i \geq 0$, we must have $x^j_i = 0$, and hence there is no maximization in (4.11) over x^j_i, $j = 1,\dots,J$. Moreover if the largest possible value of x_i is finite, i.e., if d^j_i, the right end point of D^j_i, is finite for *all* j, equivalently, $d^j_i \in \{0,1\}$, equivalently, $i \notin \bigcup_j \mathcal{M}^j$, equivalently, $i \in \bigcap_j (\tilde{\mathcal{M}}^j \cup \mathcal{M}^j_o)$, then $0 \leq x_i \leq \sum_j d^j_i = \operatorname{card}\{j : i \in \mathcal{M}^j\}$. Now suppose that for such an i, $x_i = \operatorname{card}\{j : i \in \mathcal{M}^j\}$; then for any decomposition $x = \sum_j x^j$, we must have $x^j_i = d^j_i$; i.e., again in (4.11) there is no maximization over x^j_i, $j = 1,\dots,J$.

We summarize as follows: Set $\mathcal{I}(x) := \{i \in \{1,\dots,n\} : x_i = 0\} \cup \{i \in \bigcap_j (\tilde{\mathcal{M}}^j \cup \mathcal{M}^j_o) : x_i = \sum_j d^j_i\}$, and let $n_o = \operatorname{card}(\mathcal{I}(x))$, where card denotes cardinality. Observe

that $\mathcal{I}(x) = \emptyset$, i.e., $n_o = 0$, if and only if $x \in \text{int}(A)$, and $n_o = n$ if and only if x is an extreme point of A.

Set $\bar{n} = n - n_o$, decompose $\Re^n = \Re^{\bar{n}} \oplus \Re^{n_o}$, and write $x = \bar{x} \oplus x^o$. For any decomposition $x = \sum_j x^j \in \sum_j A^j$ and any $i \in \mathcal{I}(x)$ we have $x_i^{jo} = 0$ if $x_i = 0$ or $x_i^{jo} = d_i^j$ if $x_i = \sum_j d_i^j$, i.e., x^{jo} fixed, so there is no maximization to be done over these components of x^j. Define $\bar{A}^j$ as the orthogonal projection of A^j onto $\Re^{\bar{n}}$. We can then replace the maximization problem by

$$\begin{cases} \sup\left\{f(\bar{x}^1, \ldots, \bar{x}^J) : (\bar{x}^1, \ldots, \bar{x}^J) \in \bar{C}, \quad \sum_{j=1}^J \bar{x}^j = \bar{x}\right\}, \\ f(\bar{x}^1, \ldots, \bar{x}^J) := \sum_{j=1}^J \lambda_j u_{A^j}^j(\bar{x}^j \oplus x^{jo}), \quad \bar{C} := \prod_{j=1}^J \bar{A}^j. \end{cases}$$

We may think of the new problem as having dropped the coordinates in $\mathcal{I}(x)$. If $\hat{x}^j$ decomposes as $\hat{x}^j = \bar{x}^j \oplus x^{jo}$, then $(\bar{x}^1, \ldots, \bar{x}^J) \in \bar{C}$ is feasible for the new problem. For any $i \notin \mathcal{I}(x)$ not all $\bar{x}_i^j$ can be 0, so if any are, we can perturb the $\bar{x}_i^j$ slightly to make them all positive while maintaining the constraint $\sum_j \bar{x}_i^j = \bar{x}_i$. Similarly not all $\bar{x}_i^j$ can be d_i^j, and hence we can perturb the nonzero $\bar{x}_i^j$ (i.e., j such that $i \in \tilde{\mathcal{M}}^j$ since for $i \in \mathcal{M}_o^j$ we must have $\bar{x}_i^j = 0$) so that all are less than one. This provides a feasible solution in $\text{ri}(\bar{C})$. The problem can be put in standard form as follows: $\bar{x} \in \Re^{\bar{n}}$ is fixed; let $f_i(\bar{x}^1, \ldots, \bar{x}^J) := \bar{x}_i - \sum_{j=1}^J \bar{x}_i^j,\ i = 1, \ldots, \bar{n}$. Then $\text{dom}(f_i) = \Re^{\bar{n}J}$. Moreover if $g_j(\bar{x}^1, \ldots, \bar{x}^J) := \lambda_j u_{A^j}^j(\bar{x}^j \oplus x^{jo})$, then the problem is to maximize $\sum_j g_j$ subject to $f_i = 0,\ i = 1, \ldots, \bar{n}$.

Below we will introduce the Lagrange multiplier $\bar{y}$. The following calculations will aid us in identifying it. Observe that $\bigcap_{j=1}^J \text{ri}(\text{dom}(g_j)) = \bigcap_{j=1}^J \text{ri}(\bar{A}^j \oplus \Re^{\bar{n}(J-1)}) = \prod_{j=1}^J \text{ri}(\bar{A}^j) \neq \emptyset$. From this and Theorem 23.8 of [12] follows that

$$\partial\left[\sum_{j=1}^J g_j + \sum_{i=1}^{\bar{n}} \bar{y}_i f_i\right] = \sum_{j=1}^J \partial g_j + \sum_{i=1}^{\bar{n}} \bar{y}_i \partial f_i.$$

We compute

$$\sum_j \partial g_j(\bar{x}^1, \ldots, \bar{x}^J) = \left(\lambda_1 \partial_{\bar{x}^1} u_{A^1}^1(\bar{x}^1 \oplus x^{o1}), \ldots, \lambda_J \partial_{\bar{x}^J} u_{A^J}^J(\bar{x}^J \oplus x^{Jo})\right),$$
$$\sum_{i=1}^{\bar{n}} \bar{y}_i \partial f_i(\bar{x}^1, \ldots, \bar{x}^J) = -\sum_{i=1}^{\bar{n}} \bar{y}_i(\vec{e}_i, \ldots, \vec{e}_i) = -(\bar{y}, \ldots, \bar{y}).$$

Here $\partial_{\bar{x}^j}$ refers to supergradients with respect to $\bar{x}^j \in \Re^{\bar{n}}$.

Now Corollary 28.2.2 and Theorem 28.3 of [12] allow us to infer that there exists a multiplier $\bar{y} \in \Re^{\bar{n}}$ such that $0 \in \partial[\sum_j g_j + \sum_i \bar{y}_i f_i]$; i.e., for every j (recalling $\hat{x}^j = \bar{\hat{x}}^j \oplus \hat{x}^{jo}$)

$$\bar{y} \in \lambda_j \partial_{\bar{x}^j} u_{A^j}^j(\hat{x}^j) = \lambda_j[\partial_{\bar{x}^j} u^j(\hat{x}^j) - \mathcal{N}_{\bar{A}^j}(\bar{\hat{x}}^j)] = \lambda_j \nabla_{\bar{x}^j} u^j(\hat{x}^j) - \lambda_j \mathcal{N}_{\bar{A}^j}(\bar{\hat{x}}^j).$$

Note that $\nabla_{\bar{x}^j}$ refers to the gradient with respect to $\bar{x}^j \in \Re^{\bar{n}}$. The last equality holds by Theorem 25.6 of [12]. Hence for some $\vec{\bar{n}}^j \in \mathcal{N}_{\bar{A}^j}(\bar{\hat{x}}^j)$ we have

$$\frac{\bar{y}}{\lambda_j} + \vec{\bar{n}}^j = \nabla_{\bar{x}^j} u^j(\hat{x}^j).$$

Since $x \in \tilde{A}$, then $x = \sum_j x^j$, with $x^j \in \text{dom}(\partial u_{A^j}^j)$, for some $\{x^j\}$. But $x^{jo} = \hat{x}^{jo}$, so if $n_o > 0$, then both $x^j, \hat{x}^j \in \mathcal{C}_k^j$ for some k. Now (4.5) implies that

$\hat{x}^j \in \mathrm{dom}(\partial u^j_{A^j})$, and hence $\nabla u^j(\hat{x}^j) = \nabla_{\bar{x}^j} u^j(\hat{x}^j) \oplus \nabla_{x^{jo}} u^j(\hat{x}^j)$ exists. In addition, $\mathcal{N}_{A^j}(\hat{x}^j)$ decomposes into $\mathcal{N}_{\bar{A}^j}(\hat{\bar{x}}^j) \oplus \mathcal{N}_o$, where $\mathcal{N}_o$ (same for all j) is an orthant of $\Re^{n_o}$. Since $\nabla_{x^{jo}} u^j(\hat{x}^j) \in \Re^{n_o}$, then $\bigcap_j \lambda_j[\nabla_{x^{jo}} u^j(\hat{x}^j) - \mathcal{N}_o] \neq \emptyset$. Let y^o be a point in this set. If we set $y := \bar{y} \oplus y^o$, then for some $\vec{n}^j_y(\hat{x}^j) \in \mathcal{N}_{A^j}(\hat{x}^j)$,

$$\frac{y}{\lambda_j} + \vec{n}^j_y(\hat{x}^j) = \nabla u^j(\hat{x}^j). \tag{4.12}$$

Lemma 3.8 and Corollary 3.9 imply that $\hat{x}^j = I^{u^j_{A^j}}(\frac{y}{\lambda_j})$, so we have found y such that $I^u(y;\Lambda) = x$; i.e., $y \in (I^u(\cdot;\Lambda))^{-1}(x)$, and (4.7) holds.

We turn to the proof of (iii). Fix Λ. As $\partial u = (\partial u^*)^{-1} = (I^u)^{-1}$ and u is differentiable where ∂u is single valued, then u is differentiable where I^u is one-to-one. We will show that $I^u(\cdot\,;\Lambda)$ is one-to-one on $\mathcal{R}_o(\Lambda) \subset \Re^n_{++}$, cf. Lemma 4.2, by showing that it is strictly monotone on $\mathrm{cl}(\mathcal{R}_o(\Lambda))$. Fix j for now. If $y \in \lambda_j \mathcal{R}_{u^j_{A^j}}$ (cf. (3.3)), then $I^{u^j_{A^j}}(\frac{y}{\lambda_j}) = (\nabla u^j)^{-1}(\frac{y}{\lambda_j})$ is strictly monotone in y.

Recall $\mathcal{S}^j_k = (I^{u^j_{A^j}})^{-1}(\mathcal{C}^j_k)$; we can decompose $\Re^n = \mathrm{aff}(\mathcal{N}_{A^j}(\mathcal{C}^j_k))^\perp \oplus \mathrm{aff}(\mathcal{N}_{A^j}(\mathcal{C}^j_k))$ and accordingly for $y \in \mathcal{S}^j_k$, $y = y^{j\perp} \oplus y^{jo}$, and for $x \in \mathcal{C}^j_k$, $x = x^{j\perp} \oplus x^{jo}$ with x^{jo} depending only on j and k, not on x. We set $\mathcal{J}(\mathcal{C}^j_k) := \{i : x_i = x^{j\perp}_i, x \in \mathcal{C}^j_k\}$, i.e., the set of component indices of $x^{j\perp}$ or of $y^{j\perp}$.

For $y \in \lambda_j \mathcal{S}^j_k$, $I^{u^j_{A^j}}(\frac{y}{\lambda_j}) = (I^{u^j_{A^j}}(\frac{y}{\lambda_j}))^\perp \oplus (I^{u^j_{A^j}}(\frac{y}{\lambda_j}))^o := x^{j\perp}(\frac{y^{j\perp}}{\lambda_j}) \oplus x^{jo} \in \mathcal{C}^j_k$ with x^{jo} constant, i.e., $\mathcal{C}^j_k = \{x^\perp \oplus x^{jo} : x^\perp \in \prod_{i \in \mathcal{J}(C^j_k)} D^j_i\}$, and $x^{j\perp}(\frac{y^{j\perp}}{\lambda_j})$ is the solution of $\nabla_{x^\perp} u^j_{A^j}(x^\perp \oplus x^{jo}) = \frac{y^{j\perp}}{\lambda_j}$. We conclude that for $y \in \lambda_j \mathcal{S}^j_k$, $(I^{u^j_{A^j}}(\frac{y}{\lambda_j}))^\perp$ is constant in y^{jo} and strictly monotone in $y^{j\perp}$ and $(I^{u^j_{A^j}}(\frac{y}{\lambda_j}))^o$ is constant.

If $y \in \mathrm{bdy}(\lambda_j \mathcal{S}^j_k)$ and $y_n \to y$ with $y_n \in \lambda_j \mathcal{S}^j_k$, then $\mathcal{N}_{A^j}(I^{u^j_{A^j}}(\frac{y_n}{\lambda_j})) \subset \mathcal{N}_{A^j}(I^{u^j_{A^j}}(\frac{y}{\lambda_j}))$, i.e., $I^{u^j_{A^j}}(\frac{y_n}{\lambda_j}) \in \mathcal{C}^j_k$, $I^{u^j_{A^j}}(\frac{y}{\lambda_j}) \in \mathcal{C}^j_{k'}$ with $\mathrm{cl}(\mathcal{C}^j_k) \supset \mathcal{C}^j_{k'}$. Since $\mathrm{aff}(\mathcal{N}_{A^j}(I^{u^j_{A^j}}(\frac{y_n}{\lambda_j}))) = \mathrm{aff}(\mathcal{N}_{A^j}(\mathcal{C}^j_k))$ is independent of n, we can use this decomposition, i.e., the decomposition corresponding to $\mathcal{J}(\mathcal{C}^j_k)$. We may take y_n such that $y^{j\perp}_n = y^{j\perp}$. Passing to the limit as $n \to \infty$ allows us to establish the monotonicity properties of $I^{u^j_{A^j}}(\frac{y}{\lambda_j})$ on $\mathrm{cl}(\lambda_j \mathcal{S}^j_k)$.

Recall that $\mathcal{S}^j_0 := \mathrm{int}(\mathcal{R}_{u^j_{A^j}})$, possibly empty; cf. Definition 3.12. For $y \in \mathcal{D}(\Lambda)$, y lies in $\lambda_j \mathcal{S}^j_{k(j)}$ for each j and some selection $k(j) \in \{0, 1, \ldots, K\}$. For $y, y' \in \bigcap_j \mathrm{cl}(\lambda_j \mathcal{S}^j_{k(j)})$, $y \neq y'$, we have

$$\begin{aligned}
(y - y')^\top (I^u(y;\Lambda) - I^u(y';\Lambda)) &= \sum_j (y^{j\perp} - y'^{j\perp})^\top \left(\left(I^{u^j_{A^j}}\left(\frac{y^j}{\lambda_j}\right)\right)^\perp - \left(I^{u^j_{A^j}}\left(\frac{y'^j}{\lambda_j}\right)\right)^\perp \right) \\
&\quad + \sum_j (y^{jo} - y'^{jo})^\top \left(\left(I^{u^j_{A^j}}\left(\frac{y^j}{\lambda_j}\right)\right)^o - \left(I^{u^j_{A^j}}\left(\frac{y'^j}{\lambda_j}\right)\right)^o \right) \\
&= \sum_j (y^{j\perp} - y'^{j\perp})^\top \left(\left(I^{u^j_{A^j}}\left(\frac{y^j}{\lambda_j}\right)\right)^\perp - \left(I^{u^j_{A^j}}\left(\frac{y'^j}{\lambda_j}\right)\right)^\perp \right) \\
&\leq 0.
\end{aligned} \tag{4.13}$$

In fact the last inequality is strict unless $y^{j\perp} - y'^{j\perp} = 0$ for all j because $(I^{u^j_{A^j}}(\frac{y^j}{\lambda_j}))^\perp$ is strictly monotone in $\frac{y^{j\perp}}{\lambda_j}$ on $\mathcal{S}^j_k$.

If $y, y' \in \mathrm{cl}(\mathcal{R}_o(\Lambda)) \cap (\bigcap_j \mathrm{cl}(\lambda_j \mathcal{S}^j_{k(j)}))$, $y \neq y'$, then $y - y' = (y^{j\perp} - y'^{j\perp}) \oplus (y^{jo} - y'^{jo})$ and $(y^{j\perp} - y'^{j\perp})_i = (y - y')_i$ for $i \in \mathcal{J}(\mathcal{C}^j_{k(j)})$. Note we use the *same* decomposition on $\mathrm{bdy}(\mathcal{S}^j_{k(j)})$ as in the interior; cf. the monotonicity argument of $I^{u^j_{A^j}}$ above. On $\mathcal{R}_o(\Lambda)$ we have $\bigcap_j \mathrm{aff}(\mathcal{N}_{A^j}(I^{u^j_{A^j}}(\frac{y}{\lambda_j}))) = \{0_n\}$, i.e., $\sum_j \mathrm{aff}(\mathcal{N}_{A^j}(I^{u^j_{A^j}}(\frac{y}{\lambda_j})))^\perp = \Re^n$, equivalently, $\bigcup_j \mathcal{J}(\mathcal{C}^j_{k(j)}) = \{1, \ldots, n\}$. It follows that if $y^{j\perp} - y'^{j\perp} = 0$ for all j, then $y - y' = 0$, a contradiction, so $I^u(\cdot\,;\Lambda)$ is strictly monotone on $\bigcap_j \mathrm{cl}(\lambda_j \mathcal{S}^j_{k(j)})$. Thus the inequality in (4.13) must be strict.

The argument extends to any y' for which the line segment $[y, y']$ contains a subsegment which lies in $\mathrm{cl}(\mathcal{R}_o(\Lambda))$ because I^u is monotone. To see this, suppose $[y, z]$ is such a subsegment lying in one of the pieces used in the previous paragraph. Then

$$\begin{aligned}
&(y - y')^\top (I^u(y;\Lambda) - I^u(y';\Lambda)) \\
&= \frac{\|y - y'\|}{\|y - z\|}(y - z)^\top (I^u(y;\Lambda) - I^u(z;\Lambda)) + \frac{\|y - y'\|}{\|z - y'\|}(z - y')^\top (I^u(z;\Lambda) - I^u(y';\Lambda)) \\
&< 0
\end{aligned}$$

since the first term on the right is negative and the second is nonpositive.

It follows that $I^u(\cdot;\Lambda)$ is strictly monotone on $\mathrm{cl}(\mathcal{R}_o(\Lambda))$ and only there. Hence $I^u(\cdot\,;\Lambda)$ is one-to-one on $\mathcal{R}_o(\Lambda)$, and so $\partial u(x)$ is single-valued for $x \in I^u(\mathcal{R}_o(\Lambda);\Lambda) = A^u(\Lambda)$ (and only there) and u is continuously differentiable there; cf. Theorem 25.5 of [12]. As the set of points where u is differentiable, $A^u(\Lambda)$ is dense in $\mathrm{int}(\mathrm{dom}(u)) = \mathrm{int}(A)$. Since $\bigcap_j \mathcal{M}^j_o = \emptyset$, then $\mathrm{aff}(A) = \Re^n$, so $\mathrm{int}(A)$ is dense in A; i.e., $A^u(\Lambda)$ is dense in A. As $I^u(\cdot\,;\Lambda)^{-1} = \nabla u(\cdot\,;\Lambda)$ on $I^u(\mathcal{R}_o(\Lambda);\Lambda)$, then (4.8) holds on $\mathcal{R}_o(\Lambda)$.

$\nabla u(\cdot\,;\Lambda) \in \mathcal{R}_o(\Lambda) \subset \Re^n_{++}$, cf. Lemma 4.2, so u is strictly increasing on $A^u(\Lambda)$; cf. (2.1).

Next we prove (iv). Let $x \in A^u(\Lambda)$, $y := (I^u(\cdot\,;\Lambda))^{-1}(x)$, a singleton for $x \in A^u(\Lambda)$, and let $\hat{x}^j$ be given by (4.7); then (4.8) and (4.12) imply

$$\nabla u(x;\Lambda) = \; y = \lambda_j \nabla u^j(\hat{x}^j) - \lambda_j \vec{n}^j_y(\hat{x}^j) \quad \text{for all } j,$$

and hence (4.9) follows. Fix i. If $[\vec{n}^j_y(\hat{x}^j)]_i \neq 0$ for all j, then $\vec{e}_i \in \mathrm{aff}(\mathcal{N}_{A^j}(\hat{x}^j))$ for all j, since $\mathrm{aff}(\mathcal{N}_{A^j}(\hat{x}^j))$ is the subspace spanned by a subset (depending on j) of the standard basis vectors. It follows that $\vec{e}_i \in \bigcap_j \mathrm{aff}(\mathcal{N}_{A^j}(\hat{x}^j))$, a contradiction, since $y \in \mathcal{R}_o(\Lambda)$. Hence for every i there exists j such that $u_{x_i}(x;\Lambda) = \lambda_j u^j_{x_i}(\hat{x}^j)$.

Finally turning to (v), (4.2), Proposition 3.2, and Corollary 3.13 imply that $I^u(\cdot\,;\Lambda)$ is continuously differentiable except on $\bigcup_{k,j} \mathrm{bdy}(\mathcal{S}^j_k)$. By the inverse function theorem $\nabla u(\cdot\,;\Lambda)$ is piecewise continuously differentiable. □

We are able to extend $\nabla u(\cdot;\Lambda)$ to $\tilde{A}$ if we control the behavior of ∇u^j near points on $\mathrm{bdy}(A)$ not in $\mathrm{dom}(\partial u^j)$, i.e., where $\|\nabla u^j\| = \infty$. Of necessity such points (if any) are on $\mathrm{bdy}(\Re^n_{++})$, i.e., constitute some of the faces of $\mathrm{bdy}(A)$; cf. (4.5). The condition is as follows: For every j and bounded sequence $\{x^k\} \subset A^j$,

$$\text{if } \lim_{k\to\infty} u^j_{x_i}(x^k) = \infty \text{ for some } i, \text{ then } \lim_{k\to\infty} x^k_i = 0. \tag{4.14}$$

COROLLARY 4.4. *Assume* (4.14). *Then* $\nabla u(\cdot\,;\Lambda)$ *can be extended continuously to* $\tilde{A}$ *such that on* $\tilde{A}$ *we have* $I^u(\nabla u(x;\Lambda);\Lambda) = x$. *Moreover* (4.8) *holds for* $y \in \mathrm{cl}(\mathcal{R}_o(\Lambda))$ *and* (4.9) *holds on* $\tilde{A}$.

Proof. The extensions of the equations follows by continuity once $\nabla u(\cdot\,;\Lambda)$ has been extended. For $x \in \tilde{A}$ and any sequence $x^k \to x$, $x^k \in A^u(\Lambda)$, there exist $y^k \in \mathcal{R}_o(\Lambda) \subset \Re^n_{++}$ with $x^k = I^u(y^k;\Lambda) = \sum_j I^{u^j_{A^j}}(\frac{y^k}{\lambda_j}) = \sum_j \hat{x}^{k,j}$. Then $\|\hat{x}^{k,j}\| \le \|x\| + 1$ for k sufficiently large. Now $\hat{x}^{k,j} \in \mathcal{C}^j_{l_j}$ is of the form $\hat{x}^{k,j} = \hat{x}^{k,j\perp} \oplus x^{l_j o}$, with $x^{l_j o} \in \operatorname{aff}(\mathcal{N}_{A^j}(\mathcal{C}^j_{l_j}))$, and we can decompose $y^k = y^{k,j\perp} \oplus y^{k,jo}$. By extracting subsequences J times we may assume that l_j is independent of k since the number of faces is finite. Then $\lambda_j \nabla u^j(\hat{x}^{k,j}) = y^k + \vec{n}^j(\hat{x}^{k,j})$ for all j by (4.9). We claim that $\{y^k\}$ is bounded.

Assume that for a subsequence $\|y^k\| \to \infty$. Since $\sum \operatorname{aff}(\mathcal{N}_{A^j}(\mathcal{C}^j_{l_j}))^\perp = \Re^n$, then there exists j_o such that $\|y^{k,j_o\perp}\| \to \infty$, i.e., $y^k_i \to \infty$ for some i such that $\vec{n}^{j_o}_i(\hat{x}^{k,j_o}) = 0$ since $\vec{n}^{j_o}(\hat{x}^{k,j_o}) \in \operatorname{aff}(\mathcal{N}_{A^{j_o}}(\mathcal{C}^{j_o}_{l_{j_o}}))$. Then $u^{j_o}_{x_i}(\hat{x}^{k,j_o}) = y^k_i/\lambda_{j_o} \to \infty$, cf. (4.9), and hence $\hat{x}^{k,j_o}_i \to 0$ by (4.14). For all j we have $\lambda_j u^j_{x_i}(\hat{x}^{k,j}) = y^k_i + \vec{n}^j_i(\hat{x}^{k,j})$, and the geometry implies that $\vec{n}^j_i(\hat{x}^{k,j}) \ge 0$ unless $\hat{x}^{k,j}_i = 0$. Again by taking subsequences, we can arrange that either $\hat{x}^{k,j}_i = 0$ for all k so that $\lim_k \hat{x}^{k,j}_i = 0$ or $\vec{n}^j_i(\hat{x}^{k,j}) \ge 0$ for all k so $u^j_{x_i}(\hat{x}^{k,j}) \ge y^k_i/\lambda_j \to \infty$, and hence again $\lim_k \hat{x}^{k,j}_i = 0$. It follows that $x_i = \lim_k x^k_i = \lim_k \sum_j \hat{x}^{k,j}_i = 0$ so $x \notin \tilde{A} = \sum_j \operatorname{dom}(\partial u^j_{A^j})$ since (4.14) implies that $\{x_i = 0\} \cap \operatorname{dom}(\partial u^{j_o}_{A^{j_o}}) = \emptyset$. This contradiction implies that $\{y^k\}$ is bounded.

We can conclude that $\{y^k\}$ contains convergent subsequences, again denoted by $\{y^k\}$, with limit $y \in \operatorname{cl}(\mathcal{R}_o(\Lambda))$. Hence

$$x = \lim_k x^k = \lim_k I^u(y^k;\Lambda) = \lim_k \sum_j I^{u^j_{A^j}}\Big(\frac{y^k}{\lambda_j}\Big) = \sum_j I^{u^j_{A^j}}\Big(\frac{y}{\lambda_j}\Big) = I^u(y;\Lambda).$$

Moreover this limit is unique since $I^u(\cdot\,;\Lambda)$ is strictly monotone on $\operatorname{cl}(\mathcal{R}_o(\Lambda))$. It follows that $\nabla u(x;\Lambda) := y$ is a continuous extension of ∇u. ∎

We note that Cobb–Douglas utility functions satisfy (4.14), as do functions such that $\operatorname{cl}(A^j) \subset \operatorname{dom}(\partial u^j)$ since the assumption is void in this case.

Remark 4.5. Since $\tilde{A} = \sum_j \operatorname{dom}(\partial u^j_{A^j})$ and $\operatorname{dom}(\partial u^j_{A^j})$ is closed except possibly on some of the $\mathcal{C}^j_k \subset \operatorname{bdy}(\Re^n_{++})$, then $\tilde{A}$ has the same structure. In fact a face (in $\operatorname{bdy}(\Re^n_{++})$) of $\operatorname{cl}(\tilde{A})$ is contained in $\tilde{A}$ if and only if the corresponding face of each $\operatorname{cl}(\operatorname{dom}(\partial u^j_{A^j}))$ is contained in $\operatorname{dom}(\partial u^j_{A^j})$, i.e., for any set $\mathcal{I}_o \subset \{0,1,\dots,n\}$,

$$\{x \in A : x_i = 0,\ i \in \mathcal{I}_o\} \subset \tilde{A} \Leftrightarrow \{x \in A^j : x_i = 0,\ i \in \mathcal{I}_o\} \subset \operatorname{dom}(\partial u^j_{A^j}) \text{ for all } j.$$

Example 4.6. Let us consider a setting which arises in [3]. We take $J = 2$, and we make Assumptions 1, 2′, and 3, with $\mu^j = 0$ in (3.2), and the "Inada" condition, i.e., $\Re^n_{++} = \operatorname{dom}(\partial u^j) \subset \operatorname{dom}(u^j) \subset \Re^n_+$. (If u has the form $u(x^1 \oplus x^2) = g(x^1)$, cf. Example 3.1(iii), we require only that g satisfies this condition.)

We can conclude that

$$\lim_{x \to \operatorname{bdy}(\Re^n_{++})} \|\nabla u^j(x)\| = \infty. \tag{4.15}$$

We see this as follows: For $x_o \in \operatorname{bdy}(\Re^n_{++})$ but $x_o \notin \operatorname{dom}(u^j)$ and for $w \in \Re^n_+$ such that $x_o + w \in \operatorname{dom}(\partial u^j)$, ∇u^j monotone implies

$$\nabla u^j(x_o + \varepsilon w)^\top w(1-\varepsilon) \ge \int_\varepsilon^1 \nabla u^j(x_o + tw)^\top w\, dt = u^j(x_o + w) - u^j(x_o + \varepsilon w) \to +\infty$$

as $\varepsilon \to 0$. If $x_o \in \mathrm{bdy}(\Re^n_{++}) \cap \mathrm{dom}(u^j)$, then $\emptyset = \partial u^j(x_o) = U(x_o) - \mathcal{N}_{\mathrm{dom}(u^j)}(x_o)$, where $U(x_o)$ is the closure of the convex hull of all limits of $\nabla u^j(x_m)$ with $x_m \to x_o$; cf. Theorem 25.6 of [12]. Since $\mathcal{N}_{\mathrm{dom}(u^j)}(x_o) \neq \emptyset$ for $x_o \in \mathrm{dom}(u^j)$, then $U(x_o) = \emptyset$, so $\|\nabla u^j(x_m)\|$ must be unbounded.

We take $n = 3$, with $A^1 = ([0,\infty)^2 \times |0,1]) \cap \mathrm{dom}(u^1)$ and $A^2 = ([0,\infty)^2 \times \{0\}) \cap \mathrm{dom}(u^2)$. A and $\tilde{A}$ are easy to compute, but $A^u(\Lambda)$ is more complicated. Of interest are the pieces of A^j which lie in $\mathrm{dom}(\partial(u^j))$; these are $\mathcal{C}^1_o = \mathrm{int}(A^1)$, $\mathcal{C}^1_1 = \Re^2_{++} \times \{1\}$, and $\mathcal{C}^2_1 = \Re^2_{++} \times \{0\}$. We find $\mathcal{R}_o(\Lambda)$ which is composed of $\lambda_j \mathrm{int}(\mathcal{R}_{u^j_{A^j}})$, $j = 1, 2$, plus sets of the form $\lambda_1 \mathcal{S}^1_{k_1} \cap \lambda_2 \mathcal{S}^2_{k_2}$ with k_1, k_2 such that $\mathcal{N}_{A^1}(\mathcal{C}^1_{k_1})$ is orthogonal to $\mathcal{N}_{A^2}(\mathcal{C}^2_{k_2})$.

Since $\mathcal{M}^j = \{1, 2\}$, $\tilde{\mathcal{M}}^1 = \mathcal{M}^2_o = \{3\}$ and $\tilde{\mathcal{M}}^2 = \mathcal{M}^1_o = \emptyset$, then

$$\begin{aligned} \mathcal{S}^1_0 = \mathcal{R}_{u^1_{A^1}} &= \{y \in \Re^3_{++} : y_3 \geq u^1_{x_3}(x_1(y_1, y_2), x_2(y_1, y_2), 1)\}, \\ \mathcal{R}_{u^2_{A^2}} &= \Re^2_{++} \times \{0\},\ \mathcal{D}(\Lambda) = \Re^2_{++} \times \Re, \end{aligned}$$

where $x_1(y_1, y_2), x_2(y_1, y_2)$ satisfy $(u^1_{x_1}(x_1, x_2, 1), u^1_{x_2}(x_1, x_2, 1)) = (y_1, y_2)$, and

$$\mathcal{S}^1_1 = \{y \in \Re^3 : y_1 > 0,\ y_2 > 0,\ y_3 \leq u^1_{x_3}(x_1(y_1, y_2), x_2(y_1, y_2), 1)\}, \quad \mathcal{S}^2_1 = \Re^2_{++} \times \Re,$$

since $\mathcal{S}^j_k = (I^{u^j_{A^j}})^{-1}(\mathcal{C}^j_k)$. There are no other $\mathcal{S}^j_k$'s. We also have $\mathrm{aff}(\mathcal{N}_{A^1}(\mathcal{C}^1_1)) = \mathrm{aff}(\mathcal{N}_{A^2}(\mathcal{C}^2_1)) = \{(0,0)\} \times \Re$, and hence $\mathcal{R}_o(\Lambda) = \lambda_1 \mathrm{int}(\mathcal{R}_{u^1_{A^1}})$ and $A^u(\Lambda) = I^u(\mathcal{R}_o(\Lambda); \Lambda) = I^{u^1_{A^1}}(\mathrm{int}(\mathcal{R}_{u^1_{A^1}})) + I^{u^2_{A^2}}(\frac{\lambda_1}{\lambda_2} \mathrm{int}(\mathcal{R}_{u^1_{A^1}}))$,

$$\begin{aligned} A^u(\Lambda) &= (0,\infty)^2 \times (0,1) + (0,\infty)^2 \times \{0\} = (0,\infty)^2 \times (0,1), \\ \tilde{A} &= (0,\infty)^2 \times (0,1] + (0,\infty)^2 \times \{0\} = (0,\infty)^2 \times (0,1], \\ A &= [0,\infty)^2 \times [0,1] \cap \mathrm{dom}(u^1) \cap \mathrm{dom}(u^2). \end{aligned}$$

Note that $A^u(\Lambda)$ is all of $(0,\infty)^2 \times (0,1)$ because for x in this set we can solve

$$\begin{aligned} \lambda_1 u^1_{x_1}(\xi_1, \xi_2, x_3) - \lambda_2 u^2_{x_1}(x_1 - \xi_1, x_2 - \xi_2, 0) = 0, \\ \lambda_1 u^1_{x_2}(\xi_1, \xi_2, x_3) - \lambda_2 u^2_{x_2}(x_1 - \xi_1, x_2 - \xi_2, 0) = 0 \end{aligned}$$

for $(\xi_1, \xi_2) \in (0, x_1) \times (0, x_2)$. Then $x = I^u(y; \Lambda)$ for $y = \lambda_1 \nabla u^1(\xi_1, \xi_2, x_3)$.

For $x \in A^u(\Lambda)$, $\hat{x}^1 \in (0,\infty)^2 \times (0,1)$ and $\hat{x}^2 \in (0,\infty)^2 \times \{0\}$ since $\hat{x}^j \in \mathrm{dom}(\partial u^j_{A^j})$ always and $\hat{x}^1_3 = x_3$. Then the normals in (4.9) are $\vec{n}^1(\hat{x}^1) = 0$ and $\vec{n}^2(\hat{x}^2) = (0, 0, n_3)^\top$, and (4.9) becomes

$$\begin{aligned} u_{x_1}(x; \Lambda) &= \lambda_1 u^1_{x_1}(\hat{x}^1) = \lambda_2 u^2_{x_1}(\hat{x}^2), \\ u_{x_2}(x; \Lambda) &= \lambda_1 u^1_{x_2}(\hat{x}^1) = \lambda_2 u^2_{x_2}(\hat{x}^2), \\ u_{x_3}(x; \Lambda) &= \lambda_1 u^1_{x_3}(\hat{x}^1), \end{aligned} \tag{4.16}$$

where we have dropped the uninteresting equality $u_{x_3}(x; \Lambda) = -\vec{n}^2_3(\hat{x}^2)$. The extension given by Corollary 4.4 is obvious.

Example 4.7. Again take $J = 2$, $\mu^j = 0$, $\mathrm{dom}(u^j) = \Re^3_+$, and $\mathrm{dom}(\partial u^j) = \Re^3_{++}$, with $u^2(x) := u^2(x_1, x_2)$ (we are thinking of something like $u^1(x_1, x_2, x_3) = x_1^{\frac{1}{3}} x_2^{\frac{1}{3}} x_3^{\frac{1}{6}}$ and $u^2(x_1, x_2, x_3) = x_1^{\frac{1}{3}} x_2^{\frac{1}{3}}$). We take $A^1 = [0,\infty) \times [0,1] \times [0,\infty)$ and $A^2 = [0,\infty) \times [0,1] \times \{0\}$.

Then $\mathcal{M}^1 = \{0,2\}$, $\mathcal{M}^2 = \{0\}$, $\tilde{\mathcal{M}}^1 = \tilde{\mathcal{M}}^2 = \{1\}$, $\mathcal{M}_o^2 = \{2\}$, and $\mathcal{M}_o^1 = \emptyset$. The pieces of A^1 in $\mathrm{dom}(\partial u^1)$ are $\mathcal{C}_0^1 = \mathrm{int}(A^1)$ and $\mathcal{C}_1^1 = (0,\infty)\times\{1\}\times(0,\infty)$, a face, and for A^2 they are $\mathcal{C}_1^2 = (0,\infty)\times(0,1)\times\{0\}$, a face, and $\mathcal{C}_2^2 = (0,\infty)\times\{1\}\times\{0\}$, an edge. Moreover

$$\mathcal{S}_1^1 = \{(y_1,y_2,y_3) : y_1 = u^1_{x_1}(x_1,1,x_3),\ y_2 \le u^1_{x_2}(x_1,1,x_3),\ y_3 = u^1_{x_3}(x_1,1,x_3), \\ (x_1,x_3)\in\Re^2_{++}\},$$
$$\mathcal{S}_1^2 = \{(y_1,y_2) : y_1 = u^2_{x_1}(x_1,x_2,0),\ y_2 = u^2_{x_2}(x_1,x_2,0),\ (x_1,x_2)\in(0,\infty)\times(0,1)\} \\ \times\Re,$$
$$\mathcal{S}_2^2 = \{(y_1,y_2) : y_1 = u^2_{x_1}(x_1,1,0),\ y_2 \le u^2_{x_2}(x_1,1,0),\ x_1\in(0,\infty)\}\times\Re$$

and $\mathrm{aff}(\mathcal{N}_{A^1}(\mathcal{C}_1^1)) = \{0\}\times\Re\times\{0\}$, $\mathcal{N}_{A^2}(\mathcal{C}_1^2) = \{(0,0)\}\times\Re$, and $\mathrm{aff}(\mathcal{N}_{A^2}(\mathcal{C}_2^2)) = \{0\}\times\Re^2$. Observe that

$$\mathcal{R}_{u^1_{A^1}} = \{(y_1,y_2,y_3) : y_1 = u^1_{x_1}(x_1,1,x_3),\ y_2 \ge u^1_{x_2}(x_1,1,x_3),\ y_3 = u^1_{x_3}(x_1,1,x_3), \\ (x_1,x_3)\in\Re^2_{++}\},$$
$$\mathcal{R}_{u^2_{A^2}} = \{(y_1,y_2) : y_1 = u^2_{x_1}(x_1,x_2,0),\ y_2 = u^2_{x_2}(x_1,x_2,0),\ (x_1,x_2)\in(0,\infty)\times(0,1]\} \\ \times\{0\},$$

and $\mathcal{D}(\Lambda) = \Re_{++}\times\Re\times\Re_{++}$. Thus

$$\mathcal{R}_o(\Lambda) = \lambda_1\mathrm{int}(\mathcal{R}_{u^1_{A^1}}) \cup [\lambda_1\mathcal{S}_1^1 \cap \lambda_2\mathcal{S}_1^2].$$

If $y\in\lambda_1\mathrm{int}(\mathcal{R}_{u^1_{A^1}})$, then $I^{u^1}(\frac{y}{\lambda_1})\in\mathrm{int}(A^1)$, so $I^u(y)\in\mathrm{int}(A^1)+A^2 = (0,\infty)\times(0,2)\times(0,\infty)$. If $y\in\lambda_1\mathcal{S}_1^1\cap\lambda_2\mathcal{S}_1^2$, then $I^{u^1}(\frac{y}{\lambda_1})\in\mathcal{C}_1^1$ and $I^{u^2}(\frac{y}{\lambda_2})\in\mathcal{C}_1^2$, and hence $I^u(y)\in\mathcal{C}_1^1+\mathcal{C}_1^2 = (0,\infty)\times(1,2)\times(0,\infty)$.

Let us show that $A^u(\Lambda) = (0,\infty)\times(0,2)\times(0,\infty)$. For $x\in(0,\infty)\times(0,1]\times(0,\infty)$ we can find $\hat{x}^j$ by solving $\nabla_\xi[\lambda_1u^1(\xi_1,\xi_2,x_3) + \lambda_2u^2(x_1-\xi_1,x_2-\xi_2,0)] = 0$, i.e., $\hat{x}^1 = (\xi_1,\xi_2,x_3)^\top$, $\hat{x}^2 = x-\hat{x}^1$, where

$$\begin{aligned}\lambda_1u^1_{x_1}(\xi_1,\xi_2,x_3) - \lambda_2u^2_{x_1}(x_1-\xi_1,x_2-\xi_2,0) &= 0,\\ \lambda_1u^1_{x_2}(\xi_1,\xi_2,x_3) - \lambda_2u^2_{x_2}(x_1-\xi_1,x_2-\xi_2,0) &= 0,\end{aligned} \tag{4.17}$$

and $(\xi_1,\xi_2)\in(0,x_1)\times(0,x_2)$. For each ξ_2 the intermediate value and implicit function theorems give a continuous function $\xi_1(\xi_2)$ satisfying the first equation. Similarly from the second we obtain $\xi_2(\xi_1)$. Their intersection in the rectangle gives the solution. Then $\hat{x}^1 = (\xi_1,\xi_2,x_3)$ and $x = I^u(y;\Lambda)$ with $y = \lambda_1\nabla u^1(\xi_1,\xi_2,x_3) = \lambda_2\nabla u^2(x_1-\xi_1,x_2-\xi_2,0)$.

For $x\in(0,\infty)\times(1,2]\times(0,\infty)$ the same argument gives a solution $(\xi_1,\xi_2)\in(0,x_1)\times[x_2-1,1]$ if we work with supergradients; i.e., we replace $u^j_{x_2}(\xi_1,1,x_3)$ whenever it occurs by the interval (multivalued function) $(-\infty,u^j_{x_2}(\xi_1,1,x_3))$ (due to the constraint $x_2=1$; cf. the normal in (4.9)). So again $x = I^u(y;\Lambda)$ with $y = \lambda_1\nabla u^1(\xi_1,\xi_2,x_3)$, and I^u maps onto $(0,\infty)\times(0,2)\times(0,\infty)$. Then

$$A = [0,\infty)\times[0,2]\times[0,\infty),\quad \tilde{A} = (0,\infty)\times(0,2]\times(0,\infty),\quad A^u(\Lambda) = \mathrm{int}(A).$$

Observe that $\vec{n}^1(\hat{x}^1) = (0,n_2^1,0)$ where $n_2^1 = 0$ unless $\hat{x}^1\in\mathcal{C}_3^1$; otherwise $\vec{n}^1 = 0$. Moreover $\vec{n}^2(\hat{x}^2) = (0,n_2^2,n_3^2)$ with $n_2^2 = 0$ unless $\hat{x}^2\in\mathcal{C}_2^2$. From (4.17) we see that

$n_2^1 > 0$ if and only if $\lambda_1 u^1_{x_2}(\xi_1, 1, x_3) > \lambda_2 u^2_{x_2}(x_1 - \xi_1, x_2 - 1, 0)$ and $n_2^2 > 0$ if and only if $\lambda_1 u^1_{x_2}(\xi_1, x_2 - 1, x_3) < \lambda_2 u^2_{x_2}(x_1 - \xi_1, 1, 0)$. Here ξ_1 is the solution of (4.17), with $\xi_2 = 1$ in the first case and $\xi_2 = x_2 - 1$ in the second. Note that $n_2^j = 0$ if $x_2 \le 1$.

Theorem 4.3(iv) gives

$$\begin{aligned} u_{x_1}(x;\Lambda) &= \lambda_1 u^1_{x_1}(\hat{x}^1) \quad = \quad \lambda_2 u^2_{x_1}(\hat{x}^2), \\ u_{x_2}(x;\Lambda) &= \lambda_1 u^1_{x_2}(\hat{x}^1) - n_2^1 \quad = \quad \lambda_2 u^2_{x_2}(\hat{x}^2) - n_2^2, \\ u_{x_3}(x;\Lambda) &= \lambda_1 u^1_{x_3}(\hat{x}^1), \end{aligned}$$

where we have dropped the uninformative equality $u_{x_3}(x;\Lambda) = n_3^2 \in \Re$.

Example 4.8. In this example we do not assume the Inada condition at 0, so take $J = 2$ and assume that u^j satisfy Assumptions 1, 2′, and 3 with $\mu^j = 0$ and $A^1 = [0,\infty) \times [0,1]$, $A^2 = [0,\infty) \times \{0\}$. We assume that $\mathrm{dom}(\partial u^j)$ contains $[0,\infty)^2$. Then $n = 2$, $\mathcal{M}^j = \{1\}$, $\tilde{\mathcal{M}}^1 = \mathcal{M}_o^2 = \{2\}$, and $\tilde{\mathcal{M}}^2 = \mathcal{M}_o^1 = \emptyset$.

The sets $\mathcal{C}_k^1$, $k = 1,\dots,5$, are $(0,\infty)\times\{0\}, \{(0,0)\}, \{0\}\times(0,1), \{(0,1)\}, (0,\infty)\times\{1\}$ in order, and $\mathcal{C}_1^2 = (0,\infty)\times\{0\}$ and $\mathcal{C}_2^2 = \{(0,0)\}$. Moreover $\mathcal{R}_{u^1_{A^1}}$ is the subset of $\Re^2_{++}$ with boundaries given by the three curves $\{y = \nabla u^1(\xi,0) : \xi \ge 0\}$, $\{y = \nabla u^1(0,\xi) : \xi \in (0,1)\}$, $\{y = \nabla u^1(\xi,1) : \xi \ge 0\}$, $\mathcal{R}_{u^2_{A^2}} = \{\nabla u^2(\xi,0) : \xi \ge 0\}$ and $\mathcal{D}(\Lambda) = \Re^1_{++} \times \Re$. Figure 1 of [3] shows $\bigcup_k \mathcal{S}_k^1 \cap \Re^2_{++}$.

$$\begin{aligned} \mathcal{S}_1^1 &= \{y \in \Re^2 : y_1 = u^1_{x_1}(\xi,0),\ y_2 \ge u^1_{x_2}(\xi,0),\ \xi > 0\}, \\ \mathcal{S}_2^1 &= \{y \in \Re^2 : y \ge \nabla u^1(0,0)\}, \\ \mathcal{S}_3^1 &= \{y \in \Re^2 : y_1 \ge u^1_{x_1}(0,\xi),\ y_2 = u^1_{x_2}(0,\xi),\ 1 > \xi > 0\}, \\ \mathcal{S}_4^1 &= \{y \in \Re^2 : y_1 \ge u^1_{x_1}(0,1),\ y_2 \le u^1_{x_2}(0,1)\}, \\ \mathcal{S}_5^1 &= \{y \in \Re^2 : y_1 = u^1_{x_1}(\xi,1),\ y_2 \le u^1_{x_2}(\xi,1),\ \xi > 0\}, \\ \mathcal{S}_1^2 &= \{y \in \Re^2 : 0 < y_1 < u^2_{x_1}(0,0)\}, \\ \mathcal{S}_2^2 &= \{y \in \Re^2 : y_1 \ge u^2_{x_1}(0,0)\}. \end{aligned}$$

Now $\mathrm{aff}(\mathcal{N}_{A^1}(\mathcal{C}_2^1)) = \mathrm{aff}(\mathcal{N}_{A^1}(\mathcal{C}_4^1)) = \mathrm{aff}(\mathcal{N}_{A^2}(\mathcal{C}_2^2)) = \Re^2$, $\mathrm{aff}(\mathcal{N}_{A^1}(\mathcal{C}_1^1)) = \mathrm{aff}(\mathcal{N}_{A^1}(\mathcal{C}_5^1))$ $= \mathrm{aff}(\mathcal{N}_{A^2}(\mathcal{C}_1^2)) = \{0\}\times\Re$, and $\mathrm{aff}(\mathcal{N}_{A^1}(\mathcal{C}_3^1)) = \Re\times\{0\}$. Thus

$$\mathcal{R}_o(\Lambda) = \lambda_1 \mathrm{int}(\mathcal{R}_{u^1_{A^1}}) \cup [\lambda_1\mathcal{S}_3^1 \cap \lambda_2\mathcal{S}_1^2].$$

If $y \in \lambda_1 \mathrm{int}(\mathcal{R}_{u^1_{A^1}})$, then $I^{u^1}(\frac{y}{\lambda_1}) \in \mathrm{int}(A^1)$, so $I^u(y) \in \mathrm{int}(A^1) + A^2 = (0,\infty)\times(0,1)$. If $y \in \lambda_1\mathcal{S}_3^1 \cap \lambda_2\mathcal{S}_1^2$, then $I^{u^1}(\frac{y}{\lambda_1}) \in \mathcal{C}_3^1$ and $I^{u^2}(\frac{y}{\lambda_2}) \in \mathcal{C}_1^2$. Hence $I^u(y) \in \mathcal{C}_3^1 + \mathcal{C}_1^2 = (0,\infty)\times(0,1)$, and in fact

$$A^u(\Lambda) = (0,\infty)\times(0,1), \qquad \tilde{A} = A = [0,\infty)\times[0,1].$$

For $x = (x_1,x_2) \in A^u(\Lambda)$, $\hat{x}_2^1 = x_2 \in (0,1)$, so $\vec{n}_2^1(\hat{x}^1) = 0$. Hence

$$\begin{aligned} u_{x_1}(x;\Lambda) &= \lambda_1 u^1_{x_1}(\hat{x}^1) - \vec{n}_1^1(\hat{x}^1) = \lambda_2 u^2_{x_1}(\hat{x}^2) - \vec{n}_1^2(\hat{x}^2), \\ u_{x_2}(x;\Lambda) &= \lambda_1 u^1_{x_2}(\hat{x}^1), \end{aligned}$$

where $\vec{n}_1^j(\hat{x}^j) \le 0$ and is nonzero only if $\hat{x}_1^j = 0$. We have dropped the uninformative equality involving the arbitrary $\vec{n}_2^2(\hat{x}^2)$. Since $\hat{x}_1^1 + \hat{x}_1^2 = x_1 > 0$, then at least one of

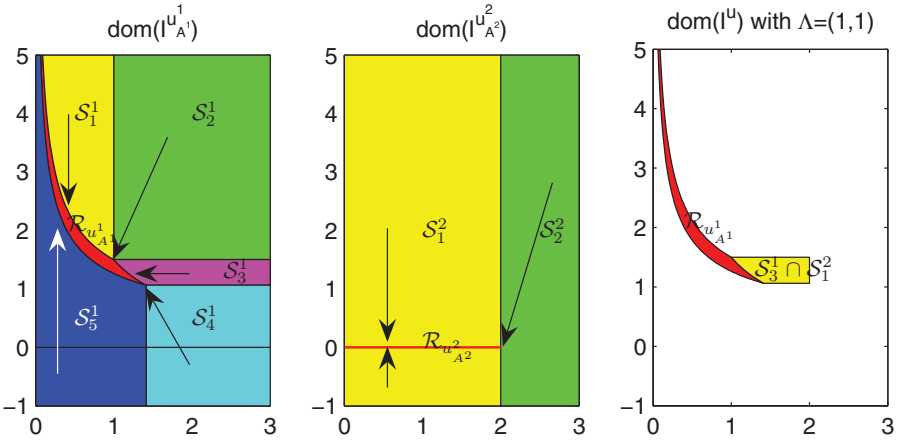


FIG. 2. *The domains of* $I^{u^1_{A^1}}$, $I^{u^2_{A^2}}$, I^u.

$\hat{x}_1^j > 0$. In fact x determines when $\vec{n}_1^j(\hat{x}^j) < 0$ as follows: $\vec{n}_1^1(\hat{x}^1) < 0$ if and only if $\lambda_1 u^1_{x_1}(0, x_2) < \lambda_2 u^2_{x_1}(x_1, 0)$, and $\vec{n}_1^2(\hat{x}^2) < 0$ if and only if $\lambda_1 u^1_{x_1}(x) > \lambda_2 u^2_{x_1}(0, 0)$ (look at $F'(0)$, $F'(x_1)$, where $F(\xi) := \lambda_1 u^1(\xi, x_2) + \lambda_2 u^2(x_1 - \xi, 0)$; cf. the definition of $u(\cdot\,; \Lambda)$).

We have computed some of the sets in the y-space for $u^1(x) := 3(x_1 + 1)^{1/3}\,(x_2 + 1)^{1/2}$ and $u^2(x) = 2(x_1 + \frac{1}{4})^{1/2}$ and plotted them in Figure 2. The arrows correspond to the projections; i.e., they are outward normals to $\mathcal{C}_k^j$. The third panel exhibits $\mathcal{R}_o$ as part of $\mathrm{dom}(I^u)$ when $\Lambda = (1, 1)$.

Remark 4.9. If we replace u^2 in Example 4.8 by $u^2(x_1, x_2) := \mu_1 x_1 - x_1^{-1}$, A^2 by $(0, \infty) \times \{0\}$, then $\mathrm{dom}(I^{u^2_{A^2}}) = \{y \in \Re^2 : y_1 > \mu_1\} = \mathcal{S}_1^2$ and $\mathcal{R}_{u^2_{A^2}} = (\mu_1, \infty) \times \{0\}$. There are no other $\mathcal{S}_k^2$. With $\Lambda = (1, 1)$ again, $\mathrm{dom}(I^u(\cdot\,, \Lambda)) = \{y \in \Re^2 : y_1 > \mu_1\}$ and $\mathcal{R}_o(\Lambda)$ is the intersection of $\mathrm{dom}(I^u(\cdot\,, \Lambda))$ with $\mathcal{R}_o(\Lambda)$ of the third panel of Figure 2, after the lightly shaded part has been extended infinitely to the right. Moreover $A = (0, \infty) \times [0, 1]$, $\tilde{A} = (0, \infty) \times (0, 1]$, and $A^u = (0, \infty) \times (0, 1)$.

REFERENCES

[1] A.B. ABEL AND J.C. EBERLY, *An exact solution for the investment and value of a firm facing uncertainty, adjustment costs, and irreversibility*, J. Econom. Dynam. Control, 21 (1997), pp. 831–852.

[2] P. BANK AND F. RIEDEL, *Optimal Dynamic Choice of Durable and Perishable Goods*, Discussion paper 03-009, Department of Economics, Stanford University, Palo Alto, CA, 2003.

[3] M.B. CHIAROLLA AND U.G. HAUSSMANN, *Equilibrium in a stochastic model with consumption, wages and investment*, J. Math. Econom., 35 (2001), pp. 1–31. A version without typos can be found at http://www.math.ubc.ca/~uhaus/wage.pdf.

[4] M.B. CHIAROLLA AND U.G. HAUSSMANN, *Equilibrium in an economy with endogenous production and consumption*, preprint, 2007.

[5] M.B. CHIAROLLA AND U.G. HAUSSMANN, *A stochastic equilibrium economy with irreversible investment*, to appear, 2008.

[6] F.H. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley & Sons, New York, 1983.

[7] R.-A. DANA AND M. PONTIER, *On existence of an Arrow-Radner equilibrium in the case of complete markets. A remark*, Math. Oper. Res., 17 (1992), pp. 148–163.

[8] G. DEELSTRA, H. PHAM, AND N. TOUZI, *Dual formulation of the utility maximization problem under transaction costs*, Ann. Appl. Probab., 11 (2001), pp. 1353–1383.

[9] I. Karatzas, J.P. Lehoczky, and S.E. Shreve, *Existence and uniqueness of multi-agent equilibrium in a stochastic, dynamic consumption/investment model*, Math. Oper. Res., 15 (1990), pp. 80–128.
[10] Q. Meng and C.K. Yip, *Investment, interest rate rules and equilibrium determinacy*, Econom. Theory, 23 (2004), pp. 863–878.
[11] A. Nagurney, *Variational inequalities in the analysis and computation of multi-sector, multi-instrument financial equilibria*, J. Econom. Dynam. Control, 18 (1994), pp. 161–184.
[12] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

SIAM J. OPTIM.
Vol. 19, No. 4, pp. 1534–1558

SPARSE SOS RELAXATIONS FOR MINIMIZING FUNCTIONS THAT ARE SUMMATIONS OF SMALL POLYNOMIALS*

JIAWANG NIE† AND JAMES DEMMEL‡

Abstract. This paper discusses how to find the global minimum of functions that are summations of small polynomials ("small" means involving a small number of variables). Some sparse sum of squares (SOS) techniques are proposed. We compare their computational complexity and lower bounds with prior SOS relaxations. Under certain conditions, we also discuss how to extract the global minimizers from these sparse relaxations. The proposed methods are especially useful in solving sparse polynomial system and nonlinear least squares problems. Numerical experiments are presented which show that the proposed methods significantly improve the computational performance of prior methods for solving these problems. Lastly, we present applications of this sparsity technique in solving polynomial systems derived from nonlinear differential equations and sensor network localization.

Key words. polynomials, sum of squares (SOS), sparsity, nonlinear least squares, polynomial system, nonlinear differential equations, sensor network localization

AMS subject classifications. 65H10, 65K10, 65N22, 90C22, 90C26, 90C59

DOI. 10.1137/060668791

1. Introduction.

Global optimization of multivariate polynomial functions contains quite a broad class of optimization problems. It has wide and important applications in science and engineering. Recently, there has been much work on globally minimizing polynomial functions using representation theorems from real algebraic geometry for positive polynomials. The basic idea is to approximate nonnegative polynomials by sum of squares (SOS) polynomials. This approximation is also called *SOS relaxation*, since not every nonnegative polynomial is SOS. Here a polynomial is said to be SOS if it can be written as a sum of squares of other polynomials. The advantage of SOS polynomials is that a polynomial is SOS if and only if a certain semidefinite program (SDP) formed by its coefficients is feasible. Since SDP [29] has efficient numerical methods, we can check whether a polynomial is SOS by solving a particular SDP.

To be more specific, suppose we wish to find the global minimum value f^* of a polynomial function $f(x)$ of vector $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$. The SOS relaxation finds a lower bound γ for f^* such that the polynomial $f(x) - \gamma$ is SOS. Obviously, $f(x) - \gamma$ being SOS implies that $f(x) - \gamma$ is nonnegative for every real vector x. Hence such a γ is a lower bound. The maximum γ found this way is called the SOS lower bound, which is often denoted by f^*_{sos}. The relation $f^*_{sos} \le f^*$ always holds for every polynomial $f(x)$ (it is possible that $f^*_{sos} = -\infty$). When $f^*_{sos} = f^*$, we say the SOS relaxation is *exact*. We refer to [13, 22, 23] for more details on SOS relaxations for polynomial optimization problems. There are two important issues for applying SOS relaxation in global optimization of polynomial functions: the *quality* and *computational complexity*.

*Received by the editors August 31, 2006; accepted for publication (in revised form) July 28, 2008; published electronically December 31, 2008.
http://www.siam.org/journals/siopt/19-4/66879.html
†Department of Mathematics, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093 (njw@math.ucsd.edu).
‡Department of Mathematics and EECS, University of California, Berkeley, CA 94720 (demmel@cs.berkeley.edu).

The quality means how good is the SOS lower bound f^*_{sos}. In practice, as observed in [22], many nonnegative polynomials that are not "artificially" constructed are SOS. However, Blekherman [4] pointed out that there are much more nonnegative polynomials than SOS polynomials. But usually SOS relaxation provides very good approximations, although theoretically it can fail with high probability. When SOS relaxation is not exact, i.e., $f^*_{sos} < f^*$, there are methods to fix it by applying modified SOS relaxations. Nie, Demmel, and Sturmfels [20] proposed to use SOS representations of $f(x)-\gamma$ modulo the gradient ideal of $f(x)$, and show that the minimum value f^* can be obtained when f^* is attained at some point. Schweighofer [26] proposes to minimize $f(x)$ over a semialgebraic set called the *gradient tentacle*, and shows that the minimum value f^* can be computed when $f^* > -\infty$ but not attainable. Jibetean and Laurent [10] and Lasserre [14] propose perturbing $f(x)$ by adding a higher degree polynomial with tiny coefficients, and showed that the lower bounds will converge to the minimum value f^*. Recently, Laurent [16] gave a survey on solving polynomial optimization by using semidefinite relaxations. We refer to [10, 14, 16, 20, 26] for related work.

Another important issue for SOS relaxation is the computational complexity. Suppose $f(x)$ has degree $2d$ (it must be even for $f(x)$ to have a finite minimum). Then $f(x)$ has up to $\binom{n+2d}{2d}$ monomials. The condition that $f(x) - \gamma$ being SOS reduces to an SDP the size of whose linear matrix inequality (LMI) is $\binom{n+d}{d}$ with $\binom{n+2d}{2d}$ variables. These numbers can be huge for moderate n and d, say, $n = 2d = 10$. For large scale polynomial optimization problems, the general SOS relaxation is very difficult to implement numerically. Sometimes this complexity makes the applicability of SOS relaxation very limited. We refer to [22, 23] for the connection between SOS relaxation and SDP.

Prior work. There is some work on exploiting sparsity in polynomial optimization when the polynomials are *sparse*. In such situations, sparse SOS relaxations are available and the resulting SDPs have reduced sizes, and hence larger problems can be solved. Here being sparse means that the number of monomials with nonzero coefficients is much smaller than the maximum possible number $\binom{n+2d}{2d}$. Kojima et al. [12] and Parrilo [24] discussed how to exploit sparsity of SOS relaxations in unconstrained polynomial optimization. Kim et al. [11] and Lasserre [15] discussed sparse SOS relaxations for constrained polynomial optimization problems and showed convergence under certain conditions. Waki et al. [28] proposed a heuristic procedure to exploit sparsity for minimizing polynomials by *chordal extension* of the *correlation sparsity pattern graph* (csp graph): the vertices of the csp graph are the variables $x_1, \ldots, x_n$; the edge (x_i, x_j) exists whenever $x_i x_j$ appears in one monomial of $f(x)$. To find one chordal extension, [28] proposed to use the symbolic sparse Cholesky factorization of the csp matrix with minimum degree ordering. If the chordal extension of the csp graph is also sparse, then the sparsity technique in [28] works well. However, if the chordal extension of the csp graph is much less sparse, then that sparsity technique might still be too expensive to be implementable for some practical problems.

Contributions. In many practical applications, the polynomials are not only sparse, but also given with certain sparsity patterns. For instance, the polynomials are often summations of other "small" polynomials, i.e., polynomials involving only a small number of variables. Sometimes, these representations contain useful information that might help us save computations significantly. These sparsity patterns are often ignored in prior work, where these polynomials would be treated using the usual "dense" algorithms. The main contribution of this paper is to propose new

sparse SOS relaxation techniques taking the given sparsity pattern into account, and to show numerical experiments demonstrating their accuracy and speed.

In this paper, we consider the polynomial optimization problem of the form

$$\min_{x\in\mathbb{R}^n} \quad f(x) = \sum_{i=1}^{m} f_i(x_{\Delta_i}), \tag{1.1}$$

where $\Delta_i \subset [n] = \{1, 2, \ldots, n\}$. Here each $f_i(x_{\Delta_i})$ is a polynomial in $x_{\Delta_i} = (x_j | j \in \Delta_i)$. Let $\deg(f_i) = 2d_i$ and $2d = \deg(f) = \max\{2d_1, \ldots, 2d_m\}$ (we assume each f_i has even degree along with f). One basic and natural idea for solving problem (1.1) is to find the maximum γ such that

$$f(x) - \gamma = \sum_{i=1}^{m} s_i(x_{\Delta_i}),$$

where each $s_i(x_{\Delta_i})$ is an SOS polynomial in x_{Δ_i} instead of all the variables $x_1, \ldots, x_n$. Exploiting this sparsity pattern can save significant computation without sacrificing much solution quality for many practical problems. In addition to presenting its numerical implementation, this paper will also discuss the theoretical properties of this sparse SOS relaxation and its variations.

The main distinction of our sparse SOS technique from earlier work like Waki et al. [28] is that we do not use the chordal extension of csp graphs. In the case that the csp graph of $f(x)$ in (1.1) is chordal, our sparsity technique is almost the same as the one in [28]. However, if the csp graph of $f(x)$ is not chordal and its chordal extension is much less sparse, then our sparsity technique is significantly more efficient. If the csp graph of $f(x)$ is not chordal and its chordal extension is also sparse, then our sparsity technique is slightly more efficient while not losing much quality of solution. Furthermore, our sparsity technique can be applied to solve bigger dense polynomial optimization problems which cannot be solved by other existing methods. This is due to the observation that every polynomial $g(x)$ is a summation of monomials whose number of variables is at most the degree $\deg(g)$. So, when $\deg(g)$ is small, like 4 or 6, then the formulation (1.1) is a good sparse model. The numerical computations show that our sparsity technique is usually more efficient than other existing methods in solving problems of the form (1.1).

We remark that for a given polynomial $f(x)$, there is a flexibility to choose the summands $f_i(x_{\Delta_i})$ in (1.1). Sometimes, this flexibility is very useful, since it allows us to choose among various sparse relaxations and select the most efficient one from them. The best choice of $f_i(x_{\Delta_i})$ is usually problem dependent and there is no general rule. However, for practical problems like solving polynomial systems or nonlinear least squares, there are natural choices for $f_i(x_{\Delta_i})$. This is illustrated in section 4 and section 5.

Polynomial optimization problems of the form (1.1) have important practical applications: (i) *Solving polynomial systems:* Many large scale polynomial equations are often sparse, and each equation might involve just a few variables, e.g., the polynomial equations obtained from discretization in nonlinear differential equations. Such polynomial systems can be equivalently transformed to global polynomial optimization problems of the form (1.1). We will show that the proposed sparse SOS relaxation is exact when the polynomial system has at least one real solution. (ii) *Nonlinear least squares:* Many difficult problems in statistics, biology, engineering, or other applications require solving certain nonlinear least squares problems and finding their

global optimal solutions. If each equation is sparse, then sparse polynomial optimization (1.1) is a very natural model and our sparsity technique is very suitable. Sensor network localization is one important application of this kind.

Outline. This paper is organized as follows. Section 2 introduces some notation and background for SOS relaxations, section 3 discusses properties of the sparse SOS relaxation and its variations, section 4 presents some numerical implementations, and section 5 shows applications. Lastly, section 6 draws some conclusions and discusses future work in this area.

2. Preliminaries. This section introduces some notations and backgrounds in SOS relaxation methods for minimizing polynomial functions.

Throughout this paper, we will use the following notation: $\mathbb{R}$ is the field of real numbers; $\mathbb{N}$ is the set of nonnegative integers; $\mathbb{R}^{\Delta_i} = \{(x_{k_1}, \ldots, x_{k_\ell}) : x_{k_j} \in \mathbb{R}\}$ when $\Delta_i = \{k_1, k_2, \ldots, k_\ell\}$; $\mathbb{R}[X]$: the ring of real polynomials in $X = (x_1, x_2, \ldots, x_n)$; $\mathbb{R}[X_{\Delta_i}]$: the ring of real polynomials in $X_{\Delta_i} = (x_k)_{k\in\Delta_i}$; $\sum\mathbb{R}[X]^2$: SOS polynomials in $\mathbb{R}[X]$; $\sum\mathbb{R}[X_{\Delta_i}]^2$: SOS polynomials in $\mathbb{R}[X_{\Delta_i}]$; $\sum\mathbb{R}_N[X]^2$: SOS polynomials in $\mathbb{R}[X]$ with degree at most $2N$; $\sum\mathbb{R}_N[X_{\Delta_i}]^2$: SOS polynomials in $\mathbb{R}[X_{\Delta_i}]$ with degree at most $2N$; $\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$; $x^\alpha = x_1^{\alpha_1}x_2^{\alpha_2}\ldots x_n^{\alpha_n}$ for $\alpha \in \mathbb{N}^n$; $\mathrm{supp}(\alpha) = \{i \in [n] : \alpha_i \neq 0\}$; $\mathrm{supp}(f) = \{\alpha \in \mathbb{N}^n :$ the coefficient of x^α in $f(x)$ is nonzero$\}$; $|F|$ denotes the cardinality of set F; A^T denotes the transpose of matrix A; $A \succeq (\succ) 0$ means matrix A is positive semidefinite (definite); $\mathcal{M}_d(y)$ is the moment matrix of order d about $x \in \mathbb{R}^n$; $\mathcal{M}_d^{\Delta_i}(y)$ is the moment matrix of order d about $x_{\Delta_i} \in \mathbb{R}^{\Delta_i}$; $\mathcal{M}_F(y)$ is the moment matrix generated monomials with support F.

2.1. SOS and semidefinite programming (SDP). A polynomial $p(x)$ in $x = (x_1, \ldots, x_n)$ is said to be sum of squares (SOS) if $p(x) = \sum_i p_i^2(x)$ for some polynomials $p_i(x)$. Obviously, if $p(x)$ is SOS, then $p(x)$ is *nonnegative*; i.e., $p(x) \geq 0$ for all $x \in \mathbb{R}^n$. However, the converse is not true. If $p(x)$ is nonnegative, then $p(x)$ is not necessarily SOS. In other words, the set of SOS polynomials (which forms a cone) is properly contained in the set of nonnegative polynomials (which forms a larger cone). The process of approximating nonnegative polynomials by SOS polynomials is called SOS relaxation. For instance, the polynomial

$$\begin{aligned} &x_1^4 + x_2^4 + x_3^4 + x_4^4 - 4x_1x_2x_3x_4 \\ &\quad = \tfrac{1}{3}\Big\{\big(x_1^2 - x_2^2 - x_4^2 + x_3^2\big)^2 + \big(x_1^2 + x_2^2 - x_4^2 - x_3^2\big)^2 + \big(x_1^2 - x_2^2 - x_3^2 + x_4^2\big)^2 \\ &\qquad + 2(x_1x_4 - x_2x_3)^2 + 2(x_1x_2 - x_3x_4)^2 + 2(x_1x_3 - x_2x_4)^2\Big\} \end{aligned}$$

is SOS. This identity immediately implies that

$$x_1^4 + x_2^4 + x_3^4 + x_4^4 - 4x_1x_2x_3x_4 \geq 0, \ \ \forall (x_1, x_2, x_3, x_4) \in \mathbb{R}^4,$$

which is one arithmetic-geometric mean inequality.

The advantage of SOS polynomials over nonnegative polynomials is that it is more tractable to check whether a polynomial is SOS. To test whether a polynomial is SOS is equivalent to testing the feasibility of some SDP [22, 23], which has efficient numerical solvers. To illustrate this, suppose polynomial $p(x)$ has degree $2d$ (SOS polynomials must have even degree). Then $p(x)$ is SOS if and only if [22, 23] there exists a symmetric matrix $W \succeq 0$ such that

$$p(x) = \mathbf{m}_d(x)^T\, W\, \mathbf{m}_d(x),$$

where $\mathbf{m}_d(x)$ is the column vector of monomials up to degree d. For instance,

$$\mathbf{m}_2(x_1, x_2) = \left[1,\, x_1,\, x_2,\, x_1^2,\, x_1x_2,\, x_2^2\right]^T.$$

As is well known, the number of monomials in x up to degree d is $\binom{n+d}{d}$. Thus the size of matrix W is $\binom{n+d}{d}$. This number can be very large. For instance, when $n = d = 10$, $\binom{n+d}{d} \geq 10^5$. However, for fixed d (e.g., $d = 2$), $\binom{n+d}{d}$ is polynomial in n. On the other hand, it is NP-hard (with respect to n) to tell whether a polynomial is nonnegative whenever $2d \geq 4$ (even when d is fixed) [13].

2.2. SOS relaxation in polynomial optimization. Let $f(x) = \sum_\alpha f_\alpha x^\alpha$ be a polynomial in x. Consider the global optimization problem

$$f^* := \min_{x \in \mathbb{R}^n} \quad f(x).$$

This problem is NP-hard when $\deg(f) \geq 4$. The standard SOS relaxation is

$$\begin{aligned} f^*_{sos} := \max \quad & \gamma \\ s.t. \quad & f(x) - \gamma \text{ is SOS.} \end{aligned}$$

Obviously we have that $f^*_{sos} \leq f^*$. In practice, SOS provides very good approximations, and often gives exact global minimum, i.e., $f^*_{sos} = f^*$, even though theoretically there are many more nonnegative polynomials than SOS polynomials [4].

In terms of SDP, the SOS relaxation can also be written as

$$f^*_{sos} := \max \quad \gamma \tag{2.1}$$

$$s.t. \quad f(x) - \gamma = \mathbf{m}_d(x)^T W \mathbf{m}_d(x) \tag{2.2}$$

$$W \succeq 0 \tag{2.3}$$

where $2d = \deg(f)$. The decision variable in the above is (γ, W) instead of x. The above program is convex about (γ, W). A lower bound f^*_{sos} can be computed by solving the resulting SDP. It can be shown [13] that the dual of (2.1)–(2.3) is

$$f^*_{mom} := \min_y \quad \sum_{|\alpha| \leq 2d} f_\alpha y_\alpha \tag{2.4}$$

$$s.t. \quad \mathcal{M}_d(y) \succeq 0 \tag{2.5}$$

$$y_{0,\ldots,0} = 1. \tag{2.6}$$

Here $\mathcal{M}_d(y)$ is the *moment matrix* generated by $y = (y_\alpha)$, a vector indexed by monomials of degree at most $2d$. The rows and columns of moment matrix $\mathcal{M}_d(y)$ are indexed by integer vectors. Each entry of $\mathcal{M}_d(y)$ is defined as

$$\mathcal{M}_d(y)(\alpha, \beta) := y_{\alpha+\beta}, \ \forall |\alpha|, |\beta| \leq d.$$

For instance, when $d = 2$ and $n = 2$, the vector

$$y = [\, y_{0,0},\, y_{1,0},\, y_{0,1},\, y_{2,0},\, y_{1,1},\, y_{0,2},\, y_{3,0},\, y_{2,1},\, y_{1,2},\, y_{0,3},\, y_{4,0},\, y_{3,1},\, y_{2,2},\, y_{1,3},\, y_{0,4}\,]$$

defines moment matrix

$$\mathcal{M}_2(y) = \begin{bmatrix} y_{0,0} & y_{1,0} & y_{0,1} & y_{2,0} & y_{1,1} & y_{0,2} \\ y_{1,0} & y_{2,0} & y_{1,1} & y_{3,0} & y_{2,1} & y_{1,2} \\ y_{0,1} & y_{1,1} & y_{0,2} & y_{2,1} & y_{1,2} & y_{0,3} \\ y_{2,0} & y_{3,0} & y_{2,1} & y_{4,0} & y_{3,1} & y_{2,2} \\ y_{1,1} & y_{2,1} & y_{1,2} & y_{3,1} & y_{2,2} & y_{1,3} \\ y_{0,2} & y_{1,2} & y_{0,3} & y_{2,2} & y_{1,3} & y_{0,4} \end{bmatrix}.$$

For SOS relaxation (2.1)–(2.3) and its dual problem (2.4)–(2.6), strong duality holds [13]; i.e., their optimal values are equal ($f^*_{sos} = f^*_{mom}$). Hence f^*_{mom} is also a lower bound for the global minimum f^* of $f(x)$.

Now let us see how to extract minimizer(s) from optimal solutions to (2.4)–(2.6). Let y^* be one optimal solution. If moment matrix $\mathcal{M}_d(y^*)$ has rank one, then there exists one vector w such that $\mathcal{M}_d(y^*) = ww^T$. Normalize w so that $w_{(0,\dots,0)} = 1$. Set $x^* = w(2 : n+1)$. Then the relation $\mathcal{M}_d(y^*) = ww^T$ immediately implies that $y^* = \mathbf{m}_{2d}(x^*)$, i.e., $y^*_\alpha = (x^*)^\alpha$, so $f^*_{mom} = f(x^*)$. This says that a lower bound of $f(x)$ is attained at one point x^*. So x^* is one global minimizer.

When moment matrix $\mathcal{M}_d(y^*)$ has rank more than one, the process described above does not work. However, if $\mathcal{M}_d(y^*)$ satisfies the so-called *flat extension condition*

$$\operatorname{rank} \mathcal{M}_k(y^*) = \operatorname{rank} \mathcal{M}_{k+1}(y^*)$$

for some $0 \le k \le m-1$, we can extract more than one minimizer (in this case the global solution is not unique). When the flat extension condition is met, it can be shown [7] that there exist distinct vectors $u_1, \dots, u_r$ such that

$$\mathcal{M}_d(y^*) = \lambda_1 \mathbf{m}_d(u_1) \cdot \mathbf{m}_d(u_1)^T + \cdots + \lambda_r \mathbf{m}_d(u_r) \cdot \mathbf{m}_d(u_r)^T$$

for some $\lambda_i > 0$, $\sum_{i=1}^r \lambda_i = 1$. Here $r = \operatorname{rank} \mathcal{M}_d(y^*)$. The set $\{u_1, \dots, u_r\}$ is called an *r-atomic representing support* for moment matrix $\mathcal{M}_d(y^*)$. All the vectors $u_1, \dots, u_r$ can be shown to be global minimizers. They can be computed by solving some particular eigenvalue problem. We refer to [7] for flat extension conditions in moment problems and [9] for extracting minimizers.

2.3. Exploiting sparsity in SOS relaxation. As mentioned in the previous subsections, the size of matrix W in SOS relaxation is $\binom{n+d}{d}$, which can be very large. So SOS relaxation is expensive when either n or d is large. This is true for general dense polynomials. However, if $f(x)$ is sparse, i.e., its support $\mathcal{F} = \operatorname{supp}(f)$ is small, the size of the resulting SDP can be reduced significantly. Without loss of generality, assume $(0, \dots, 0) \in \mathcal{F}$. Then $\operatorname{supp}(f) = \operatorname{supp}(f - \gamma)$ for any number γ.

Suppose $f(x) - \gamma = \sum_i \phi_i(x)^2$ is an SOS decomposition. Then by Theorem 1 in [25] we have

$$\operatorname{supp}(\phi_i) \subset \mathcal{F}^0 := \left(\text{the convex hull of } \frac{1}{2}\mathcal{F}^e \right)$$

where $\mathcal{F}^e = \{\alpha \in \mathcal{F} : \ \alpha \text{ is an even integer vector}\}$. There exists some work [12, 28] on exploiting sparsity further. Here we briefly describe the technique introduced in [28].

For polynomial $f(x)$, define its *csp graph* $G = ([n], E)$ such that $(i, j) \in E$ if and only if $x_i x_j$ appears in some monomial of $f(x)$. Let $\{C_1, C_2, \dots, C_K\}$ be the set of all maximal cliques of graph G. Waki et al. [28] proposed to represent $f(x) - \gamma$ as

$$f(x) - \gamma = \sum_{i=1}^K s_i(x), \quad \text{each } s_i(x) \text{ being SOS } \operatorname{supp}(s_i) \subset C_i.$$

Theoretically, when $f(x) - \gamma$ is SOS, the above representation may not hold (see Example 3.5). It is also difficult to find all the maximal cliques of graph G. Waki

et al. [28] propose to replace $\{C_1, C_2, \ldots, C_K\}$ by the set of all maximal cliques of one *chordal extension* of G. We refer to [3] for properties of chordal graphs. For chordal graphs, there are efficient methods to find all the maximal cliques. Chordal extension is essentially the *sparse symbolic Cholesky factorization*; the sparsity of matrix factors represents the chordal extension. To find the minimum chordal extension requires sparse Cholesky factorization with the smallest number of fill-ins, which is difficult generally. However, some heuristics like minimum degree ordering are usually efficient in practice in finding a good approximation. We refer to [28] for more details on how to get an efficient chordal extension.

We remark that in the worst case the sparse SOS relaxation above might be weaker than the general dense SOS relaxation even when the chordal extension is applied, as shown by Example 3.5.

There is much work in exploiting sparsity in SOS relaxations. We refer to [8, 11, 12, 24, 28] and the references therein.

3. The sparse SOS relaxation. Throughout this paper, we assume $f(x) = \sum_{i=1}^m f_i(x_{\Delta_i})$. Let $\|\Delta\|$ be the maximum cardinality of Δ_i, i.e., $\|\Delta\| = \max_i |\Delta_i|$. We are interested in the case that $\|\Delta\| \ll n$. To find the global minimum f^* of $f(x)$, we propose the following sparse SOS relaxation:

$$\begin{aligned} f_\Delta^* := \max \quad & \gamma \\ s.t. \quad & f(x) - \gamma \in \sum_{i=1}^m \sum \mathbb{R}_d[x_{\Delta_i}]^2. \end{aligned}$$

In terms of SDP, the above SOS relaxation is essentially the same as

$$f_\Delta^* := \max \quad \gamma \tag{3.1}$$

$$s.t. \quad f(x) - \gamma = \sum_{i=1}^m \mathbf{m}_d(x_{\Delta_i})^T W_i \mathbf{m}_d(x_{\Delta_i}), \tag{3.2}$$

$$W_i \succeq 0,\ i = 1, \ldots, m. \tag{3.3}$$

Notice that (3.2) is an identity. Let

$$\mathcal{F}_i = \{\alpha \in \mathbb{N}^n : \mathrm{supp}(\alpha) \subset \Delta_i,\ |\alpha| \le 2d\}, \quad \mathcal{F} = \bigcup \mathcal{F}_i. \tag{3.4}$$

Write $f(x) = \sum_\alpha f_\alpha x^\alpha$. Since $f(x) = \sum_i f_i(x_{\Delta_i})$, $f_\alpha \neq 0$ implies that $\alpha \in \mathcal{F}$. By comparing coefficients of both sides of (3.2), we have equality constraints

$$f_0 - \gamma = \sum_{i=1}^m W_i(0,0), \quad f_\alpha = \sum_{i=1}^m \sum_{\eta+\tau=\alpha} W_i(\eta,\tau),\ \ \forall\, \alpha \neq 0. \tag{3.5}$$

Now we derive the dual problem for (3.1)–(3.3). Notice that constraint (3.2) is equivalent to the equality constraints (3.5). Let $y = (y_\alpha)_{\alpha\in\mathcal{F}}$ be the Lagrange multipliers for equations in (3.5), and U_i be the Lagrange multipliers for inequalities in (3.3). Each U_i is also positive semidefinite. The Lagrange function for problem (3.1)–(3.3) is

$$\begin{aligned} \mathcal{L} &= \gamma + \left(f_0 - \gamma - \sum_i W_i(0,0)\right) y_0 + \sum_{0\neq\alpha\in\mathcal{F}} \left(f_\alpha - \sum_{i=1}^m \sum_{\eta+\tau=\alpha} W_i(\eta,\tau)\right) y_\alpha + \sum_i W_i \bullet U_i \\ &= \gamma(1-y_0) + \sum_{\alpha\in\mathcal{F}} f_\alpha y_\alpha + \sum_{i=1}^m \sum_{\alpha\in\mathcal{F}} \sum_{\eta+\tau=\alpha} W_i(\eta,\tau)\Big(U_i(\eta,\tau) - y_\alpha\Big). \end{aligned}$$

So we can see that

$$\max_{\gamma, W_i} \mathcal{L}(\gamma, W_i, y_\alpha, U_i) = \begin{cases} \sum_\alpha f_\alpha y_\alpha & \text{if } y_0 = 1, U_i = \mathcal{M}_d^{\Delta_i}(y) \succeq 0\,; \\ +\infty & \text{otherwise.} \end{cases}$$

Therefore the dual of (3.1)–(3.3) is

$$f_\Sigma^* := \min \quad \sum_{\alpha \in \mathcal{F}} f_\alpha y_\alpha \tag{3.6}$$

$$s.t. \quad \mathcal{M}_d^{\Delta_i}(y) \succeq 0,\ i = 1, \ldots, m \tag{3.7}$$

$$y_0 = 1. \tag{3.8}$$

3.1. Complexity comparison. Since the dual of the standard or sparse SOS relaxation not only returns the SOS lower bound but also provides the moment matrix to help extract minimizers, we compare the computational complexity of (2.4)–(2.6) and (3.6)–(3.8). The LMI (2.5) is of size $\binom{n+d}{d} = \mathcal{O}(n^d)$ and has $\binom{n+2d}{2d} = \mathcal{O}(n^{2d})$ decision variables. At each step of an interior-point method (e.g., the dual scaling method [2]), the complexity for solving (2.4)–(2.6) is $\mathcal{O}(n^{6d})$. On the other hand, (3.7) has m LMIs, which are of sizes at most $\binom{\|\Delta\|+d}{d} = \mathcal{O}(\|\Delta\|^d)$, and $\mathcal{O}\big(m\binom{\|\Delta\|+2d}{2d}\big) = \mathcal{O}(m\|\Delta\|^{2d})$ decision variables. At each step of interior-point methods, the complexity for solving (3.6)–(3.8) is $\mathcal{O}(m^3\|\Delta\|^{6d})$. When $\|\Delta\|$ is independent of n and $m = \mathcal{O}(n^p)$ with $p < 2d$, then

$$\mathcal{O}\left(m^3\|\Delta\|^{6d}\right) \ll \mathcal{O}\left(n^{6d}\right).$$

Therefore (3.6)–(3.8) is much easier to solve than (2.4)–(2.6).

The complexity of sparse SOS relaxation in [28] depends on the chordal extension of the csp graph. In the worst case, it can be as big as for the general SOS relaxation (2.4)–(2.6). Let Ω be the maximum size of the maximal cliques of the chordal extension. In practice, Ω is often bigger than or equal to $\|\Delta\|$. When $\Omega > \|\Delta\|$, the SOS relaxation (3.6)–(3.8) is usually more efficient.

3.2. Lower bound analysis. Recall that $\mathcal{F}_i = \{\alpha \in \mathbb{N}^n : \text{supp}(\alpha) \subset \Delta_i,\ |\alpha| \le 2d\}$. From the representation (1.1) of $f(x)$, we have

$$\text{supp}(f) \subseteq \bigcup_{i=1}^m \mathcal{F}_i.$$

This leads us to think that the relaxation (3.6)–(3.8) should give reasonable lower bounds, although it might be weaker than the general SOS (see Example 3.5).

THEOREM 3.1. *The optimal values $f_\Sigma^*, f_\Delta^*, f_{sos}^*$, and f^* satisfy the relationship*

$$f_\Sigma^* = f_\Delta^* \le f_{sos}^* \le f^*.$$

Proof. The latter two inequalities are obvious because the feasible region defined by (3.7)–(3.8) contains the one defined by (2.5)–(2.6). To prove the first equality, by the standard duality argument for convex program, it suffices to show that (3.7) admits a strict interior point. Define $\hat{y} = (\hat{y}_\alpha)_{\alpha \in \mathcal{F}}$ as

$$\hat{y}_\alpha := \frac{\int_{\mathbb{R}^n} x^\alpha e^{-\|x\|_2^2}\, dx}{\int_{\mathbb{R}^n} e^{-\|x\|_2^2}\, dx}.$$

For every nonzero vector $\xi = (\xi_\alpha)_{\alpha \in \mathcal{F}_i}$, we have

$$\xi^T M_d^{\Delta_i}(\hat{y})\, \xi = \frac{\int_{\mathbb{R}^n} \left(\sum_{|\alpha| \le d} \xi_\alpha x^\alpha \right)^2 e^{-\|x\|_2^2} dx}{\int_{\mathbb{R}^n} e^{-\|x\|_2^2} dx} > 0.$$

So $M_d^{\Delta_i}(\hat{y}) \succ 0$ for every $1 \le i \le m$. Therefore $\hat{y}$ is an interior point for (3.6)–(3.8), which implies the strong duality $f_\Sigma^* = f_\Delta^*$. ∎

Remark 3.2. Theorem 3.1 implies that the lower bound f_Δ^* given by (3.1)–(3.3) is weaker than the SOS lower bound f_{sos}^*. There are examples such that $f_\Delta^* < f_{sos}^*$ (see Example 3.5). However, in many numerical simulations, the lower bound f_Δ^* is very useful. For randomly generated polynomials, as shown in section 4, it frequently happens that $f_\Delta^* = f_{sos}^*$. On the other hand, under some conditions, we can prove $f_\Delta^* = f_{sos}^*$.

Suppose $\Delta_1, \Delta_2, \ldots, \Delta_m$ satisfy the *running intersection property*:

$$\text{For every } 1 \le i \le m-1,\ \exists\, k \le i \text{ such that } \quad \Delta_{i+1} \cap \left(\bigcup_{j=1}^{i} \Delta_j \right) \subsetneq \Delta_k. \tag{3.9}$$

THEOREM 3.3. *Suppose* (3.6)–(3.8) *has an optimal solution* y^* *such that each* $\mathcal{M}_{d_i}^{\Delta_i}(y^*)$ *has a representing measure* μ_i *on* $\mathbb{R}^{\Delta_i}$. *If condition* (3.9) *holds, then* $f_\Delta^* = f_{sos}^*$.

Proof. For any Δ_i, Δ_j, $\mathcal{M}_d^{\Delta_i \cap \Delta_j}(y^*)$ is a common principle submatrix of $\mathcal{M}_d^{\Delta_i}(y^*)$ and $\mathcal{M}_d^{\Delta_j}(y^*)$. So the marginals of measures μ_i are consistent; i.e., the restrictions of these measures on the common subspaces are the same. By Lemma 6.4 in [15], there exists a measure on $\mathbb{R}^n$ such that μ_i is the marginal of μ with respect to Δ_i for all $i = 1, \ldots, m$. Define vector $\tilde{y}$ such that

$$\mathcal{M}_d(\tilde{y}) = \int_{\mathbb{R}^n} \mathbf{m}_d(x) \mathbf{m}_d(x)^T \mu(dx).$$

Then every $\mathcal{M}_d^{\Delta_i}(y^*)$ is a principle submatrix of $\mathcal{M}_d(\tilde{y})$. So $\tilde{y}_\alpha = y_\alpha^*$ whenever $\mathrm{supp}(\alpha) \subset \Delta_j$ for some j. Since the $f_\alpha \ne 0$ implies $\mathrm{supp}(\alpha) \subset \Delta_j$ for some j, we know the objective value of (3.6) is the same for y^* and $\tilde{y}$. Thus $f_{sos}^* \le f_\Delta^*$. Since $f_{sos}^* \ge f_\Delta^*$, we get $f_{sos}^* = f_\Delta^*$. ∎

Remark 3.4. The running intersection property (3.9) alone is not sufficient to guarantee the equality $f_\Delta^* = f_{sos}^*$, as shown by the following example.

Example 3.5. $f(x) = f_1(x_1, x_2) + f_2(x_2, x_3)$ where $f_1 = x_1^4 + (x_1 x_2 - 1)^2$ and $f_2 = x_2^2 x_3^2 + (x_3^2 - 1)^2$. Solving dense SOS relaxation (2.1)–(2.3) and sparse SOS relaxation (3.1)–(3.3) numerically, we find that

$$f_\Delta^* \approx 5.0 \cdot 10^{-5} \ < \ f_{sos}^* \approx 0.8499.$$

Actually the minimum $f^* \approx 0.8650$. First, solve equation $\nabla f(x) = 0$, and evaluate $f(x)$ on these critical points, and then we find the minimum of these critical values is about 0.8650. So $f^* < 1$. Second, we prove that the minimum f^* is attainable. Let $\{x^{(k)}\}$ be a sequence such that $f(x^{(k)}) \to f^*$ as k goes to infinity. We claim that the sequence $\{x^{(k)}\}$ must be bounded.

Otherwise, suppose $x^{(k)} \to \infty$. Thus at least one of coordinates $x_1^{(k)}$, $x_2^{(k)}$, $x_3^{(k)}$ should go to infinity. If either $x_1^{(k)}$ or $x_3^{(k)}$ goes to infinity, then $f(x^{(k)})$ goes to infinity, which is not possible. So $x_2^{(k)} \to \infty$. Since $\{f(x^{(k)})\}$ is bounded, without loss of generality, we assume $x_1^{(k)} \to a_1$, $x_1^{(k)}x_2^{(k)} \to a_{12}$, $x_2^{(k)}x_3^{(k)} \to a_{23}$, $x_3^{(k)} \to a_3$ for some numbers a_1, a_{12}, a_{23}, a_3. If $a_3 = 1$, then $x_2^{(k)}$ is convergent to a_2, which is not possible. And, if $a_3 \neq 0$, then $x_2^{(k)}x_3^{(k)}$ goes to infinity, which is also not possible. So $a_3 = 0$, and hence

$$f\left(x^{(k)}\right) \geq \left(x_3^2 - 1\right)^2 \to 1 > f^*,$$

which is a contradiction.

So the sequence $\{x^{(k)}\}$ is bounded and has an accumulation point x^*. Then we must have $f(x^*) = f^*$, which means that f^* is attained at some point. From the computation of critical values, we know $f^* \approx 0.8650$. For this polynomial, both the dense and sparse SOS relaxation are not exact: $f_\Delta^* < f_{sos}^* < f^*$, and the method in [28] gives the same lower bound f_Δ^*.

COROLLARY 3.6. *If all f_i are quadratic and condition* (3.9) *holds, then $f_{sos}^* = f_\Delta^*$.*

Proof. When all f_i are quadratic, i.e., $d_i = 1$, the entries of moment matrix $\mathcal{M}_1^{\Delta_i}$ are the first and second order moments. The positive semidefiniteness of $\mathcal{M}_1^{\Delta_i}$ implies $\mathcal{M}_1^{\Delta_i}$ has a representing measure. Then the conclusion is immediately implied by Theorem 3.3. ∎

Remark 3.7. If the running intersection condition (3.9) fails, then Corollary 3.6 is no longer true, as shown by the example below.

Example 3.8. Consider the polynomial $f(x) = f_1(x_1, x_2) + f_2(x_2, x_3) + f_3(x_1, x_3)$ where $f_1 = \frac{1}{2}(x_1^2 + x_2^2) + 2x_1x_2$, $f_2 = \frac{1}{2}(x_2^2 + x_3^2) + 2x_2x_3$, and $f_3 = \frac{1}{2}(x_1^2 + x_3^2) + 2x_1x_3$. In this case

$$\Delta_1 = \{1, 2\}, \quad \Delta_2 = \{2, 3\}, \quad \Delta_3 = \{1, 3\}.$$

The running intersection property (3.9) fails. But we have $f_\Delta^* = -\infty < f_{sos}^* = f^* = 0$.

3.3. Extraction of minimizers. In this subsection, we discuss how to extract minimizer(s) $x^* = (x_1^*, \ldots, x_n^*)$. Suppose $y^* = (y_\alpha^*)_{\alpha \in \mathcal{F}}$ is one optimal solution to (3.6)–(3.8). Let $\delta_i = \{i\}$ for every i. The entries of y^* whose indices are supported in δ_i are

$$y_0^*, y_{e_i}^*, y_{2e_i}^*, \ldots, y_{2de_i}^*,$$

which are the entries of the moment matrix $M_d^{\delta_i}(y^*)$. So coordinate x_i^* can be extracted from moment matrix $M_d^{\delta_i}(y^*)$ if it satisfies the flat extension condition. Let $\mathcal{V}_i$ be the set of all the points that can be extracted from the moment matrix $M_d^{\delta_i}(y^*)$. If $\mathcal{V}_i$ is a singleton, then x_i^* has a unique choice.

The situation is more subtle if some $\mathcal{V}_i$ has cardinality greater than one. Suppose for some $i, j \in [n]$ we have $|\mathcal{V}_i| > 1$ and $|\mathcal{V}_j| > 1$. Can $x_i^* x_j^*$ appear simultaneously in the optimal solution x^* for arbitrarily chosen $x_i^* \in \mathcal{V}_i$, $x_j^* \in \mathcal{V}_j$? The answer is obviously no! For instance, the polynomial

$$\left(x_1^2 - 1\right)^2 + \left(x_2^2 - 1\right)^2 + (x_1 - x_2)^2$$

has only two global minimizers $\pm(1,1)$. We find that $\mathcal{V}_1 = \mathcal{V}_2 = \{1,-1\}$. But obviously $(1,-1)$ and $(-1,1)$ are not global minimizers.

Now what is the rule for matching x_i^* and x_j^* if $|\mathcal{V}_i| > 1$ or $|\mathcal{V}_j| > 1$? So far we have not yet used the information of moment matrix $M_d^{\Delta_i}(y^*)$. If $M_d^{\Delta_i}(y^*)$ also satisfies the flat extension condition, we can extract the tuples $x_{\Delta_i}^* = (x_k^*)_{k\in\Delta_i}$ from $M_d^{\Delta_i}(y^*)$. Let $\mathcal{X}_{\Delta_i}$ be the set of all such tuples that can be extracted from $M_d^{\Delta_i}(y^*)$. One might ask whether $\mathcal{V}_i$ and $\mathcal{X}_{\Delta_i}$ are consistent, that is, does $x_{\Delta_i}^* \in \mathcal{X}_{\Delta_i}$ imply that $x_k^* \in \mathcal{V}_k$ for all $k \in \Delta_i$? Under the flat extension assumption, the answer is yes, which is due to the following theorem.

THEOREM 3.9. *Suppose y^* is one optimal solution to* (3.6)–(3.8) *such that all $M_d^{\Delta_i}(y^*)$ satisfy the flat extension condition. Then for any $x_{\Delta_i}^* \in \mathcal{X}_{\Delta_i}$, it holds that $x_k^* \in \mathcal{V}_k$ for all $k \in \Delta_i$.*

Proof. Let $\mathcal{X}_{\Delta_i} = \{x_{\Delta_i}^{(1)}, x_{\Delta_i}^{(2)}, \ldots, x_{\Delta_i}^{(r)}\}$ be the r-atomic representing support for $\mathcal{M}_d^{\Delta_i}(y^*)$. Then we have decomposition

$$M_d^{\Delta_i}(y^*) = \sum_{\ell=1}^{r} \lambda_\ell \, \mathbf{m}_2\left(x_{\Delta_i}^{(\ell)}\right) \mathbf{m}_2\left(x_{\Delta_i}^{(\ell)}\right)^T$$

for some $\lambda_1, \ldots, \lambda_r > 0, \sum_{\ell=1}^r \lambda_\ell = 1$. Notice that $M_d^{\delta_k}(y^*)$ is a principle submatrix of $M_d^{\Delta_i}(y^*)$. So we also have that for every $k \in \Delta_i$

$$M_d^{\delta_k}(y^*) = \sum_{\ell=1}^{r} \lambda_\ell \, \mathbf{m}_2\left(x_k^{(\ell)}\right) \mathbf{m}_2\left(x_k^{(\ell)}\right)^T.$$

This means that $\{x_k^{(1)}, x_k^{(2)}, \ldots, x_k^{(r)}\}$ is a r-atomic representing support for moment matrix $M_d^{\delta_k}(y^*)$ (some $x_k^{(\ell)}$ might be the same). By the definition of $\mathcal{V}_i$, we have $\{x_k^{(1)}, \ldots, x_k^{(r)}\} \subseteq \mathcal{V}_k$. ∎

THEOREM 3.10. *Suppose y^* is one optimal solution to* (3.6)–(3.8) *such that all $M_d^{\Delta_i}(y^*)$ satisfy the flat extension condition. Then any $x^* = (x_1^*, \ldots, x_n^*)$ with $x_k^* \in \mathcal{V}_k$ and $x_{\Delta_i}^* \in \mathcal{X}_{\Delta_i}$ for all k and i is a global optimal minimizer of $f(x)$.*

Proof. Fix x^* as in the theorem. Since $M_d^{\Delta_i}(y^*)$ satisfies the flat extension condition, we have the decomposition

$$M_d^{\Delta_i}(y^*) = \lambda_{\Delta_i} \mathbf{m}_d(x_{\Delta_i}^*)\mathbf{m}_2(x_{\Delta_i}^*)^T + \hat{M}_{\Delta_i},$$

where $1 \geq \lambda_{\Delta_i} > 0$ and $\hat{M}_{\Delta_i} \succeq 0$. Now let $\lambda = \min_i \lambda_{\Delta_i} > 0$ and

$$M_{\Delta_i} = (\lambda_{\Delta_i} - \lambda)\mathbf{m}_2(x_{\Delta_i}^*)\mathbf{m}_2(x_{\Delta_i}^*)^T + \hat{M}_{\Delta_i} \succeq 0.$$

Notice that $\hat{M}_{\Delta_i}$ and M_{Δ_i} are also moment matrices. Without loss of generality, we can assume $\lambda < 1$, since otherwise each $M_2^{\Delta_i}(y^*)$ has rank one and then x^* is obviously a global minimizer. For every $\alpha \in \mathcal{F}_i$, define $\hat{y}_\alpha = \left(x_{\Delta_i}^*\right)^\alpha$ and $\hat{y} = (\hat{y}_\alpha)_{\alpha\in\mathcal{F}}$. Let $\tilde{y} = (\tilde{y}_\alpha)_{\alpha\in\mathcal{F}}$ be the vector such that $y^* = \lambda\hat{y} + (1-\lambda)\tilde{y}$. Then it holds

$$M_d^{\Delta_i}(y^*) = \lambda M_d^{\Delta_i}(\hat{y}) + (1-\lambda) M_d^{\Delta_i}(\tilde{y}).$$

Obviously vector $\tilde{y}$ is feasible for (3.7)–(3.8) since

$$M_d^{\Delta_i}(\tilde{y}) = \frac{1}{1-\lambda}\left(M_d^{\Delta_i}(y^*) - \lambda M_d^{\Delta_i}(\hat{y})\right) = \frac{1}{1-\lambda} M_{\Delta_i} \succeq 0.$$

Since y^* is optimal, we can see $\sum_{\alpha\in\mathcal{F}} f_\alpha y_\alpha^* \le \sum_{\alpha\in\mathcal{F}} f_\alpha \hat{y}_\alpha$ and $\sum_{\alpha\in\mathcal{F}} f_\alpha y_\alpha^* \le \sum_{\alpha\in\mathcal{F}} f_\alpha \tilde{y}_\alpha$. By linearity, it holds

$$f_\Delta^* = \sum_{\alpha\in\mathcal{F}} f_\alpha y_\alpha^* = \lambda \sum_{\alpha\in\mathcal{F}} f_\alpha \hat{y}_\alpha + (1-\lambda) \sum_{\alpha\in\mathcal{F}} f_\alpha \tilde{y}_\alpha.$$

Therefore, we must have $\sum_{\alpha\in\mathcal{F}} f_\alpha \hat{y}_\alpha = f_\Delta^*$ since $0 < \lambda < 1$. On the other hand, by the definition of $\hat{y}$, we know $f(x^*) = \sum_{\alpha\in\mathcal{F}} f_\alpha \hat{y}_\alpha = f_\Delta^*$. Thus x^* is one point at which the polynomial $f(x)$ attains its lower bound f_Δ^*, which implies that x^* is a global minimizer of $f(x^*)$. □

The algorithm for minimizing $f(x)$ via sparse SOS relaxation (3.1)–(3.3) is as follows.

ALGORITHM 3.11 (Minimizing sum of polynomials).
Input: $n, m, \Delta_i, f_i(x_{\Delta_i})\,(i = 1, \ldots, m)$
Output: $\mathcal{V}_i$ *and* $\mathcal{X}_{\Delta_i}\,(i = 1, \ldots, m)$
Begin
Step 1: *Solve the dual problem* (3.6)–(3.8). *Get the optimal solution* y^*.
Step 2: *For each* $1 \le k \le n$, *find the set* $\mathcal{V}_k$ *of points that can be extracted from* $M_d^{\delta_k}(y^*)$.
Step 3: *For every* k *with* $|\mathcal{V}_k| > 1$, *find the set* $\mathcal{X}_{\Delta_i}$ *from* $M_d^{\Delta_i}(y^*)$ *whenever* $k \in \Delta_i$.
End

As an example, let us illustrate how to solve the global optimization problem

$$\min_{x\in\mathbb{R}^3} \quad \underbrace{\left(x_1^2 - 1\right)^2 + (x_1 - x_2)^4}_{f_1(x_{\Delta_1})} + \underbrace{(x_2 - x_3)^4}_{f_2(x_{\Delta_2})}$$

and find global minimizers. Here $\Delta_1 = \{1, 2\}$ and $\Delta_2 = \{2, 3\}$. Solve the dual problem (3.6)–(3.8) and we get solutions

$$\mathcal{M}_1^{\Delta_1}(y^*) = \mathcal{M}_1^{\Delta_2}(y^*) = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}.$$

Both $\mathcal{M}_1^{\Delta_1}(y^*)$ and $\mathcal{M}_1^{\Delta_2}(y^*)$ have rank two and satisfy the flat extension condition. Using the technique from [9], we can extract

$$\mathcal{V}_1 = \mathcal{V}_2 = \mathcal{V}_3 = \{-1, 1\}$$

and

$$\mathcal{X}_{\Delta_1} = \left\{ \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\}, \quad \mathcal{X}_{\Delta_2} = \left\{ \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\}.$$

Since the x_2-component from $\mathcal{X}_{\Delta_1}$ and $\mathcal{X}_{\Delta_2}$ must be the same, we know there are two global minimizers $x^* = \pm(1, 1, 1)$.

3.4. Nonlinear least squares problems. Now we consider the special case that each $f_i(x_{\Delta_i})$ is a square of some polynomial, say, $f_i(x_{\Delta_i}) = g_i^2(x_{\Delta_i})$. Then the global minimization of $f(x) = \sum_i f_i(x_{\Delta_i})$ is equivalent to solving the nonlinear least squares (NLS) problem associated with the polynomial system:

$$g_1(x_{\Delta_1}) = g_2(x_{\Delta_2}) = \cdots = g_m(x_{\Delta_m}) = 0. \tag{3.10}$$

In this situation, the polynomial function is often nonconvex and it is very difficult for general numerical optimization schemes like branch-bound to find the global minimizer of $f(x)$.

THEOREM 3.12. *If the polynomial system* (3.10) *admits a solution, then the sparse SOS relaxation* (3.1)–(3.3) *is exact; i.e.,* $f_\Delta^* = f_{sos}^* = f^*$.

Proof. Obviously $f^* = 0$. And $\gamma = 0$ is a feasible solution to problem (3.1)–(3.3), since $f(x)$ itself is a sparse SOS representation as in (3.2)–(3.3). So $f_\Delta^* \geq 0$, and hence all the inequalities in the Theorem 3.1 become equalities. ∎

Remark 3.13. When the polynomial system (3.10) admits a solution, we necessarily have $f^* = 0$. This might be trivial in some sense. However, the optimal solution y^* to the dual problem (3.6)–(3.8) can help recover the real zeros of polynomial system (3.10), which are absolutely the global minimizers of $f(x)$. See the example below.

Example 3.14. Consider the sparse polynomial system

$$\begin{aligned} 2x_1^2 - 3x_1 + 2x_2 - 1 &= 0, \\ 2x_i^2 + x_{i-1} - 3x_i + 2x_{i+1} - 1 &= 0\,(i = 2, \ldots, n-1), \\ 2x_n^2 + x_{n-1} - 3x_n - 1 &= 0. \end{aligned}$$

This polynomial system is consistent and has at least two real solutions. Set $n = 20$. We apply sparse SOS relaxation (3.1)–(3.3) to solve the least squares problem and get the lower bound $f_\Delta^* \approx -2.0 \cdot 10^{-11}$. Using the optimal dual solution, we obtain two real solutions (only the first four digits are shown):

$\hat{x} = (1.8327,\ -0.1097,\ -0.5929,\ -0.6860,\ -0.7032,\ -0.7064,\ -0.7070,\ -0.7071,\ -0.7071,\ -0.7071,$
$-0.7071,\ -0.7070,\ -0.7068,\ -0.7064,\ -0.7051,\ -0.7015,\ -0.6919,\ -0.6658,\ -0.5960,\ -0.4164)$
$\tilde{x} = (-0.5708,\ -0.6819,\ -0.7025,\ -0.7063,\ -0.7070,\ -0.7071,\ -0.7071,\ -0.7071,\ -0.7071,\ -0.7071,$
$-0.7071,\ -0.7070,\ -0.7068,\ -0.7064,\ -0.7051,\ -0.7015,\ -0.6919,\ -0.6658,\ -0.5960,\ -0.4164).$

3.5. A sparser SOS relaxation. From Theorem 3.12, we know the sparse SOS relaxation (3.1)–(3.3) is exact whenever the polynomial system (3.10) admits a solution, and the optimal dual solution can help recover the real zeros. This fact makes it possible to exploit the sparsity of each $f_i(x_{\Delta_i})$ further. In (3.1)–(3.3), we assume each $f_i(x_{\Delta_i})$ is a dense polynomial. However, if each $f_i(x_{\Delta_i})$ is sparse, we can get a sparser SOS relaxation. It is obvious that

$$\operatorname{supp}(f_i) \subseteq \mathcal{G}_i + \mathcal{G}_i,$$

where $\mathcal{G}_i$ is the convex hull of $\{\alpha \in \mathbb{N}^n : 2\alpha \in \operatorname{supp}(f_i)\}$. This motivates us to propose the sparser SOS relaxation

$$f_{\Delta_s}^* := \max \quad \gamma \tag{3.11}$$

$$s.t. \quad f(x) - \gamma = \sum_{i=1}^{m} \mathbf{m}_{\mathcal{G}_i}(x)^T W_i \mathbf{m}_{\mathcal{G}_i}(x), \tag{3.12}$$

$$W_i \succeq 0,\ i = 1, \ldots, m. \tag{3.13}$$

Here $\mathbf{m}_{\mathcal{G}_i}(x_{\Delta_i})$ is the column vector of all monomials in x with exponents from $\mathcal{G}_i$. The size of matrix W_i is equal to the cardinality of $\mathcal{G}_i$. Similar to (3.1)–(3.3), the dual of (3.11)–(3.13) can be derived to be

$$f^*_{\Sigma_s} := \min \quad \sum_\alpha f_\alpha y_\alpha \tag{3.14}$$

$$s.t. \quad \mathcal{M}_{\mathcal{G}_i}(y) \succeq 0,\ i = 1, \ldots, m, \tag{3.15}$$

$$y_0 = 1. \tag{3.16}$$

Here the sparse moment matrix $\mathcal{M}_{\mathcal{G}_i}(y)$ is indexed by vectors from $\mathcal{G}_i$ and defined as

$$\mathcal{M}_{\mathcal{G}_i}(y)(\alpha, \beta) = y_{\alpha+\beta}$$

for all $\alpha, \beta \in \mathcal{G}_i$.

THEOREM 3.15. *The optimal values $f^*_{\Sigma_s}$, $f^*_{\Delta_s}$, f^*_{Σ}, f^*_{Δ}, f^*_{sos} and f^* satisfy the relationship*

$$f^*_{\Sigma_s} = f^*_{\Delta_s} \le f^*_{\Sigma} = f^*_{\Delta} \le f^*_{sos} \le f^*.$$

Proof. Applying the standard duality theory in convex programming as in the proof of Theorem 3.1, we can get the first equality from the left by proving (3.12)–(3.13) has a strict interior point. Since the relaxation (3.11)–(3.13) is a special case of (3.1)–(3.3), we obtain the first inequality from the left. The other relations follow Theorem 3.1. □

THEOREM 3.16. *Suppose $f_i(x_{\Delta_i}) = g_i^2(x_{\Delta_i})$. If the polynomial system* (3.10) *admits a solution, then the sparse SOS relaxation* (3.11)–(3.13) *is exact, i.e., $f^*_{\Delta_s} = f^*_{sos} = f^*$.*

Proof. The proof is almost the same as for Theorem 3.12. Obviously $f^* = 0$. And $\gamma = 0$ is a feasible solution, since $f(x)$ itself is a sparse SOS representation as in (3.12)–(3.13). So $f^*_{\Delta_s} \ge 0$, and hence all the inequalities in the Theorem 3.15 become equalities. □

Remark 3.17. When the polynomial system (3.10) admits a solution, we must have $f^* = 0$. This lower bound itself might not be interesting. However, the optimal dual solution y^* to (3.14)–(3.16) can help recover the real zeros of polynomial system (3.10), which are absolutely the global minimizers of $f(x)$. This observation is very important and has many applications. See examples in subsection 5.1.

4. Numerical examples. In this section, we present some numerical experiments using sparse SOS relaxations (3.1)–(3.3) and (3.11)–(3.13). First, we use them to solve some test problems from unconstrained optimization. Second, we generate various random polynomials, test the performance of these sparse SOS relaxations, and compare with other methods. All the computations are implemented on a Linux machine with 0.98 GB memory and 1.46 GHz CPU. The SOS relaxations are solved by the software *SeDuMi* [27] using the *YALMIP* [17] interface. Throughout this section, the computation time is in CPU seconds. The accuracy of relaxations is measured by $\frac{|f(\hat{x})-\hat{f}|}{\max\{1,|f(\hat{x})|\}}$, where $\hat{x}$ is one extracted solution and $\hat{f}$ is the computed lower bound.

4.1. Some global optimization test problems. In this subsection, we apply SOS relaxations (3.1)–(3.3) and (3.11)–(3.13) to solve some global optimization test problems from [6, 18, 19]. The relaxation (3.1)–(3.3) is usually applied when each $f_i(\Delta_i)$ is almost dense, and the sparser relaxation (3.11)–(3.13) is usually applied

TABLE 1
The performance of sparse SOS relaxation (3.1)–(3.3).

	Chained singular		Chained wood		Gen. Rosen.	
n	accu.	time	accu.	time	accu.	time
100	3.2e − 09	2.72	3.5e − 10	1.52	9.0e − 8	0.95
200	3.0e − 10	5.29	3.7e − 10	2.25	1.8e − 7	1.46
300	5.0e − 09	8.01	3.8e − 10	3.19	2.7e − 7	2.24
400	5.0e − 10	11.64	3.9e − 10	4.12	3.6e − 7	2.88
500	4.9e − 09	33.09	3.9e − 10	5.12	4.5e − 7	3.45

when each $f_i(\Delta_i)$ is sparse. All the test functions in this subsection have global minimum $f^* = 0$. So we use the absolute value of the lower bounds f^*_Δ or $f^*_{\Delta_s}$ to measure the accuracy of the relaxation.

First, consider the following test functions.

- The chained singular function [6]:

$$f(x) = \frac{1}{10^5}\sum_{i\in J}\big((\,x_i + 10x_{i+1})^2 + 5(x_{i+2} - x_{i+3})^2 + (x_{i+1} - 2x_{i+2})^4 + 10(x_i - 10x_{i+3})^4\big)$$

  where $J = \{1, 3, 5, \ldots, n-3\}$ and n is a multiple of 4. The factor $\frac{1}{10^5}$ is used to scale the coefficients to avoid numerical troubles.

- The chained wood function [6]

$$f(x) = \sum_{i\in J}\big(\,100\left(x_{i+1} - x_i^2\right)^2 + (1 - x_i)^2 + 90\left(x_{i+3} - x_{i+2}^2\right)^2 + (1 - x_{i+2})^2 + 10(x_{i+1} + x_{i+3} - 2)^2 + 0.1(x_{i+1} - x_{i+3})^2\big)$$

  where $J = \{1, 3, 5, \ldots, n-3\}$ and n is a multiple of 4.

- The generalized Rosenbrock function [19]:

$$f(x) = \sum_{i=2}^{n}\left\{100\left(x_i - x_{i-1}^2\right)^2 + (1 - x_i)^2\right\}.$$

We apply SOS relaxation (3.1)–(3.3) to minimize these polynomial functions. The relaxation (3.1)–(3.3) is solved by the software *SeDuMi* using the *YALMIP* interface. The accuracy and consumed CPU time are in Table 1. The problems are solved from size 100 to 500. For these polynomials, the relaxation (3.1)–(3.3) is almost the same as the one in [28]. This is because the csp graphs of these polynomials are chordal graphs. However, if the csp graphs are sparse but their chordal extensions are much dense, then the relaxation in [28] is very similar to the dense SOS relaxation. In such situations, the relaxation (3.1)–(3.3) might be more suitable. For example, to minimize the sparse polynomial

$$\left(x_1^2 + x_2^2 - 1\right)^2 + \left(x_2^2 + x_3^2 - 1\right)^2 + \cdots + \left(x_{n-1}^2 + x_n^2 - 1\right)^2 + \left(x_n^2 + x_1^2 - 1\right)^2,$$

the chordal extension of the csp graph is the complete graph, and hence the sparse SOS relaxation using chordal extension is the same as the dense SOS relaxation. However, the sparse relaxation (3.1)–(3.3) is very suitable for this problem.

TABLE 2
The performance of sparse SOS relaxation (3.11)–(3.13).

Broyden tridiagonal			Broyden banded			Disc. bound val.		
n	accu.	time	n	accu.	time	n	accu.	time
100	1.2e − 7	2.65	10	3.6e − 11	9.72	10	6.0e − 12	0.92
200	2.3e − 7	2.69	15	2.2e − 10	17.28	20	3.4e − 11	1.57
300	5.0e − 7	3.58	20	1.6e − 10	25.27	25	1.6e − 11	2.28
400	3.0e − 6	4.53	25	1.8e − 10	35.19	30	1.1e − 11	2.47
500	4.1e − 6	5.44	30	4.9e − 10	45.30	35	3.9e − 11	3.00

Second, consider the following test functions.

- Broyden tridiagonal function [18]:

$$f(x) = \sum_{i=1}^{n} \left((3 - 2x_i)x_i - x_{i-1} - 2x_{i+1} + 1\right)^2,$$

  where $x_0 = x_{n+1} = 0$.
- Broyden banded function [18]:

$$f(x) = \sum_{i=1}^{n} \left(x_i \left(2 + 10x_i^2\right) + 1 - \sum_{j \in J_i} (1 + x_j)x_j \right)^2,$$

  where $J_i = \{j : j \neq i, \max(1, i-5) \leq j \leq \min(n, i+1)\}$.
- Discrete boundary value function [6]:

$$f(x) = \sum_{i=1}^{n} \left(2x_i - x_{i-1} - x_{i+1} + \frac{1}{2} h^2 (x_i + t_i + 1)^3 \right)^2,$$

  where $h = \frac{1}{n+1}$, $t_i = ih$, and $x_0 = x_{n+1} = 0$.

These three polynomials have sparse summand polynomial $f^*_{\Delta_i}$. So we apply the sparser SOS relaxation (3.11)–(3.13) and solve it by the software *SeDuMi* using the *YALMIP* interface. The computational results are in Table 2. All the problems are solved quite well in a few seconds.

For the Broyden tridiagonal function, we can also apply the sparse relaxation (3.1)–(3.3) or chordal extension from [28]. They are slightly more expensive. For $n = 500$, the problem can be solved in about ten seconds with similar accuracy. However, for the Broyden banded function and discrete boundary value function, the relaxation (3.1)–(3.3) and the method in [28] are much more expensive. For instance, when n has values 10 or bigger, they are usually difficult to implement due to computer memory restrictions.

We should mention that the Broyden banded function and discrete boundary value function have different representations as sums of small polynomials. For instance, if we expand all the squares, then they are sums of small polynomials which all have only two variables. However, the sparse SOS relaxations based on these new representations are usually too loose and not useful in practice, because they are very likely to be primarily infeasible and often no lower bounds can be obtained.

One interesting observation in Table 2 is that the accuracy for the Broyden tridiagonal function is not as high as for the other two functions. One possible reason is that the global minimizer of Broyden tridiagonal function is not unique and there

TABLE 3
Computational results for quartic polynomials with different sizes.

	$\|\Delta\| = 3$				$\|\Delta\| = 4$			
	CPU seconds			accu	CPU seconds			accu
n	max	avr.	min	max	max	avr.	min	max
20	0.85	0.62	0.54	4.1e − 9	1.46	1.15	0.91	2.4e − 9
40	1.22	1.07	0.91	1.9e − 9	2.86	2.49	2.25	2.9e − 9
60	1.80	1.55	1.45	2.9e − 9	4.43	4.17	3.91	3.1e − 9
80	2.30	2.18	2.02	2.3e − 9	6.26	5.94	5.24	3.7e − 9
100	3.02	2.70	2.33	2.8e − 9	7.85	7.41	7.01	5.0e − 9

are additional numerical troubles caused from extracting minimizers. This illustrates that the computation is more numerically difficult when there are multiple global solutions.

4.2. Randomly generated test problems. In this subsection, we present the computational results for randomly generated polynomials. The aim is to test the performance of the sparse SOS relaxation (3.1)–(3.3) for minimizing random polynomials and compare with other sparse SOS methods. For these randomly generated polynomials, solve the sparse relaxation (3.1)–(3.3) by the software *SeDuMi* using the *YALMIP* interface. Then we get lower bounds f_Δ^* and extract minimizers $\hat{x}$. Since we do not know the true global minimizers in advance, the accuracy of $\hat{x}$ can be measured by $err = \frac{|f(\hat{x})-f_\Delta^*|}{\max\{1,|f(\hat{x})|\}}$. The smaller err is, the more accurate $\hat{x}$ is, since f_Δ^* is a guaranteed lower bound.

4.2.1. Randomly generated sums of small polynomials. In this subsubsection, we randomly generate sparse polynomials $f(x)$ of the form (1.1) and use them to test the performance of the sparse relaxation (3.1)–(3.3). Then the csp graph of $f(x)$ is usually not chordal, and its chordal extension is often much less sparse. So the method in [28] is usually expensive for these polynomials. We let $m = n$ and choose f_i to have the form

$$f_i(x_{\Delta_i}) = \mathbf{m}_d(x_{\Delta_i})^T \cdot A_i \cdot \mathbf{m}_d(x_{\Delta_i}) + b_i^T \mathbf{m}_{2d-1}(x_{\Delta_i}),$$

where Δ_i are chosen to be random subsets of $[n]$ with cardinality at most $\|\Delta\|$. Here $N_i = \binom{|\Delta_i|+d}{d}$, $A_i = nI_{N_i} + BB^T$, $B \in \mathbb{R}^{N_i \times N_i}$, and $b_i \in \mathbb{R}^{\binom{|\Delta_i|+d-1}{d-1}}$ are random. So each A_i is positive definite. This choice guarantees that the global minimizers of $f(x)$ are contained in some compact set.

First, let $2d = 4$ and n be $20, 40, 60, 80, 100$. For each $\|\Delta\|$ (3 or 4) and n, we generate 100 random polynomials in the way described above. For each one, we solve the sparse SOS relaxation (3.1)–(3.3) by the software *SeDuMi* using the *YALMIP* interface, and get the lower bound f_Δ^* and optimal dual solution $\hat{y}$. For all these randomly generated polynomials, the moment matrices $\mathcal{M}_d^{\Delta_i}(\hat{y})$ have numerical rank one. So we can easily extract the minimizer $\hat{x}$. The maximum, average, and minimum of consumed CPU time are in Table 3. For these random polynomials, we just record the maximum error of the extracted minimizers. From Table 3, for $\Delta = 4$, we can find the global minimizer of a quartic sparse polynomial of 100 variables with error $O(10^{-9})$ within about 8 CPU seconds.

Second, let $n = 30$ and $2d$ be $4, 6, 8$. For each $\|\Delta\|$ (3 or 4) and $2d$, we generate 100 random polynomials in the way described in the above. For each one, solve the sparse SOS relaxation (3.1)–(3.3) by the software *SeDuMi* using the *YALMIP* interface,

TABLE 4
Computational results for polynomials of size $n = 30$ with different degrees.

	$\|\Delta\| = 3$				$\|\Delta\| = 4$			
	CPU seconds			err	CPU seconds			err
$2d$	max	avr	min	max	max	avr	min	max
4	1.01	0.87	0.77	2.5e − 9	2.33	1.93	1.65	2.4e − 9
6	3.22	2.96	2.67	1.8e − 9	17.16	14.92	11.71	2.2e − 9
8	13.07	11.44	10.13	1.7e − 8	136.67	119.90	107.28	9.4e − 8

and get the lower bounds f_Δ^* and optimal dual solution $\hat{y}$. Similarly, all moment matrices $\mathcal{M}_d^{\Delta_i}(\hat{y})$ have rank one, and the minimizer $\hat{x}$ can be extracted easily. The maximum, average, and minimum of the consumed CPU time, and the maximum error of extracted minimizers are in Table 4. For $\|\Delta\| = 4$, the global minimizer of such generated polynomials of degree 8 and 30 variables can be found with error $O(10^{-8})$ within about 120 seconds.

We remark that the sparsity technique in [28] is too expensive to be implementable for minimizing these random polynomials generated in the way as above because of computer memory limitations. For these random polynomials, the sparsity technique using chordal extension is almost as expensive as the general dense SOS relaxation. This is because the chordal extensions of csp graphs of these polynomials are usually much more dense than the original csp graphs. However, as we have seen in the above, the SOS relaxation (3.1)–(3.3) is very suitable for these polynomials.

4.2.2. Random sparse polynomials with given chordal extension. In this subsubsection, we generate random sparse polynomials in a similar way as in [28], and compare the performance of our sparse SOS relaxation (3.1)–(3.3) with the one in [28] using chordal extension. Generate a chordal graph randomly such that the size of every maximal clique is at most 6. Let $\{C_1, \ldots, C_m\}$ be the set of maximal cliques. If we choose $\Delta_i = C_i$, then the sparse SOS relaxation (3.1)–(3.3) is the same as the one using chordal extension. Therefore, to make a reasonable comparison, for each C_i, we choose a random subset $\Delta_i \subseteq C_i$. Choose each small polynomial f_i to have the form

$$f_i(x_{\Delta_i}) = \mathbf{m}_d(x_{\Delta_i})^T \cdot A_i \cdot \mathbf{m}_d(x_{\Delta_i}) + b_i^T \mathbf{m}_{2d-1}(x_{\Delta_i}).$$

Here $N_i = \binom{|\Delta_i|+d}{d}$, $A_i = nI_{N_i} + BB^T$, $B \in \mathbb{R}^{N_i \times N_i}$, and $b_i \in \mathbb{R}^{\binom{|\Delta_i|+d-1}{d-1}}$ are random. The global minimizers of $f(x) = \sum_i f_i(x_{\Delta_i})$ generated as above always exist and are contained in some compact set.

For polynomials randomly generated as above, the technique in [28] using chordal extension is a good choice, because there exists one sparse chordal extension of the csp graph. Now we compare the computational results for these two methods.

First, let $2d = 4$ and n be $20, 40, 60, 80, 100$. For each n, generate 50 random polynomials as above. For each of these random polynomials, solve the relaxation (3.1)–(3.3), find a chordal extension of the csp graph of $f(x)$, and then apply the sparse relaxation in [28]. Both relaxations are solved by the software *SeDuMi* using the *YALMIP* interface. Then we extract minimizers $\hat{x}$ from moment matrices. The computational results are in Table 5. For these solved problems, we just record the maximum error of the relaxation. Second, let $n = 30$ and $2d$ be $4, 6, 8$. For each $2d$, generate 50 random polynomials as above. For each one, solve the problem by the relaxation (3.1)–(3.8) and the one in [28] using chordal extension. They are solved by

TABLE 5
Comparison with chordal extension on quartic polynomials.

	Relaxation (3.1)–(3.3)				Relaxation using chordal extension			
	CPU seconds			accu	CPU seconds			accu
n	max	avr.	min	max	max	avr.	min	max
20	1.75	1.21	0.96	6.8e − 9	2.15	1.78	1.43	5.5e − 9
40	3.07	2.69	2.24	7.5e − 9	4.08	3.51	3.12	4.9e − 9
60	4.99	4.54	3.82	6.7e − 9	7.88	6.93	5.65	6.4e − 9
80	6.59	5.87	5.23	6.3e − 9	10.84	9.57	8.59	5.7e − 9
100	9.34	7.64	7.11	7.2e − 9	13.45	12.76	11.74	4.3e − 9

TABLE 6
Comparison with chordal extension on polynomials with 30 *variables.*

	Relaxation (3.1)–(3.3)				Relaxation using chordal extension			
	CPU seconds			accu	CPU seconds			accu
$2d$	max	avr.	min	max	max	avr.	min	max
4	2.87	1.98	1.35	7.2e − 9	3.06	2.21	1.69	4.3e − 9
6	22.61	16.78	10.53	6.9e − 9	32.15	23.91	13.51	5.1e − 9
8	193.45	131.17	98.75	6.7e − 9	253.79	186.84	112.37	5.8e − 9

the software *SeDuMi* using the *YALMIP* interface. The computational results are in Table 6.

From Tables 5 and 6, we observe that for polynomials randomly generated as above the sparse SOS relaxation (3.1)–(3.3) is slightly more computationally efficient than the one using chordal extension. As we can see, for these random polynomials, there is not much difference between the qualities of these two kinds of sparse SOS relaxations. The distinction between their qualities depends on specific problems. Of course, theoretically the sparse relaxation using chordal extension in [28] is at least as tight as the relaxation (3.1)–(3.3).

4.2.3. Random dense polynomials. In this subsubsection, we test the performance of our sparse SOS relaxation on minimizing general dense polynomials. We observe that every polynomial $f(x)$ is a summation of monomials whose number of variables is at most the degree $\deg(f)$. So the sparse SOS relaxation (3.1)–(3.3) is attractive when the degree $2d$ is small like 4. We generate the random dense polynomials as follows:

$$f(x) = \mathbf{m}_d(x)^T \cdot A \cdot \mathbf{m}_d(x) + b^T \mathbf{m}_{2d-1}(x).$$

Here $N = \binom{n+d}{d}$, $A = nI_N + BB^T$, $B \in \mathbb{R}^{N\times N}$ is a random matrix, and $b \in \mathbb{R}^{\binom{n+2d-1}{n}}$ is a random vector. So the global minimizers of $f(x)$ generated this way are contained in some compact set. Note that $f(x)$ is also a summation of small polynomials. Let Δ_i be the subsets of $[n]$ with cardinality $2d$. Then we can write

$$f(x) = \sum_{i=1}^{\binom{n}{2d}} f_i(x_{\Delta_i})$$

for some small polynomials $f_i(x_{\Delta_i})$.

Since $\|\Delta\| = 2d$, which should not be big for the effectiveness of the sparse relaxation (3.1)–(3.3), we test for the case that $2d = 4$. Let n be $16, 17, 18, 19, 20, 21, 22, 23$. For each pair (n, d) of these values, generate 50 random examples as above. For each

TABLE 7
Computational results for dense quartic polynomials.

n—	16	17	18	19	20	21	22	23
max time	335.29	569.74	901.32	1505.45	2249.19	3257.86	4734.25	7060.72
avr. time	241.48	455.32	751.69	1245.22	2070.70	2989.45	4497.84	6419.53
min time	205.60	397.11	688.58	1052.70	1893.02	2676.62	4197.95	5874.28
accuracy	7.3e − 9	6.7e − 9	7.4e − 9	6.9e − 9	8.1e − 9	6.5e − 9	7.9e − 9	8.5e − 9

random polynomial, solve the sparse relaxation (3.1)–(3.3) by the software *SeDuMi* using the *YALMIP* interface. The consumed CPU time and the accuracy of relaxation are in Table 7. We can see that the obtained solutions are very good within reasonably acceptable time. When $n \geq 24$, the sparse relaxation (3.1)–(3.3) is then also too expensive to be implementable due to computer memory restrictions.

For these randomly generated dense polynomials, the general dense SOS relaxation and sparse SOS relaxations like in [28] are not implementable for $n \geq 16$, due to either computer memory shortage or unacceptable long running time. However, when $2d$ is small like 4, the sparse SOS relaxation (3.1)–(3.3) can solve bigger dense polynomial optimization problems which can not be solved by other methods.

5. Applications. Minimizing a summation of small polynomials arises in various applications. Many big polynomials in applications often come in this form. In such situations, the sparse SOS relaxation (3.1)–(3.3) or (3.11)–(3.13) is very useful. In this section, we show some applications in solving sparse polynomial systems and sensor network localization.

5.1. Solving sparse polynomial system. Suppose we are trying to solve the sparse polynomial system

$$g_1(x_{\Delta_1}) = 0,\, g_2(x_{\Delta_2}) = 0,\, \ldots,\, g_m(x_{\Delta_m}) = 0.$$

In some applications, these equations are redundant or even inconsistent. When the polynomial system does not admit a solution, we want to seek a least squares solution, which is often useful in applications.

This problem can be formulated as finding the global minimizer of the sparse polynomial

$$f^* := \min_{x\in\mathbb{R}^n} f(x) = \sum_{i=1}^{m} g_i^2(x_{\Delta_i}).$$

The polynomial system has a real zero if and only if $f^* = 0$. When $f^* = 0$, the global minimizers are precisely the real zeros of the polynomial system. When $f^* > 0$, the global minimizers are the least squares solutions.

One important sparse polynomial system of the above form is from computing the numerical solutions of nonlinear differential equations. Consider the two-point BVP

$$F(t, x, x', x'') = 0,\;\; x(a) = \alpha,\, x(b) = \beta,$$

where $F(t, x, x', x'')$ is a polynomial function in t, x, x', x''. To find the numerical solution, the central difference approximation with a uniform mesh is often used to discretize the derivatives. Let N be a positive integer and set $h = \frac{b-a}{N+1}$. Then we get

TABLE 8
The performance of (3.11)–(3.13) *solving the equations in Example* 5.1.

N	Eqn. error	$\|x_k - x(t_k)\|_\infty$	$\|x_k - x(t_k)\|_\infty / h^2$	Time
5	2.8937e − 07	7.0252e − 05	2.5291e − 003	0.52
10	2.3329e − 07	1.9570e − 05	2.3680e − 003	0.77
20	5.2879e − 07	1.5041e − 05	6.6331e − 003	1.18
30	2.6194e − 07	1.9413e − 05	1.8656e − 002	2.09
40	3.0304e − 07	4.3344e − 05	7.2861e − 002	3.99
50	6.5375e − 07	1.5124e − 04	3.9338e − 001	6.82
60	1.5271e − 06	4.8695e − 04	1.8119e + 00	7.77
70	1.2555e − 06	5.2428e − 04	2.6429e + 00	9.16
80	9.7315e − 07	6.1330e − 04	4.0239e + 00	9.78
90	2.7519e − 06	1.9311e − 03	1.5991e + 01	10.81
100	1.8628e − 06	8.1425e − 04	8.3062e + 00	8.79

polynomial difference equations

$$F\left(t_k, x_k, \frac{x_{k+1} - x_{k-1}}{2h}, \frac{x_{k-1} - 2x_k + x_{k+1}}{h^2}\right) = 0,\ k = 1, \ldots, N,$$

where $x_0 = \alpha$, $x_{N+1} = \beta$, and $t_k = a + hk$. Every polynomial on the left involves 2 or 3 variables x_{k-1}, x_k, x_{k+1}. So this is a sparse polynomial system. There are several methods for solving this kind of polynomial system, like Newton's method and homotopy methods. Newton's method is very fast, but often require an accurate initial guess. Homotopy methods do not require a "satisfactory" guess and work well for small N, but are expensive to implement for large N. We refer to [1] and the references therein for work in this area. When N is large, this polynomial system is large but sparse. We solve this system by applying the sparse SOS relaxation (3.11)–(3.13) for big N (up to 100 or even bigger).

Example 5.1 ([1]). Consider a basic BVP

$$x'' - 2x^3 = 0,\ \ x(0) = \frac{1}{2},\ x(1) = \frac{1}{3}.$$

The exact solution to this problem is $x(t) = \frac{1}{t+2}$. Now we discretize the differential equation with mesh size $h = \frac{1}{N+1}$, then get the difference equation

$$\begin{aligned} \frac{1}{2} - 2x_1 + x_2 - 2h^2 x_1^3 &= 0, \\ x_{k-1} - 2x_k + x_{k+1} - 2h^2 x_k^3 &= 0,\ k = 2, \ldots, N-1 \\ x_{N-1} - 2x_N + \frac{1}{3} - 2h^2 x_N^3 &= 0. \end{aligned}$$

This is a polynomial system about $x_1, x_2, \ldots, x_N$. We can solve this polynomial system as a nonlinear least squares problem by applying sparse SOS relaxation (3.11)–(3.13). The computational results are in Table 8. The equation error is defined to be the infinity norm of the residuals of the left-hand side of the polynomial system, which measures the quality of how the polynomial systems are solved. The obtained solutions have equation error from $\mathcal{O}(10^{-6})$ to $\mathcal{O}(10^{-7})$. If we want to make them more accurate, they can be used as the initial guesses in Newton's methods for refining. The accuracy of the discretization is defined to be the difference between computed solution x_k and true solution $x(t_k) = \frac{1}{2+t_k}$ where $t_k = \frac{k}{N+1}$. Since the discretization

has error $\mathcal{O}(h^2)$, we expect that $\|x_k - x(t_k)\|_\infty/h^2$ is a constant. When $N \le 40$, we can see that $\|x_k - x(t_k)\|_\infty/h^2$ is almost constant. When $N \ge 50$, *SeDuMi* experienced numerical troubles, and the returned solutions are not as accurate as for the smaller Ns. This explains why $\|x_k - x(t_k)\|_\infty$ and $\|x_k - x(t_k)\|_\infty/h^2$ becomes bigger for $N \ge 50$. Time records the CPU seconds consumed by the SDP solver *SeDuMi*. For $N = 100$, the computation takes less CPU time than for $N = 80$ or $N = 90$. This is because the numerical troubles make *SeDuMi* terminate earlier.

Example 5.2. Consider another BVP

$$x'' + \frac{1}{2}(x+t)^3 = 0, \;\; x(0) = 0, \, x(1) = 0.$$

Now we discretize the differential equation with mesh size $h = \frac{1}{N+1}$ and get the difference equation

$$\begin{aligned} 2x_1 - x_2 + \frac{1}{2}h^2(x_1 + t_1)^3 &= 0, \\ 2x_i - x_{i-1} - x_{i+1} + \frac{1}{2}h^2(x_i + t_i)^3 &= 0, \, i = 2, \dots, N-1, \\ 2x_N - x_{N-1} + \frac{1}{2}h^2(x_N + t_N)^3 &= 0. \end{aligned}$$

This is a polynomial system about $x_1, x_2, \dots, x_N$. We can solve this polynomial system as a nonlinear least squares problem by applying sparse SOS relaxation (3.11)–(3.13). When $N = 30$, we get the following real solution within about 2.5 CPU seconds (only the first four digits are shown):

$$\begin{aligned} &(-0.0159, \, -0.0312, \, -0.0459, \, -0.0600, \, -0.0735, \, -0.0864, \, -0.0985, \, -0.1099, \, -0.1205, \, -0.1302, \\ &\, -0.1391, \, -0.1470, \, -0.1540, \, -0.1599, \, -0.1646, \, -0.1682, \, -0.1705, \, -0.1715, \, -0.1710, \, -0.1689, \\ &\, -0.1651, \, -0.1596, \, -0.1521, \, -0.1425, \, -0.1307, \, -0.1164, \, -0.0995, \, -0.0796, \, -0.0567, \, -0.0302). \end{aligned}$$

For solving polynomial systems arising from BVPs, the sparse SOS method based on chordal extension like in [28] and (3.1)–(3.3) have similar performance, because the csp matrices are banded and the Cholesky factors are sparse. But they are much more expensive than the further sparse SOS relaxation (3.11)–(3.13). This is because (3.11)–(3.13) has further used the sparsity of each small polynomial $g_i(x_{\Delta_i})$ resulting from BVPs.

5.2. Sensor network localization. The *sensor network location* problem is basically described as follows: find a sequence of unknown vectors $x_1, x_2, \dots, x_n \in \mathbb{R}^k (k = 1, 2, \dots)$ (they are called *sensors*) such that distances between these sensors and some other known vectors $a_1, \dots, a_m$ (they are called *anchors*) are equal to some given numbers. Now each x_i itself is a k-dimensional vector. To be more specific, let $\mathcal{A} = \{(i,j) \in [n] \times [n] : \, i < j, \, \|x_i - x_j\|_2 = d_{ij}\}$, and $\mathcal{B} = \{(i,k) \in [n] \times [m] : \, \|x_i - a_k\|_2 = e_{ik}\}$, where d_{ij}, e_{ik} are given distances. Then the sensor network localization problem is to find vectors $x_1, x_2, \dots, x_n$ such that $\|x_i - x_j\|_2 = d_{ij}$ for all $(i,j) \in \mathcal{A}$ and $\|x_i - a_k\|_2 = e_{ik}$ for all $(i,k) \in \mathcal{B}$. Notice that $\mathcal{A}$ and $\mathcal{B}$ give only some partial pairs of distances. $\mathcal{A}$ does not contain all the pairs (i,j) such that $i < j$, and neither does $\mathcal{B}$.

Sensor network localization is also known as the *graph realization problem* or the *distance geometry problem*. Given a graph $G = (V, E)$ along with a real number associated with each edge, graph realization is to assign each vertex a coordinate so that the Euclidean distance between any two adjacent vertices is equal to the real number associated with that edge.

The locations of sensors can be determined from the polynomial system

$$\|x_i - x_j\|_2^2 = d_{ij}^2, \forall\, (i,j) \in \mathcal{A},$$
$$\|x_i - a_k\|_2^2 = e_{ik}^2, \forall\, (i,k) \in \mathcal{B}.$$

Usually solving this polynomial system directly is very expensive. Here we solve this polynomial system as a nonlinear least squares problem. Minimize the quartic polynomial function

$$f(x) := \sum_{(i,j)\in\mathcal{A}} \left(\|x_i - x_j\|_2^2 - d_{ij}^2\right)^2 + \sum_{(i,k)\in\mathcal{B}} \left(\|x_i - a_k\|_2^2 - e_{ik}^2\right)^2, \tag{5.1}$$

where $x = [\,x_1, \ldots, x_n]$. x^* is a solution to sensor network localization problem if and only if x^* is a global minimizer of $f(x)$ such that $f(x^*) = 0$. When x^* is a global minimizer such that $f(x^*) > 0$, the distances $d_{i,j}$ and e_{ik} are not consistent, and x^* is a solution in the least squares sense. This polynomial $f(x)$ is of the form (1.1), and our sparse SOS relaxation (3.1)–(3.3) can be applied to solve the problem.

We randomly generate test problems which are similar to those given in [5]. First, we randomly generate $n = 500$ sensor locations $x_1^*, \ldots, x_n^*$ from the unit square $[-0.5, 0.5] \times [-0.5, 0.5]$. The anchors $\{a_1, a_2, a_3, a_4\}\,(m = 4)$ are chosen to be four fixed points $(\pm 0.45, \pm 0.45)$. Choose edge set $\mathcal{A}$ such that for every sensor x_i^* there are at most 10 sensors $x_j^*\,(j > i)$ with $(i,j) \in \mathcal{A}$ and $\|x_i^* - x_j^*\|_2 \le 0.3$. For every $(i,j) \in \mathcal{A}$, compute the distance $\|x_i^* - x_j^*\|_2 = d_{ij}$. Choose edge $\mathcal{B}$ such that every anchor is connected to all the sensors within distance 0.3. For every $(i,k) \in \mathcal{B}$, compute the distance $\|x_i^* - a_k\|_2 = e_{ik}$. Then we apply sparse SOS relaxation (3.1)–(3.8) to minimize polynomial function (5.1). The accuracy of computed sensor locations $\hat{x}_1, \ldots, \hat{x}_n$ will be measured by the Root Mean Square Distance (RMSD), which is defined as $\text{RMSD} = \left(\frac{1}{n}\sum_{i=1}^n \|\hat{x}_i - x_i^*\|_2^2\right)^{\frac{1}{2}}$. We use *SeDuMi* to solve the sparse SOS relaxation (3.1)–(3.8) on a Linux machine with 1.46 GHz CPU and 0.98GB memory. The problem can be solved within about 18 CPU minutes with accuracy $\mathcal{O}(10^{-6})$.

The sparse SOS method based on chordal extension like in [28] is usually not practical for solving sensor network localization problems, because the corresponding csp matrices usually do not have sparse Cholesky factorizations and the resulting chordal extensions are often too dense to be useful. We refer to [21] for more details about sparse SOS methods for sensor network localization.

6. Conclusions and discussions. This paper proposes sparse SOS relaxations for minimizing polynomial functions that are summations of small polynomials. We discuss various properties of these relaxations and the computational issues. We also present applications of this sparsity technique in solving polynomial equations derived from nonlinear differential equations and sensor network localization. As a special case, this sparsity technique provides a heuristic approach to solve bigger dense polynomial optimization problems.

In order to exploit the sparsity, the polynomial and its SOS representation must be sparse. In many applications, the polynomials are often given with sparsity pattern (1.1), and then the sparsity technique proposed in this paper is very suitable. If the sparsity pattern is not given, one important future work is how to represent the polynomial in a sparse pattern such that the technique proposed in this paper is most efficient. Of course, one simple choice is to consider each monomial as a small polynomial.

The idea of this sparse SOS relaxation can be applied in a similar way to solve constrained polynomial optimization problems, provided the objective and constraint polynomials are also sums of small polynomials. See Kim et al. [11] and Lasserre [15] for related work. To get the global minimum, high order relaxations are usually necessary. Lasserre [15] proved the convergence under the running intersection property. However, unlike the general dense SOS relaxation for minimizing polynomials over compact sets, the convergence might fail when the running intersection property does not hold. As a counterexample, consider the *Minimum Cover Set Problem*. Let $G = (V, E)$ be a graph with vertex set $V = [3]$ and edge set $E = \{(1,2), (1,3), (2,3)\}$. To find the minimum cover set is equivalent to solving

$$\begin{aligned} \min_{x \in \mathbb{R}^3} \quad & f_1(x_{\Delta_1}) + f_2(x_{\Delta_2}) + f_3(x_{\Delta_3}) \\ s.t. \quad & x_1^2 = x_1,\, x_2^2 = x_2,\, x_3^2 = x_3, \\ & x_1 + x_2 \geq 1,\, x_1 + x_3 \geq 1, x_2 + x_3 \geq 1, \end{aligned}$$

where $\Delta_1 = \{1,2\}$, $\Delta_2 = \{1,3\}$, $\Delta_3 = \{2,3\}$ and $f_1(x_{\Delta_1}) = \frac{1}{2}(x_1 + x_2)$, $f_2(x_{\Delta_2}) = \frac{1}{2}(x_1 + x_3)$, $f_3(x_{\Delta_3}) = \frac{1}{2}(x_2 + x_3)$. The running intersection property now fails. However, it can be shown that the global minimum $f^* = 2$ and the lower bounds given by sparse SOS relaxations are at most $\frac{3}{2}$. The sparse SOS relaxations do not converge for this example.

Another important future work is to apply the sparse SOS relaxations in solving big real sparse polynomial systems arising from nonlinear differential equations.

Acknowledgments. The authors wish to thank Bernd Sturmfels and the referees for helpful suggestions to improve this paper.

REFERENCES

[1] A.L. Allgower, D.J. Bates, A.J. Sommese, and C.W. Wampler, *Solution of polynomial systems derived from differential equations*, Computing, 76 (2005), pp, 1–10.

[2] S.J. Benson and Y. Ye, *DSDP3: Dual scaling algorithm for general positive semidefinite programming*, Technical report ANL/MCS-P851-1000, Mathematics and Computer Science Division, Argonne National Laboratory, 2001.

[3] J. Blair and B. Peyton, *An introduction to chordal graphs and clique trees*, in Graph Theory and Sparse Matrix Computations, J. George, J. Gilbert, and J. Liu, eds., Springer Verlag, New York, 1993, pp. 1–30.

[4] G. Blekherman, *There are significantly more nonnegative polynomials than sums of squares*, Israel J. Math., 153 (2006), pp. 355–380.

[5] P. Biswas, T.C. Liang, K.C. Toh, T.C. Wang, and Y. Ye, *Semidefinite programming approaches for sensor network localization with noisy distance measurements*, IEEE Trans. Automat. Sci. Eng., 3 (2006), pp. 360–371.

[6] A.R. Conn, N.I.M. Gould, and P.L. Toint, *Testing a class of methods for solving minimization problems with simple bounds on the variables*, Math. Comp., 50 (1988), pp. 399–430.

[7] R.E. Curto and L.A. Fialkow, *The truncated complex K-moment problem*, Trans. Amer. Math. Soc., 352 (2000), pp. 2825–2855.

[8] K. Gatermann and P. Parrilo, *Symmetry groups, semidefinite programs, and sums of squares*, J. Pure Appl. Algebra, 192 (2004), pp. 95–128.

[9] D. Henrion and J. Lasserre, *Detecting global optimality and extracting solutions in GloptiPoly*, in Positive Polynomials in Control, D. Henrion and A. Garulli, eds., Lecture Notes on Control and Information Sciences, Springer Verlag, New York, 2005.

[10] D. Jibetean and M. Laurent, *Semidefinite approximations for global unconstrained polynomial optimization*, SIAM J. Optim., 16 (2005), pp. 490–514.

[11] S. Kim, M. Kojima, and H. Waki, *Generalized Lagrangian duals and sums of squares relaxations of sparse polynomial optimization problems*, SIAM J. Optim., 15 (2005), pp. 697–719.

[12] M. Kojima, S. Kim, and H. Waki, *Sparsity in sums of squares of polynomials*, Math. Program., 103 (2005), pp. 45–62.
[13] J. Lasserre, *Global optimization with polynomials and the problem of moments*, SIAM J. Optim., 11 (2001), pp. 796–817.
[14] J. Lasserre, *A sum of squares approximation of nonnegative polynomials*, SIAM J. Optim., 16 (2006), pp. 751–765.
[15] J. Lasserre, *Convergent SDP-relaxations in polynomial optimization with sparsity*, SIAM J. Optim., 17 (2006), pp. 822–843.
[16] M. Laurent, *Moment matrices and optimization over polynomials - A survey on selected topics*, Emerging Applications of Algebraic Geometry, IMA Volumes in Mathematics and its Applications 149, (des. M. Putinar and S. Sullivant), Springer, to appear.
[17] J. Löfberg, *YALMIP: A toolbox for Modeling and Optimization in MATLAB*, in Proceedings of the CACSD Conference, Taipei, Taiwan, 2004. http://control.ee.ethz.ch/~joloef/yalmip.php.
[18] J. More, B. Garbow, and K. Hillstrom, *Testing unconstrained optimization software*, ACM Trans. Math. Software, 7 (1981), pp. 17–41.
[19] S.G. Nash, *Newton-type minimization via the Lanczos method*, SIAM J. Numer. Anal., 21 (1984), pp. 770–788.
[20] J. Nie, J. Demmel, and B. Sturmfels, *Minimizing polynomials via sum of squares over the gradient ideal*, Math. Program., Ser. A, 106 (2006), pp. 587–606.
[21] J. Nie, *Sum of squares methods for sensor network localization*, Comput. Optim. Appl., to appear.
[22] P. Parrilo and B. Sturmfels, *Minimizing polynomial functions*, in Proceedings of the DIMACS Workshop on Algorithmic and Quantitative Aspects of Real Algebraic Geometry in Mathematics and Computer Science 2001, S. Basu and L. Gonzalez-Vega, eds., American Mathematical Society, Providence, RI, 2003, pp. 83–100.
[23] P. Parrilo, *Semidefinite programming relaxations for semialgebraic problems*, Math. Program. Ser. B, 96 (2003), pp. 293–320.
[24] P. Parrilo, *Exploiting structure in sum of squares programs*, in Proceedings for the 42nd IEEE Conference on Decision and Control, Maui, Hawaii, 2003.
[25] B. Reznick, *Extremal psd forms with few terms*, Duke Math. J., 45 (1978), pp. 363–374.
[26] M. Schweighofer, *Global optimization of polynomials using gradient tentacles and sums of squares*, SIAM J. Optim., 17 (2006), pp. 920–942.
[27] J.F. Sturm, *SeDuMi* 1.02*, a MATLAB toolbox for optimization over symmetric cones*, Optim. Methods Softw., 11&12 (1999), pp. 625–653.
[28] H. Waki, S. Kim, M. Kojima, and M. Muramatsu, *Sums of squares and semidefinite programming relaxations for polynomial optimization problems with structured sparsity*, SIAM J. Optim., 17 (2006), pp. 218–242.
[29] H. Wolkowicz, R. Saigal, and L. Vandenberghe, eds., *Handbook of semidefinite programming*, Kluwer Publishers, Boston, 2000.

SIAM J. OPTIM.
Vol. 19, No. 4, pp. 1559–1573

ON SEMIDEFINITE PROGRAMMING RELAXATIONS OF THE TRAVELING SALESMAN PROBLEM*

ETIENNE DE KLERK†, DMITRII V. PASECHNIK‡, AND RENATA SOTIROV†

Abstract. We consider a new semidefinite programming (SDP) relaxation of the symmetric traveling salesman problem (TSP) that may be obtained via an SDP relaxation of the more general quadratic assignment problem (QAP). We show that the new relaxation dominates the one in [D. Cvetković, M. Cangalović, and V. Kovačević-Vujčić, *Semidefinite programming methods for the symmetric traveling salesman problem*, in Proceedings of the 7th International IPCO Conference on Integer Programming and Combinatorial Optimization, Springer-Verlag, London, UK, 1999, pp. 126–136]. Unlike the bound of Cvetković et al., the new SDP bound is not dominated by the Held–Karp linear programming bound, or vice versa.

Key words. traveling salesman problem, semidefinite programming, quadratic assignment problem, association schemes

AMS subject classifications. 90C22, 20Cxx, 70-08

DOI. 10.1137/070711141

1. Introduction.

The quadratic assignment problem (QAP) may be stated in the following form:

$$\min_{X \in \Pi_n} trace\left(AXBX^T\right), \tag{1}$$

where A and B are given symmetric $n \times n$ matrices, and Π_n is the set of $n \times n$ permutation matrices.

It is well known that the QAP contains the symmetric traveling salesman problem (TSP) as a special case. To show this, we denote the complete graph on n vertices with edge lengths (weights) $D_{ij} = D_{ji} > 0$ $(i \neq j)$, by $K_n(D)$, where D is called the matrix of edge lengths (weights). The TSP is to find a Hamiltonian circuit of minimum length in $K_n(D)$. The n vertices are often called *cities*, and the Hamiltonian circuit of minimum length the *optimal tour*.

To see that TSP is a special case of QAP, let C_1 denote the adjacency matrix of $\mathcal{C}_n$ (the standard circuit on n vertices):

$$C_1 := \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 1 \\ 1 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \\ 0 & & & & 0 & 1 \\ 1 & 0 & \cdots & 0 & 1 & 0 \end{bmatrix}.$$

Now the TSP problem is obtained from the QAP problem (1) by setting $A = \frac{1}{2}D$ and $B = C_1$. To see this, note that every Hamiltonian circuit in a complete graph has

*Received by the editors December 17, 2007; accepted for publication (in revised form) August 28, 2008; published electronically December 31, 2008.
http://www.siam.org/journals/siopt/19-4/71114.html
†Department of Econometrics and OR, Tilburg University, Tilberg 5000 LE, The Netherlands (e.deklerk@uvt.nl, r.sotirov@uvt.nl).
‡School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore (dima@ntu.edu.sg).

adjacency matrix XC_1X^T for some $X \in \Pi_n$. Thus we may concisely state the TSP as

$$TSP_{opt} := \min_{X \in \Pi_n} trace\left(\frac{1}{2} DXC_1X^T\right). \tag{2}$$

The symmetric TSP is NP-hard in the strong sense [20], and therefore so is the more general QAP. In the special case where the distance function of the TSP instance satisfies the triangle inequality (metric TSP), there is a celebrated 3/2-approximation algorithm due to Christofides [9]. It is a long-standing (since 1975) open problem to improve on the 3/2 constant, since the strongest negative result is that a $(1+1/219)$-approximation algorithm is not possible, unless P=NP [21].

In the case when the distances are Euclidean in fixed dimension (the so-called planar or geometric TSP), the problem allows a polynomial-time approximation scheme [1]. A recent survey of the TSP is given by Schrijver [22, Chapter 58].

Main results and outline of this paper. In this paper we will consider semidefinite programming (SDP) relaxations of the TSP. We will introduce a new SDP relaxation of TSP in section 2, which is motivated by the theory of association schemes. Subsequently, we will show in section 3 that the new SDP relaxation coincides with the SDP relaxation for QAP introduced in [26] when applied to the QAP reformulation of TSP in (2). Then we will show in section 4 that the new SDP relaxation dominates the relaxation due to Cvetković et al. [5]. The relaxation of Cvetković et al. is known to be dominated by the Held–Karp linear programming bound [6, 15], but we show in section 5 that the new SDP bound is not dominated by the Held–Karp bound (or vice versa).

Notation. The space of $p \times q$ real matrices is denoted by $\mathbb{R}^{p\times q}$, the space of $k \times k$ symmetric matrices is denoted by $\mathcal{S}_k$, and the space of $k \times k$ symmetric positive semidefinite matrices by $\mathcal{S}_k^+$. We will sometimes also use the notation $X \succeq 0$ instead of $X \in \mathcal{S}_k^+$, if the order of the matrix is clear from the context. By $\mathrm{diag}(X)$ we mean the n-vector composed of the diagonal entries of $X \in \mathcal{S}_n$.

We use I_n to denote the identity matrix of order n. Similarly, J_n and e_n denote the $n \times n$ all-ones matrix and all ones n-vector, respectively, and $0_{n\times n}$ is the zero matrix of order n. We will omit the subscript if the order is clear from the context.

The *Kronecker product* $A \otimes B$ of matrices $A \in \mathbb{R}^{p\times q}$ and $B \in \mathbb{R}^{r\times s}$ is defined as the $pr \times qs$ matrix composed of pq blocks of size $r \times s$, with block ij given by $A_{ij}B$ $(i = 1, \ldots, p)$, $(j = 1, \ldots, q)$.

The Hadamard (component-wise) product of matrices A and B of the same size will be denoted by $A \circ B$.

2. A new SDP relaxation of TSP. In this section we show that the optimal value of the following semidefinite program provides a lower bound on the length TSP_{opt} of an optimal tour:

$$\left.\begin{array}{rll} \min & \frac{1}{2} trace\left(DX^{(1)}\right) & \\ \text{subject to} & & \\ & X^{(k)} \geq 0, & k = 1, \ldots, d \\ & \sum_{k=1}^{d} X^{(k)} = J - I, & \\ & I + \sum_{k=1}^{d} \cos\left(\frac{2\pi i k}{n}\right) X^{(k)} \succeq 0, & i = 1, \ldots, d \\ & X^{(k)} \in \mathcal{S}^n, & k = 1, \ldots, d, \end{array}\right\}, \tag{3}$$

where $d = \lfloor \frac{1}{2} n \rfloor$ is the diameter of $\mathcal{C}_n$.

Note that this problem involves nonnegative matrix variables $X^{(1)}, \ldots, X^{(d)}$ of order n. The matrix variables $X^{(k)}$ have an interesting interpretation in terms of *association schemes*.

Association schemes. We will give a brief overview of this topic; for an introduction to association schemes, see Chapter 12 in [10], and in the context of SDP, [11].

DEFINITION 2.1 (Asssociation scheme). *Assume that a given set of $n \times n$ matrices $B_0, \ldots, B_t$ has the following properties:*

(1) *B_i is a $0-1$ matrix for all i and $B_0 = I$;*
(2) *$\sum_i B_i = J$;*
(3) *$B_i = B_{i^*}^T$ for some i^*;*
(4) *$B_i B_j = B_j B_i$ for all i, j;*
(5) *$B_i B_j \in span\{B_1, \ldots, B_t\}$.*

Then we refer to $\{B_1, \ldots, B_t\}$ as an association scheme. If the B_i's are also symmetric, then we speak of a symmetric association scheme.

Note that item (4) (commutativity) implies that the matrices $B_1, \ldots, B_t$ share a common set of eigenvectors, and therefore can be simultaneously diagonalized. Note also that an association scheme is a basis of a matrix-$*$ algebra (viewed as a vector space). Moreover, one clearly has

$$trace\left(B_i B_j^T\right) = 0 \text{ if } i \neq j.$$

Since the B_i's share a system of eigenvectors, there is a natural ordering of their eigenvalues with respect to any fixed ordering of the eigenvectors. Thus the last equality may be interpreted as

$$\sum_k \lambda_k(B_i)\lambda_k(B_j) = 0 \text{ if } i \neq j, \tag{4}$$

where the $\lambda_k(B_i)$'s are the eigenvalues of B_i with respect to the fixed ordering.

The association scheme of particular interest to us arises as follows. Given a connected graph $G = (V, E)$ with diameter d, we define $|V| \times |V|$ matrices $A^{(k)}$ $(k = 1, \ldots, d)$ as follows:

$$A_{ij}^{(k)} = \begin{cases} 1 & \text{if } \operatorname{dist}(i,j) = k \\ 0 & \text{else,} \end{cases} \qquad (i, j \in V),$$

where $\operatorname{dist}(i, j)$ is the length of the shortest path from i to j.

Note that $A^{(1)}$ is simply the adjacency matrix of G. Moreover, one clearly has

$$I + \sum_{k=1}^{d} A^{(k)} = J.$$

It is well known that, for $G = \mathcal{C}_n$, the matrices $A^{(k)}$ $(k = 1, \ldots, d \equiv \lfloor n/2 \rfloor)$ together with $A^{(0)} := I$ form an association scheme, since $\mathcal{C}_n$ is a distance regular graph.

It is shown in the Appendix to this paper that for $G = \mathcal{C}_n$, the eigenvalues of the matrix $A^{(k)}$ are

$$\lambda_m\left(A^{(k)}\right) = 2\cos(2\pi mk/n), \quad m = 0, \ldots, n-1,\ k = 1, \ldots, \lfloor (n-1)/2 \rfloor,$$

and, if n is even,

$$\lambda_{n/2}\left(A^{(k)}\right) = \cos(k\pi) = (-1)^k.$$

In particular, we have

$$\lambda_m\big(A^{(k)}\big) = \lambda_k\big(A^{(m)}\big) \qquad k,m = 1,\ldots,\lfloor (n-1)/2 \rfloor. \tag{5}$$

Also note that

$$\lambda_m\big(A^{(k)}\big) = \lambda_{n-m}\big(A^{(k)}\big), \quad k,m = 1,\ldots,\lfloor (n-1)/2 \rfloor, \tag{6}$$

so that each matrix $A^{(k)}$ $(k = 1,\ldots,d)$ has only $1 + \lfloor n/2 \rfloor$ distinct eigenvalues.

Verifying the SDP relaxation (3). We now show that setting $X^{(k)} = A^{(k)}$ $(k = 1,\ldots,d)$ gives a feasible solution of (3). We only need to verify that

$$I + \sum_{k=1}^{d} \cos\left(\frac{2\pi i k}{n}\right) A^{(k)} \succeq 0, \qquad i = 1,\ldots,d.$$

We will show this for odd n, the proof for even n being similar.

Since the $A^{(k)}$'s may be simultaneously diagonalized, the last linear matrix inequality (LMI) is the same as

$$2 + \sum_{k=1}^{d} \lambda_k\big(A^{(i)}\big)\lambda_j\big(A^{(k)}\big) \geq 0, \qquad i,j = 1,\ldots,d,$$

and by using (5) this becomes

$$2 + \sum_{k=1}^{d} \lambda_k\big(A^{(i)}\big)\lambda_k\big(A^{(j)}\big) \geq 0, \qquad i,j = 1,\ldots,d.$$

Since $\lambda_0(A^{(i)}) = 2$ $(i = 1,\ldots,d)$, and using (4), one can easily verify that the last inequality holds. Indeed, one has

$$\begin{aligned}
&2 + \sum_{k=1}^{d} \lambda_k\big(A^{(i)}\big)\lambda_k\big(A^{(j)}\big) \\
&= 2 + \frac{1}{2}\sum_{k=1}^{n-1} \lambda_k\big(A^{(i)}\big)\lambda_k\big(A^{(j)}\big) \quad \text{(by (6))} \\
&= 2 - \frac{1}{2}\lambda_0\big(A^{(i)}\big)\lambda_0\big(A^{(j)}\big) + \frac{1}{2}\sum_{k=0}^{n-1} \lambda_k\big(A^{(i)}\big)\lambda_k\big(A^{(j)}\big) \\
&= \begin{cases} 2 - 2 + 0 = 0 & \text{if } (i \neq j), \text{ by (4)}, \\ 2 - 2 + \frac{1}{2}\sum_{k=0}^{n-1} \big(\lambda_k\big(A^{(i)}\big)\big)^2 \geq 0 & \text{if } (i = j). \end{cases}
\end{aligned}$$

Thus we have established the following result.

THEOREM 2.1. *The optimal value of the SDP problem* (3) *provides a lower bound on the optimal value* TSP_{opt} *of the associated TSP instance.*

3. Relation of (3) to an SDP relaxation of QAP. An SDP relaxation of the QAP problem (1) was introduced in [26], and further studied for specially structured instances in [7].

When applied to the QAP reformulation of TSP in (2), this SDP relaxation takes the form:

$$\left.\begin{array}{rl} \min & \frac{1}{2}trace(C_1 \otimes D)Y \\ \text{subject to} & \\ & trace(((I \otimes (J-I))Y + ((J-I) \otimes I)Y = 0 \\ & trace(Y) - 2e^T y = -n \\ & \begin{pmatrix} 1 & y^T \\ y & Y \end{pmatrix} \succeq 0, \;\; Y \geq 0. \end{array}\right\}. \tag{7}$$

It is easy to verify that this is indeed a relaxation of problem (2), by noting that setting $Y = \mathrm{vec}(X)\mathrm{vec}(X)^T$ and $y = \mathrm{diag}(Y)$ gives a feasible solution if $X \in \Pi_n$.

In this section we will show that the optimal value of the SDP problem (7) actually equals the optimal value of the new SDP relaxation (3). The proof is via the technique of *symmetry reduction*.

Symmetry reduction of the SDP problem (7). Consider the following form of a general SDP problem:

$$p^* := \min_{X \succeq 0, X \geq 0} \{trace(A_0 X) \; : \; trace(A_k X) = b_k, \quad k = 1, \ldots, m\}, \tag{8}$$

where the A_i $(i = 0, \ldots, m)$ are given symmetric matrices.

If we view (7) as an SDP problem in the form (8), the data matrices of problem (7) are

$$\begin{pmatrix} 0 & 0^T \\ 0 & \frac{1}{2}C_1 \otimes D \end{pmatrix}, \begin{pmatrix} 0 & 0^T \\ 0 & I \otimes (J-I) + (J-I) \otimes I, \end{pmatrix}, \begin{pmatrix} 0 & -e^T \\ -e & 2I \end{pmatrix}, \begin{pmatrix} 1 & 0^T \\ 0 & 0 \end{pmatrix}. \tag{9}$$

Definition 3.1. *We define the* automorphism group *of a matrix* $Z \in \mathbb{R}^{k \times k}$ *as*

$$aut(Z) = \{P \in \Pi_k \; : \; PZP^T = Z\}.$$

Symmetry reduction of problem (8) is possible under the assumption that the multiplicative matrix group

$$\mathcal{G} := \bigcap_{i=0}^{m} \mathrm{aut}(A_i)$$

is nontrivial. We call $\mathcal{G}$ the symmetry group of the SDP problem (8).

For the matrices (9), the group $\mathcal{G}$ is given by the matrices

$$\mathcal{G} := \left\{ \begin{pmatrix} 1 & 0^T \\ 0 & P \otimes I \end{pmatrix} \; : \; P \in \mathcal{D}_n \right\}, \tag{10}$$

where $\mathcal{D}_n$ is the (permutation matrix representation of) dihedral group of order n, i.e., the automorphism group of $\mathcal{C}_n$.

The basic idea of symmetry reduction is given by the following result.

Theorem 3.1 (see, e.g., [8]). *If X is a feasible (resp. optimal) solution of the SDP problem* (8) *with symmetry group $\mathcal{G}$, then*

$$\bar{X} := \frac{1}{|\mathcal{G}|} \sum_{P \in \mathcal{G}} P^T X P$$

is also a feasible (resp. optimal) solution of (8).

Thus there exist optimal solutions in the set

$$\mathcal{A}_{\mathcal{G}} := \left\{ \frac{1}{|\mathcal{G}|} \sum_{P \in \mathcal{G}} P^T X P \ : \ X \in \mathbb{R}^{n \times n} \right\}.$$

This set is called the centralizer ring (or commutant) of $\mathcal{G}$ and it is a matrix $*$-algebra. For the group defined in (10), it is straightforward to verify that the centralizer ring is given by

$$\mathcal{A}_{\mathcal{G}} := \left\{ \begin{pmatrix} \alpha & x^T \\ y & C \otimes Z \end{pmatrix} \ \middle| \ \alpha \in \mathbb{R},\ C = C^T \text{ circulant},\ Z \in \mathbb{R}^{n \times n},\ x, y \in \mathbb{R}^{n^2} \right\}, \tag{11}$$

where $x^T = [x_1 e^T \dots x_n e^T]$ and $y^T = [y_1 e^T \dots y_n e^T]$ for some scalars x_i and y_i $(i = 1, \dots, n)$, where $e \in \mathbb{R}^n$ is the all-ones vector, as before.

Thus we may restrict the feasible set of problem (7) to feasible solutions of the form (11).

If we divide y and Y in (7) into blocks

$$y = \left(\left(y^{(1)}\right)^T \cdots \left(y^{(n)}\right)^T \right)^T,$$

and

$$Y = \begin{pmatrix} Y^{(11)} & \dots & Y^{(1n)} \\ \vdots & \ddots & \vdots \\ Y^{(n1)} & \dots & Y^{(nn)} \end{pmatrix},$$

where $y^{(i)} \in \mathbb{R}^n$ and $Y^{(ij)} = Y^{(ji)T} \in \mathbb{R}^{n \times n}$, then feasible solutions of (7) satisfy

$$\begin{pmatrix} 1 & \left(y^{(1)}\right)^T & \dots & \left(y^{(n)}\right)^T \\ y^{(1)} & Y^{(11)} & \dots & Y^{(1n)} \\ \vdots & \vdots & \ddots & \vdots \\ y^{(n)} & Y^{(n1)} & \dots & Y^{(nn)} \end{pmatrix} \succeq 0. \tag{12}$$

Feasible solutions have the following additional structure (see [26] and Theorem 3.1 in [7]):

- $Y^{(ii)}$ $(i = 1, \dots, n)$ is a diagonal matrix;
- $Y^{(ij)}$ $(i \neq j)$ is a matrix with zero diagonal;
- $trace(JY^{(ij)}) = 1$ $(i, j = 1, \dots, n)$;
- $\sum_{i=1}^n Y^{(ij)} = e \left(y^{(j)}\right)^T \quad (j = 1, \dots, n)$;
- $\operatorname{diag}(Y) = y$.

Since $\operatorname{diag}(Y) = y$ for feasible solutions, we have $y^{(i)} = \operatorname{diag}(Y^{(ii)})$ $(i = 1, \dots, n)$. Moreover, since we may also assume the structure (11), we have that

$$y^{(i)} = y_i e \quad (i = 1, \dots, n),$$

for some scalar values y_i. This implies that the diagonal elements of $Y^{(ii)}$ all equal y_i. Since the diagonal elements of $Y^{(ii)}$ sum to 1, we have $y_i = 1/n$ and $\operatorname{diag}(Y^{(ii)}) = (1/n)e$. Thus the condition

$$\begin{pmatrix} 1 & y^T \\ y & Y \end{pmatrix} \succeq 0$$

reduces to

$$Y - \frac{1}{n^2} J \succeq 0$$

by the Shur complement theorem. This is equivalent to

$$(I \otimes Q^*) Y (I \otimes Q) - \frac{1}{n^2} (I \otimes Q^*) J (I \otimes Q) \succeq 0,$$

where Q is the discrete Fourier transform matrix defined in (25) in the Appendix.

Using the properties of the Kronecker product and of Q, we get

$$\begin{pmatrix} Q^*Y^{(11)}Q & \cdots & Q^*Y^{(1n)}Q \\ \vdots & \ddots & \vdots \\ Q^*Y^{(n1)}Q & \cdots & Q^*Y^{(nn)}Q \end{pmatrix} - J \otimes \begin{pmatrix} \frac{1}{n} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix} \succeq 0.$$

Recall that $Y^{(ii)} = \frac{1}{n} I$ and that we may assume $Y^{(ij)}$ $(i \neq j)$ to be symmetric circulant, say

$$Y^{(ij)} = \sum_{k=1}^{d} x_k^{(ij)} C_k, \quad (i \neq j),$$

where C_k $(k = 1, \ldots, d)$ forms a basis of the symmetric circulant matrices with zero diagonals (see the Appendix for the precise definition). Note that the nonnegativity of $Y^{(ij)}$ is equivalent to $x_k^{(ij)} \geq 0$ $(k = 1, \ldots, d)$. Since $trace(JY^{(ij)}) = 1$, one has

$$\sum_{k=1}^{d} x_k^{(ij)} = \frac{1}{2n} \quad (i \neq j).$$

Since $\sum_{i=1}^{n} Y^{(ij)} = e \left(y^{(j)}\right)^T = \frac{1}{n} J$, one also has

$$\sum_{k=1}^{d} \sum_{i=1}^{n} x_k^{(ij)} C_k = \frac{1}{n} J.$$

By the definition of the C_k's, this implies that

$$\sum_{i=1}^{n} x_k^{(ij)} = \begin{cases} \frac{1}{n} & \text{if } 1 \leq k \leq \lfloor (n-1)/2 \rfloor, \\ \frac{1}{2n} & \text{if } k = n/2 \ (n \text{ even}). \end{cases} \tag{13}$$

Moreover,

$$Q^* Y^{(ij)} Q = \sum_{k=1}^{d} x_k^{(ij)} D_k, \quad (i \neq j),$$

where D_k is the diagonal matrix with the eigenvalues (26) of C_k on its diagonal.

Thus the LMI becomes

$$\begin{pmatrix} \frac{1}{n} I & \cdots & \sum_{k=1}^{d} x_k^{(1n)} D_k \\ \vdots & \ddots & \vdots \\ \sum_{k=1}^{d} x_k^{(1n)} D_k & \cdots & \frac{1}{n} I \end{pmatrix} - J \otimes \begin{pmatrix} \frac{1}{n} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix} \succeq 0. \tag{14}$$

The left-hand side of this LMI is a block matrix with each block being a diagonal matrix. Thus this matrix has a chordal sparsity structure (n disjoint cliques of size n). We may now use the following lemma to obtain the system of LMI's (3).

LEMMA 3.1 (cf. [14]). *Assume a $nt \times nt$ matrix has the block structure*

$$M := \begin{pmatrix} D^{(11)} & \cdots & D^{(1n)} \\ \vdots & \ddots & \vdots \\ D^{(n1)} & \cdots & D^{(nn)} \end{pmatrix},$$

where $D^{(ij)} \in \mathcal{S}_t$ are diagonal $(i, j = 1, \ldots, n)$. Then $M \succeq 0$ if and only if

$$\begin{pmatrix} D_{ii}^{(11)} & \cdots & D_{ii}^{(1n)} \\ \vdots & \ddots & \vdots \\ D_{ii}^{(n1)} & \cdots & D_{ii}^{(nn)} \end{pmatrix} \succeq 0 \quad i = 1, \ldots, t.$$

Applying the lemma to the LMI (14), and setting

$$X_{ij}^{(k)} = 2n x_k^{(ij)}, \qquad k = 1, \ldots, \lfloor n/2 \rfloor \tag{15}$$

yields the system of LMI's in (3).

Thus we have established the following result.

THEOREM 3.2. *The optimal values of the semidefinite programs* (3) *and* (7) *are equal.*

4. Relation of (3) to an SDP relaxation of Cvetković et al. We will now show that the new SDP relaxation (3) dominates an SDP relaxation (16) due to Cvetković et al. [5]. This latter relaxation is based on the fact that the spectrum of the Hamiltonian circuit $\mathcal{C}_n$ is known. In particular, the smallest eigenvalue of its Laplacian is zero and corresponds to the all ones eigenvector, while the second smallest eigenvalue equals $2 - 2\cos\left(\frac{2\pi}{n}\right)$.

The relaxation takes the form

$$TSP_{opt} \geq \min \frac{1}{2} trace(DX)$$

subject to

$$\left.\begin{aligned} Xe &= 2e, \\ \operatorname{diag}(X) &= 0, \\ 0 \leq X &\leq J, \\ 2I - X + \left(2 - 2\cos\left(\tfrac{2\pi}{n}\right)\right)(J - I) &\succeq 0. \end{aligned}\right\} \tag{16}$$

Note that the matrix variable X corresponds to the adjacency matrix of the minimal length Hamiltonian circuit.

THEOREM 4.1. *The SDP relaxation* (3) *dominates the relaxation* (16).

Proof. Assume that given $X^{(k)} \in \mathcal{S}^n$ $(k = 1, \ldots, d)$ satisfies (3). Then, $\operatorname{diag}(X^{(1)}) = 0$, while (13) and (15) imply

$$X^{(k)}e = 2e \quad (k = 1, \ldots, \lfloor (n-1)/2 \rfloor),$$

and $X^{(n/2)}e = e$ if n is even. In particular, one has $X^{(1)}e = 2e$. It remains to show that

$$2I - X^{(1)} + \left(2 - 2\cos\left(\frac{2\pi}{n}\right)\right)(J - I) \succeq 0,$$

which is the same as showing that

$$2I - X^{(1)} + \left(2 - 2\cos\left(\frac{2\pi}{n}\right)\right)\sum_{k=1}^{d} X^{(k)} \succeq 0, \tag{17}$$

since

$$\sum_{k=1}^{d} X^{(k)} = J - I.$$

We will show that the LMI (17) may be obtained as a nonnegative aggregation of the LMI's

$$I + \sum_{k=1}^{d} X^{(k)} \succeq 0$$

and

$$I + \sum_{k=1}^{d} \cos\left(\frac{2\pi ik}{n}\right) X^{(k)} \succeq 0 \quad (i = 1, \ldots, d).$$

The matrix of coefficients of these LMI's is a $(d+1) \times (d+1)$ matrix, say A, with entries:

$$A_{ij} = \cos\left(\frac{2\pi ij}{n}\right) \quad (i, j = 0, \ldots, d).$$

Since we may rewrite (17) as

$$2I + \left(1 - 2\cos\left(\frac{2\pi}{n}\right)\right) X^{(1)} + \left(2 - 2\cos\left(\frac{2\pi}{n}\right)\right)\sum_{k=2}^{d} X^{(k)} \succeq 0,$$

we need to show that the linear system $Ax = b$ has a nonnegative solution, where

$$b := \left[2, \left(1 - 2\cos\left(\frac{2\pi}{n}\right)\right), \left(2 - 2\cos\left(\frac{2\pi}{n}\right)\right), \ldots, \left(2 - 2\cos\left(\frac{2\pi}{n}\right)\right)\right]^T.$$

One may verify that, for n odd, the system $Ax = b$ has a (unique) solution given by

$$x_i = \frac{4}{n}\begin{cases} d\left(1 - \cos\left(\frac{2\pi}{n}\right)\right) & \text{if } i = 0 \\ \cos\left(\frac{2\pi}{n}\right) - \cos\left(\frac{2\pi i}{n}\right) & \text{for } i = 1, \ldots, d. \end{cases}$$

Note that x is nonnegative, as it should be. If n is even, the solution is

$$x_i = \frac{4}{n}\begin{cases} \frac{(n-1)}{2}\left(1 - \cos\left(\frac{2\pi}{n}\right)\right) & \text{if } i = 0, \\ \cos\left(\frac{2\pi}{n}\right) - \cos\left(\frac{2\pi i}{n}\right) & \text{for } i = 1, \ldots, d-1, \\ \frac{1}{2}\cos\left(\frac{2\pi}{n}\right) - \frac{1}{2}\cos\left(\frac{2\pi i}{n}\right) & \text{for } i = d. \end{cases}$$ □

In the section with numerical examples, we will present instances where the new SDP relaxation (3) is strictly better than (16).

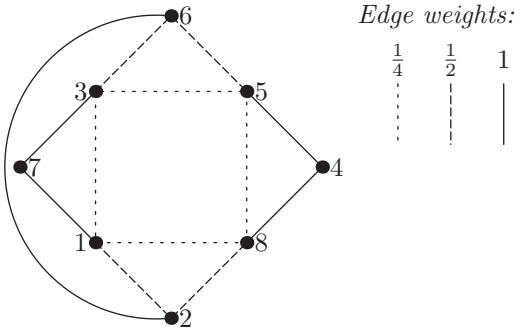


FIG. 1. *The weighted graph used in the proof of Theorem* 5.1.

5. Relation to the Held–Karp bound. One of the best-known linear programming (LP) relaxations of TSP is the LP with subtour elimination constraints:

$$TSP_{opt} \geq \min \frac{1}{2} trace(DX)$$

subject to

$$\left.\begin{array}{r} Xe = 2, \\ \mathrm{diag}(X) = 0, \\ 0 \leq X \leq J, \\ \sum_{i \in \mathcal{I},\, j \notin \mathcal{I}} X_{ij} \geq 2 \quad \forall\, \emptyset \neq \mathcal{I} \subset \{1, \ldots, n\} \end{array}\right\}. \tag{18}$$

This LP relaxation dates back to 1954 and is due to Dantzig, Fulkerson, and Johnson [6]. Its optimal value coincides with the LP bound of Held and Karp [15] (see, e.g., Theorem 21.34 in [16]), and the optimal value of the LP is commonly known as the *Held–Karp bound*.

The last constraints are called *subtour elimination inequalities* and model the fact that $\mathcal{C}_n$ is 2-connected. Although there are exponentially many subtour elimination inequalities, it is well known that the LP (18) may be solved in polynomial time using the ellipsoid method; see, e.g., Schrijver [22], section 58.5.

It was shown by Goemans and Rendl [12] that this LP relaxation dominates the SDP relaxation (16) by Cvetković et al. [5]. The next theorem shows that the LP relaxation (18) does not dominate the new SDP relaxation (3), or vice versa.

THEOREM 5.1. *The LP subtour elimination relaxation* (18) *does not dominate the new SDP relaxation* (3), *or vice versa.*

Proof. Define the 8×8 symmetric matrix $\bar{X}$ as the weighted adjacency matrix of the graph shown in Figure 1.

The matrix $\bar{X}$ satisfies the subtour elimination inequalities, since the minimum cut in the graph in Figure 1 has weight 2.

On the other hand, there does not exist a feasible solution of (3) that satisfies $X^{(1)} = \bar{X}$, as may be shown using SDP duality theory.

Conversely, in section 7 we will provide examples where the optimal value of (18) is strictly greater than the optimal value of (3) (see, e.g., the instances gr17, gr24, and bays24 there). ∎

6. An LMI cut via the number of spanning trees. In addition to the subtour elimination inequalities, there are several families of linear inequalities known for the TSP polytope; for a review, see Naddef [18] and Schrijver [22, Chapter 58].

Of particular interest to us is a valid nonlinear inequality that models the fact that $\mathcal{C}_n$ has n distinct spanning trees. To introduce the inequality we require a general form of the *matrix tree theorem*; see, e.g., Theorem VI.29 in [24] for a proof.

THEOREM 6.1 (Matrix tree theorem). *Let a simple graph $G = (V, E)$ be given and associate with each edge $e \in E$ a real variable x_e. Define the (generalized) Laplacian of G with respect to x as the $|V| \times |V|$ matrix with entries*

$$L(G)(x)_{ij} := \begin{cases} \sum_{e\,:\,e\cap i\neq\emptyset} x_e & \text{if } i = j, \\ -x_e & \text{if } \{i,j\} = e, \\ 0 & \text{else.} \end{cases}$$

Now all principal minors of $L(G)(x)$ of order $|V| - 1$ equal:

$$\sum_T \prod_{e\in T} x_e, \tag{19}$$

where the sum is over all distinct spanning trees T of G.

In particular, if $L(G)(x)$ is the usual Laplacian of a given graph, then $x_e = 1$ for all edges e of the graph, and expression (19) evaluates to the number of spanning trees in the graph.

Thus if X corresponds to the approximation of the adjacency matrix of a minimum tour, then one may require that

$$\det\,(2I - X)_{2:n,2:n} \geq n, \tag{20}$$

where $X_{2:n,2:n}$ denotes the principle submatrix of X obtained by deleting the first row and column.

The inequality (20) may be added to the above SDP relaxations (16) and (3) (with $X = X^{(1)}$), since the set

$$\{Z \succeq 0 \;:\; \det Z \geq n\}$$

is LMI representable; see, e.g., Nemirovski [19, section 3.2].

We know from numerical examples that (20) is not implied by the relaxation of Cvetković et al. (16), but do not know any examples where it is violated by a feasible $X^{(1)}$ of the new relaxation (3). Nevertheless, we have been unable to show that (20) (with $X = X^{(1)}$) is implied by (3).

7. Numerical examples. In Table 1 we give the lower bounds on some small TSPLIB[1] instances for the two SDP relaxations (3) and (16), as well as the LP relaxation with all subtour elimination constraints (18) (the Held–Karp bound). These instances have integer data, and the optimal values of the relaxations were rounded up to obtain the bounds in the table.

The SDP problems were solved by the interior point software CSDP [2] using the Yalmip interface [17] and Matlab 6.5, running on a PC with two 2.1 GHz dual-core processors and 2GB of memory.

Note that the relaxation (3) can indeed be strictly better than (16), as is clear from the gr17, bays24, and bays29 instances. Also, since the LP relaxation (18) gives better bounds than (3) for all four instances, it is worth recalling that this will not happen in general, by Theorem 5.1.

[1] http://www.iwr.uni-heidelberg.de/groups/comopt/software/TSPLIB95/

TABLE 1
Lower bounds on some small TSPLIB instances from various convex relaxations.

Problem	SDP bound (16)	SDP bound (3) (time)	LP bound (18)	TSP_{opt}
gr17	1810	2007 (39s)	2085	2085
gr21	2707	2707 (139s)	2707	2707
gr24	1230	1271 (1046s)	1272	1272
bays29	1948	2000 (2863s)	2014	2020

TABLE 2
Results for instances on $n = 8$ cities, constructed from the facet-defining inequalities.

Inequality	SDP bound (16)	SDP bound (3)	Held–Karp bound (18)	RHS
1	2	2	2	2
2	1.098	1.628	2	2
3	1.172	1.172	2	2
4	8.507	8.671	9	10
5	9	9	9	10
6	8.566	8.926	9	10
7	8.586	8.586	9	10
8	8.570	8.926	9	10
9	9	9	9	10
10	8.411	8.902	9	10
11	8.422	8.899	9	10
12	0	0	0	0
13	10.586	10.667	11	12
14	12	12	12	13
15	12.408	12.444	$12\frac{2}{3}$	14
16	14	14.078	14	16
17	16	16	16	18
18	16	16	16	18
19	16	16	16	18
20	15.185	15.926	16	18
21	18	18.025	18	20
22	20	20	20	22
23	23	23.033	23	26
24	34.586	34.739	35	38

The LMI cut from (20) was already satisfied by the optimal solutions of (16) and (3) for the four instances.

A second set of test problems was generated by considering all facet-defining inequalities for the TSP polytope on 8 nodes; see [3] for a description of these inequalities, as well as the SMAPO project web site.[2]

The facet-defining inequalities are of the form $\frac{1}{2}trace(DX) \geq RHS$ where $D \in \mathcal{S}_n$ has nonnegative integer entries and RHS is an integer. From each inequality, we form a symmetric TSP instance with distance matrix D. Thus the optimal value of the TSP instance is the value RHS. In Table 2 we give the optimal values of the LP relaxation (18) (i.e., the Held–Karp bound), the SDP relaxation of Cvetković et al. (16), and the new SDP relaxation (3) for these instances, as well as the right-hand-side RHS of each inequality $\frac{1}{2}trace(DX) \geq RHS$. For $n = 8$, there are 24 classes of facet-defining inequalities. The members of each class are equal modulo a permutation of the nodes, and we need therefore consider only one representative per class. The first three classes of inequalities are subtour elimination inequalities.

[2] http://www.iwr.uni-heidelberg.de/groups/comopt/software/SMAPO/tsp/

The numbering of the instances in Table 2 coincides with the numbering of the classes of facet-defining inequalities on the SMAPO project web site.

The new SDP bound (3) is only stronger than the Held–Karp bound (18) for the instances 16, 21, and 23 in Table 2, and for the instances 1, 5, 9, 12, 14, 17, 18, 19, and 22 the two bounds coincide. For the remaining 18 instances the Held–Karp bound is better than the SDP bound (3). However, if the bounds are rounded up, the SDP bound (3) is still better for the instances 16, 21 and 23, whereas the two (rounded) bounds are equal for all the other instances. Adding the LMI cut from (20) did not change the optimal values of the SDP relaxations (16) or (3) for any of the instances.

For $n = 9$, there are 192 classes of facet-defining inequalities of the TSP polytope [4]. Here the SDP bound (3) is better than the Held–Karp bound for 23 out of the 192 associated TSP instances. Similar to the $n = 8$ case, when rounding up, the rounded SDP bound remains better in all 23 cases and coincides with the rounded Held–Karp bound in all the remaining cases.

8. Concluding remarks. Wolsey [25] showed that the optimal value of the LP relaxation (18) is at least 2/3 the length of an optimal tour for metric TSP (see also [23]). An interesting question is whether a similar result may be proved for the new SDP relaxation (3).

Finally, the computational perspectives of the SDP relaxation (3) are somewhat limited due to its size. However, since it provides a new polynomial-time convex approximation of TSP with a rich mathematical structure, it is our hope that it may lead to a renewed interest in improving approximation results for metric TSP.

Appendix: Circulant matrices. Our discussion of circulant matrices is condensed from the review paper by Gray [13].

A circulant matrix has the form

$$C = \begin{bmatrix} c_0 & c_1 & c_2 & \cdots & & c_{n-1} \\ c_{n-1} & c_0 & c_1 & & & \\ & c_{n-1} & c_0 & c_1 & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \\ & & & & & c_1 \\ c_1 & & \cdots & & c_{n-1} & c_0 \end{bmatrix}. \tag{21}$$

Thus the entries satisfy the relation

$$C_{ij} = c_{(j-i) \mod n}. \tag{22}$$

The matrix C has eigenvalues

$$\lambda_m(C) = c_0 + \sum_{k=1}^{n-1} c_k e^{-2\pi\sqrt{-1}mk/n}, \quad m = 0, \ldots, n-1.$$

If C is symmetric with n odd, this reduces to

$$\lambda_m(C) = c_0 + \sum_{k=1}^{(n-1)/2} 2c_k \cos(2\pi mk/n), \quad m = 0, \ldots, n-1, \tag{23}$$

and when n is even we have

$$\lambda_m(C) = c_0 + \sum_{k=1}^{n/2-1} 2c_k \cos(2\pi mk/n) + c_{n/2}\cos(m\pi), \quad m = 0, \ldots, n-1. \tag{24}$$

The circulant matrices form a commutative matrix $*$-algebra, as do the symmetric circulant matrices. In particular, all circulant matrices share a set of eigenvectors, given by the columns of the *discrete Fourier transform matrix*:

$$Q_{ij} := \frac{1}{\sqrt{n}} e^{-2\pi\sqrt{-1}ij/n}, \quad i,j = 0,\ldots,n-1. \tag{25}$$

One has $Q^*Q = I$, and Q^*CQ is a diagonal matrix for any circulant matrix C. Also note that $Q^*e = \sqrt{n}e$.

We may define a basis $C^{(0)},\ldots,C^{\lfloor n/2\rfloor}$ for the symmetric circulant matrices as follows: to obtain $C^{(i)}$ we set $c_i = c_{n-i} = 1$ in (21) and all other c_j's to zero. (We set $C_0 = 2I$ and also multiply $C_{n/2}$ by 2 if n is even.)

By (23) and (24), the eigenvalues of these basis matrices are

$$\lambda_m(C^{(k)}) = 2\cos(2\pi mk/n), \quad m = 0,\ldots,n-1,\ k = 0,\ldots,\lfloor n/2\rfloor. \tag{26}$$

Also note that

$$\lambda_m(C^{(k)}) = \lambda_{n-m}(C^{(k)}), \quad m = 1,\ldots,\lfloor n/2\rfloor,\ k = 0,\ldots,\lfloor n/2\rfloor$$

so that each matrix $C^{(k)}$ has only $1 + \lfloor n/2\rfloor$ distinct eigenvalues.

Acknowledgments. Etienne de Klerk would like to thank Dragoš Cvetković and Vera Kovačević-Vujčić for past discussions on the SDP relaxation (16). The authors would also like to thank an anonymous referee for suggestions that led to a significant improvement of this paper.

REFERENCES

[1] S. Arora, *Polynomial time approximation schemes for Euclidean Traveling Salesman and other geometric problems*, J. ACM, 45 (1998), pp. 753–782.

[2] B. Borchers, *CSDP, a C library for semidefinite programming,* Optim. Methods Softw., 11/12 (1999), pp. 613–623.

[3] T. Christof, M. Jünger, and G. Reinelt, *A complete description of the traveling salesman polytope on* 8 *nodes,* Oper. Res. Lett., 10 (1991), pp. 497–500.

[4] T. Christof and G. Reinelt, *Combinatorial optimization and small polytopes,* Top, 4 (1996), pp. 1–53.

[5] D. Cvetković, M. Cangalović, and V. Kovačević-Vujčić, *Semidefinite programming methods for the symmetric traveling salesman problem,* in Proceedings of the 7th International IPCO Conference on Integer Programming and Combinatorial Optimization, Springer-Verlag, London, UK, 1999, pp. 126–136.

[6] G.B. Dantzig, D.R. Fulkerson and S.M. Johnson, *Solution of a large-scale traveling salesman problem,* Oper. Res., 2 (1954), pp. 393–410.

[7] E. de Klerk and R. Sotirov, *Exploiting Group Symmetry in Semidefinite Programming Relaxations of the Quadratic Assignment Problem*, Center Discussion Paper 2007-44, Tilburg University, The Netherlands, 2007. Available at: http://arno.uvt.nl/show.cgi?fid=60929.

[8] K. Gatermann and P.A. Parrilo, *Symmetry groups, semidefinite programs, and sum of squares,* J. Pure Appl. Algebra, 192 (2004), pp. 95–128.

[9] N. Christofides, *Worst-case analysis of a new heuristic for the travelling salesman problem,* Technical report 388, Graduate School of Industrial Administration, Carnegie-Mellon University, Pittsburgh, PA, 1976.

[10] C. Godsil, *Algebraic Combinatorics,* Chapman and Hall, London, 1993.

[11] M.X. Goemans and F. Rendl, *Semidefinite programs and association schemes,* Computing, 63 (1999), pp. 331–340.

[12] M.X. Goemans and F. Rendl, *Combinatorial Optimization,* in Handbook of Semidefinite Programming: Theory, Algorithms and Applications, H. Wolkowicz, R. Saigal, and L. Vandenberghe, eds., Kluwer, Boston, MA, 2000.

[13] R.M. Gray, *Toeplitz and circulant matrices: A review,* Found. Trends Commun. Information Theory, 2 (2006), pp. 155–239.
[14] R. Grone, C.R. Johnson, E.M. Sá, and H. Wolkowicz, *Positive definite completions of partial Hermitian matrices,* Linear Algebra Appl., 58 (1984), pp. 109–124.
[15] M. Held and R.M. Karp, *The traveling-salesman problem and minimum spanning trees,* Oper. Res., 18 (1970), pp. 1138–1162.
[16] B. Korte and J. Vygen, *Combinatorial optimization: Theory and algorithms,* 4th ed., Algorithms and Combinatorics 21, Springer-Verlag, Berlin, 2008.
[17] J. Löfberg, *YALMIP: A Toolbox for Modeling and Optimization in MATLAB,* in Proceedings of the CACSD Conference, Taipei, Taiwan, 2004. http://control.ee.ethz.ch/~joloef/yalmip.php.
[18] D. Naddef, *Polyhedral theory and branch-and-cut algorithms for the TSP,* in The Traveling Salesman Problem and Its Variations, G. Gutin and A.P. Punnen, eds., Kluwer Academic Publishers, Norwell, MA, 2002.
[19] A. Nemirovskii, *Lectures on modern convex optimization*, Lecture notes, Georgia Tech. 2005; Available at: http://www2.isye.gatech.edu/~nemirovs/Lect_ModConvOpt.pdf.
[20] P. Orponen and H. Mannila, *On approximation preserving reductions: Complete problems and robust measures,* Technical report C-1987-28, Department of Computer Science, University of Helsinki, Helsinki, 1987.
[21] C.H. Papadimitriou and S. Vempala, *On the approximability of the traveling salesman problem,* in Proceedings of the 32nd Annual ACM Symposium on Theory of Computing, 2000.
[22] A. Schrijver, *Combinatorial Optimization – Polyhedra and Efficiency*, Vol. 2, Springer-Verlag, Berlin, 2003.
[23] D.B. Shmoys and D.P. Williamson, *Analyzing the Held-Karp TSP bound: A monotonicity property with application,* Inform. Process. Lett., 35 (1990), pp. 281–285.
[24] W.T. Tutte, *Graph Theory*, Addison-Wesley, Reading, PA, 1984.
[25] L. Wolsey, *Heuristic analysis, linear programming, and branch and bound,* Math. Prog. Study, 13 (1980), pp. 121–134.
[26] Q. Zhao, S.E. Karisch, F. Rendl, and H. Wolkowicz, *Semidefinite programming relaxations for the quadratic assignment problem,* J. Combin. Optim., 2 (1998), pp. 71–109.

SIAM J. OPTIM.
Vol. 19, No. 4, pp. 1574–1609

ROBUST STOCHASTIC APPROXIMATION APPROACH TO STOCHASTIC PROGRAMMING*

A. NEMIROVSKI†, A. JUDITSKY‡, G. LAN†, AND A. SHAPIRO†

Abstract. In this paper we consider optimization problems where the objective function is given in a form of the expectation. A basic difficulty of solving such stochastic optimization problems is that the involved multidimensional integrals (expectations) cannot be computed with high accuracy. The aim of this paper is to compare two computational approaches based on Monte Carlo sampling techniques, namely, the stochastic approximation (SA) and the sample average approximation (SAA) methods. Both approaches, the SA and SAA methods, have a long history. Current opinion is that the SAA method can efficiently use a specific (say, linear) structure of the considered problem, while the SA approach is a crude subgradient method, which often performs poorly in practice. We intend to demonstrate that a properly modified SA approach can be competitive and even significantly outperform the SAA method for a certain class of convex stochastic problems. We extend the analysis to the case of convex-concave stochastic saddle point problems and present (in our opinion highly encouraging) results of numerical experiments.

Key words. stochastic approximation, sample average approximation method, stochastic programming, Monte Carlo sampling, complexity, saddle point, minimax problems, mirror descent algorithm

AMS subject classifications. 90C15, 90C25

DOI. 10.1137/070704277

1. Introduction.

In this paper we first consider the following stochastic optimization problem:

$$\min_{x\in X}\big\{f(x)=\mathbb{E}[F(x,\xi)]\big\}, \tag{1.1}$$

and then we deal with an extension of the analysis to stochastic saddle point problems. Here $X\subset\mathbb{R}^n$ is a nonempty bounded closed convex set, ξ is a random vector whose probability distribution P is supported on set $\Xi\subset\mathbb{R}^d$ and $F: X\times\Xi\to\mathbb{R}$. We assume that the expectation

$$\mathbb{E}[F(x,\xi)]=\textstyle\int_{\Xi}F(x,\xi)dP(\xi) \tag{1.2}$$

is well defined and finite valued for every $x\in X$. Moreover, we assume that the expected value function $f(\cdot)$ is *continuous* and *convex* on X. Of course, if for every $\xi\in\Xi$ the function $F(\cdot,\xi)$ is convex on X, then it follows that $f(\cdot)$ is convex. With these assumptions, (1.1) becomes a convex programming problem.

A basic difficulty of solving stochastic optimization problem (1.1) is that the multidimensional integral (expectation) (1.2) cannot be computed with a high accuracy for dimension d, say, greater than five. The aim of this paper is to compare two

*Received by the editors October 1, 2007; accepted for publication (in revised form) August 26, 2008; published electronically January 21, 2009.

http://www.siam.org/journals/siopt/19-4/70427.html

†Georgia Institute of Technology, Atlanta, Georgia 30332 (nemirovs@isye.gatech.edu, glan@isye.gatech.edu, ashapiro@isye.gatech.edu). Research of the first author was partly supported by NSF award DMI-0619977. Research of the third author was partially supported by NSF award CCF-0430644 and ONR award N00014-05-1-0183. Research of the fourth author was partly supported by NSF awards DMS-0510324 and DMI-0619977.

‡Université J. Fourier, B.P. 53, 38041 Grenoble Cedex 9, France (Anatoli.Juditsky@imag.fr).

computational approaches based on Monte Carlo sampling techniques, namely, the *stochastic approximation* (SA) and the *sample average approximation* (SAA) methods. To this end we make the following assumptions.

(A1) It is possible to generate an independent identically distributed (iid) sample $\xi_1, \xi_2, \ldots$, of realizations of random vector ξ.

(A2) There is a mechanism (an oracle), which, for a given input point $(x, \xi) \in X \times \Xi$ returns *stochastic subgradient*—a vector $\mathsf{G}(x, \xi)$ such that $\mathsf{g}(x) := \mathbb{E}[\mathsf{G}(x, \xi)]$ is well defined and is a subgradient of $f(\cdot)$ at x, i.e., $\mathsf{g}(x) \in \partial f(x)$.

Recall that if $F(\cdot, \xi)$, $\xi \in \Xi$, is convex and $f(\cdot)$ is finite valued in a neighborhood of a point x, then (cf. Strassen [28])

$$\partial f(x) = \mathbb{E}\left[\partial_x F(x, \xi)\right]. \tag{1.3}$$

In that case we can employ a measurable selection $\mathsf{G}(x, \xi) \in \partial_x F(x, \xi)$ as a stochastic subgradient. At this stage, however, this is not important, we shall see later other relevant ways for constructing stochastic subgradients.

Both approaches, the SA and SAA methods, have a long history. The SA method is going back to the pioneering paper by Robbins and Monro [21]. Since then SA algorithms became widely used in stochastic optimization (see, e.g., [3, 6, 7, 20, 22] and references therein) and, due to especially low demand for computer memory, in signal processing . In the classical analysis of the SA algorithm (it apparently goes back to the works [5] and [23]) it is assumed that $f(\cdot)$ is twice continuously differentiable and strongly convex and in the case when the minimizer of f belongs to the interior of X, exhibits asymptotically optimal rate[1] of convergence $\mathbb{E}[f(x_t) - f_*] = O(t^{-1})$ (here x_t is tth iterate and f_* is the minimal value of $f(x)$ over $x \in X$). This algorithm, however, is very sensitive to a choice of the respective stepsizes. Since "asymptotically optimal" stepsize policy can be very bad in the beginning, the algorithm often performs poorly in practice (e.g., [27, section 4.5.3.]).

An important improvement of the SA method was developed by Polyak [18] and Polyak and Juditsky [19], where longer stepsizes were suggested with consequent averaging of the obtained iterates. Under the outlined "classical" assumptions, the resulting algorithm exhibits the same optimal $O(t^{-1})$ asymptotical convergence rate, while using an easy to implement and "robust" stepsize policy. It should be mentioned that the main ingredients of Polyak's scheme—long steps and averaging—were, in a different form, proposed already in Nemirovski and Yudin [15] for the case of problems (1.1) with general-type Lipschitz continuous convex objectives and for convex-concave saddle point problems. The algorithms from [15] exhibit, in a nonasymptotical fashion, the $O(t^{-1/2})$ rate of convergence. It is possible to show that in the general convex case (without assuming smoothness and strong convexity of the objective function), this rate of $O(t^{-1/2})$ is unimprovable. For a summary of early results in this direction, see Nemirovski and Yudin [16].

The SAA approach was used by many authors in various contexts under different names. Its basic idea is rather simple: generate a (random) sample $\xi_1, \ldots, \xi_N$, of size N, and approximate the "true" problem (1.1) by the sample average problem

$$\min_{x \in X} \left\{ \hat{f}_N(x) = N^{-1} \sum_{j=1}^{N} F(x, \xi_j) \right\}. \tag{1.4}$$

[1] Throughout the paper, we speak about convergence in terms of the objective value.

Note that the SAA method is not an algorithm; the obtained SAA problem (1.4) still has to be solved by an appropriate numerical procedure. Recent theoretical studies (cf. [11, 25, 26]) and numerical experiments (see, e.g., [12, 13, 29]) show that the SAA method coupled with a good (deterministic) algorithm could be reasonably efficient for solving certain classes of two-stage stochastic programming problems. On the other hand, classical SA-type numerical procedures typically performed poorly for such problems.

We intend to demonstrate in this paper that a properly modified SA approach can be competitive and even significantly outperform the SAA method for a certain class of stochastic problems. The mirror descent SA method we propose here is a direct descendent of the stochastic mirror descent method of Nemirovski and Yudin [16]. However, the method developed in this paper is more flexible than its "ancestor": the iteration of the method is exactly the prox-step for a chosen prox-function, and the choice of prox-type function is not limited to the norm-type distance-generating functions. Close techniques, based on subgradient averaging, have been proposed in Nesterov [17] and used in [10] to solve the stochastic optimization problem (1.1). Moreover, the results on large deviations of solutions and applications of the mirror descent SA to saddle point problems, to the best of our knowledge, are new.

The rest of this paper is organized as follows. In section 2 we focus on theory of the SA method applied to (1.1). We start with outlining the relevant–to-our-goals part of the classical "$O(t^{-1})$" SA theory (section 2.1), along with its "$O(t^{-1/2})$" modifications (section 2.2). Well-known and simple results presented in these sections pave the road to our main developments carried out in section 2.3. In section 3 we extend the constructions and results of section 2.3 to the case of the convex-concave stochastic saddle point problem. In concluding section 4 we present results (in our opinion, highly encouraging) of numerical experiments with the SA algorithm (sections 2.3 and 3) applied to large-scale stochastic convex minimization and saddle point problems. Section 5 gives a short conclusion for the presented results. Finally, some technical proofs are given in the appendix.

Throughout the paper, we use the following notation. By $\|x\|_p$, we denote the ℓ_p norm of vector $x \in \mathbb{R}^n$, in particular, $\|x\|_2 = \sqrt{x^T x}$ denotes the Euclidean norm, and $\|x\|_\infty = \max\{|x_1|, \ldots, |x_n|\}$. By Π_X, we denote the metric projection operator onto the set X, that is, $\Pi_X(x) = \arg\min_{x' \in X} \|x - x'\|_2$. Note that Π_X is a nonexpanding operator, i.e.,

$$\|\Pi_X(x') - \Pi_X(x)\|_2 \le \|x' - x\|_2 \ \ \forall x', x \in \mathbb{R}^n. \tag{1.5}$$

By $O(1)$, we denote positive absolute constants. The notation $\lfloor a \rfloor$ stands for the largest integer less than or equal to $a \in \mathbb{R}$ and $\lceil a \rceil$ for the smallest integer greater than or equal to $a \in \mathbb{R}$. By $\xi_{[t]} = (\xi_1, \ldots, \xi_t)$, we denote the history of the process $\xi_1, \ldots,$ up to time t. Unless stated otherwise, all relations between random variables are supposed to hold almost surely.

2. Stochastic approximation, basic theory. In this section we discuss theory and implementations of the SA approach to the minimization problem (1.1).

2.1. Classical SA algorithm. The classical SA algorithm solves (1.1) by mimicking the simplest subgradient descent method. That is, for chosen $x_1 \in X$ and a sequence $\gamma_j > 0$, $j = 1, \ldots,$ of stepsizes, it generates the iterates by the formula

$$x_{j+1} = \Pi_X(x_j - \gamma_j \mathsf{G}(x_j, \xi_j)). \tag{2.1}$$

Of course, the crucial question of that approach is how to choose the stepsizes γ_j. Let x_* be an optimal solution of (1.1). Note that since the set X is compact and $f(x)$ is continuous, (1.1) has an optimal solution. Note also that the iterate $x_j = x_j(\xi_{[j-1]})$ is a function of the history $\xi_{[j-1]} = (\xi_1, \ldots, \xi_{j-1})$ of the generated random process and hence is random.

Denote

$$A_j = \tfrac{1}{2}\|x_j - x_*\|_2^2 \quad \text{and} \quad a_j = \mathbb{E}[A_j] = \tfrac{1}{2}\mathbb{E}\left[\|x_j - x_*\|_2^2\right]. \tag{2.2}$$

By using (1.5) and since $x_* \in X$ and hence $\Pi_X(x_*) = x_*$, we can write

$$\begin{aligned} A_{j+1} &= \tfrac{1}{2}\left\|\Pi_X(x_j - \gamma_j \mathsf{G}(x_j, \xi_j)) - x_*\right\|_2^2 \\ &= \tfrac{1}{2}\left\|\Pi_X(x_j - \gamma_j \mathsf{G}(x_j, \xi_j)) - \Pi_X(x_*)\right\|_2^2 \\ &\le \tfrac{1}{2}\left\|x_j - \gamma_j \mathsf{G}(x_j, \xi_j) - x_*\right\|_2^2 \\ &= A_j + \tfrac{1}{2}\gamma_j^2 \|\mathsf{G}(x_j, \xi_j)\|_2^2 - \gamma_j (x_j - x_*)^T \mathsf{G}(x_j, \xi_j). \end{aligned} \tag{2.3}$$

Since $x_j = x_j(\xi_{[j-1]})$ is independent of ξ_j, we have

$$\begin{aligned} \mathbb{E}\left[(x_j - x_*)^T \mathsf{G}(x_j, \xi_j)\right] &= \mathbb{E}\left\{\mathbb{E}\left[(x_j - x_*)^T \mathsf{G}(x_j, \xi_j) \,|\xi_{[j-1]}\right]\right\} \\ &= \mathbb{E}\left\{(x_j - x_*)^T \mathbb{E}\left[\mathsf{G}(x_j, \xi_j) \,|\xi_{[j-1]}\right]\right\} \\ &= \mathbb{E}\left[(x_j - x_*)^T \mathsf{g}(x_j)\right]. \end{aligned} \tag{2.4}$$

Assume now that there is a positive number M such that

$$\mathbb{E}\left[\|\mathsf{G}(x, \xi)\|_2^2\right] \le M^2 \quad \forall x \in X. \tag{2.5}$$

Then, by taking expectation of both sides of (2.3) and using (2.4), we obtain

$$a_{j+1} \le a_j - \gamma_j \mathbb{E}\left[(x_j - x_*)^T \mathsf{g}(x_j)\right] + \tfrac{1}{2}\gamma_j^2 M^2. \tag{2.6}$$

Suppose further that the expectation function $f(x)$ is differentiable and strongly convex on X, i.e., there is constant $c > 0$ such that

$$f(x') \ge f(x) + (x' - x)^T \nabla f(x) + \tfrac{1}{2}c\|x' - x\|_2^2, \quad \forall x', x \in X,$$

or equivalently that

$$(x' - x)^T (\nabla f(x') - \nabla f(x)) \ge c\|x' - x\|_2^2 \quad \forall x', x \in X. \tag{2.7}$$

Note that strong convexity of $f(x)$ implies that the minimizer x_* is unique. By optimality of x_*, we have that

$$(x - x_*)^T \nabla f(x_*) \ge 0 \quad \forall x \in X,$$

which together with (2.7) implies that $(x - x_*)^T \nabla f(x) \ge c\|x - x_*\|_2^2$. In turn, it follows that $(x - x_*)^T g \ge c\|x - x_*\|_2^2$ for all $x \in X$ and $g \in \partial f(x)$, and hence

$$\mathbb{E}\left[(x_j - x_*)^T \mathsf{g}(x_j)\right] \ge c\mathbb{E}\left[\|x_j - x_*\|_2^2\right] = 2ca_j.$$

Therefore, it follows from (2.6) that

$$a_{j+1} \le (1 - 2c\gamma_j)a_j + \tfrac{1}{2}\gamma_j^2 M^2. \tag{2.8}$$

Let us take stepsizes $\gamma_j = \theta/j$ for some constant $\theta > 1/(2c)$. Then, by (2.8), we have

$$a_{j+1} \leq (1 - 2c\theta/j)a_j + \tfrac{1}{2}\theta^2 M^2/j^2.$$

It follows by induction that

$$\mathbb{E}\left[\|x_j - x_*\|_2^2\right] = 2a_j \leq Q(\theta)/j, \tag{2.9}$$

where

$$Q(\theta) = \max\left\{\theta^2 M^2 (2c\theta - 1)^{-1}, \|x_1 - x_*\|_2^2\right\}. \tag{2.10}$$

Suppose further that x_* is an *interior* point of X and $\nabla f(x)$ is Lipschitz continuous, i.e., there is constant $L > 0$ such that

$$\|\nabla f(x') - \nabla f(x)\|_2 \leq L\|x' - x\|_2 \ \ \forall x', x \in X. \tag{2.11}$$

Then

$$f(x) \leq f(x_*) + \tfrac{1}{2}L\|x - x_*\|_2^2, \ \ \forall x \in X, \tag{2.12}$$

and hence

$$\mathbb{E}\big[f(x_j) - f(x_*)\big] \leq \tfrac{1}{2}L\mathbb{E}\left[\|x_j - x_*\|_2^2\right] \leq \tfrac{1}{2}LQ(\theta)/j, \tag{2.13}$$

where $Q(\theta)$ is defined in (2.10).

Under the specified assumptions, it follows from (2.9) and (2.13), respectively, that after t iterations, the expected error of the current solution in terms of the distance to x_* is of order $O(t^{-1/2})$, and the expected error in terms of the objective value is of order $O(t^{-1})$, provided that $\theta > 1/(2c)$. The simple example of $X = \{x : \|x\|_2 \leq 1\}$, $f(x) = \frac{1}{2}cx^T x$, and $\mathsf{G}(x, \xi) = \nabla f(x) + \xi$, with ξ having standard normal distribution $\mathcal{N}(0, I_n)$, demonstrates that the outlined upper bounds on the expected errors are tight within factors independent of t.

We have arrived at the $O(t^{-1})$ rate of convergence in terms of the expected value of the objective mentioned in the Introduction. Note, however, that the result is highly sensitive to a priori information on c. What would happen if the parameter c of strong convexity is overestimated? As a simple example, consider $f(x) = x^2/10$, $X = [-1, 1] \subset \mathbb{R}$, and assume that there is no noise, i.e., $\mathsf{G}(x, \xi) \equiv \nabla f(x)$. Suppose, further that we take $\theta = 1$ (i.e., $\gamma_j = 1/j$), which will be the optimal choice for $c = 1$, while actually here $c = 0.2$. Then the iteration process becomes

$$x_{j+1} = x_j - f'(x_j)/j = \left(1 - \frac{1}{5j}\right) x_j,$$

and hence starting with $x_1 = 1$,

$$\begin{aligned} x_j = \prod_{s=1}^{j-1}\left(1 - \frac{1}{5s}\right) &= \exp\left\{-\sum_{s=1}^{j-1} \ln\left(1 + \frac{1}{5s-1}\right)\right\} > \exp\left\{-\sum_{s=1}^{j-1}\frac{1}{5s-1}\right\} \\ &> \exp\left\{-\left(0.25 + \int_1^{j-1}\frac{1}{5t-1}dt\right)\right\} > \exp\left\{-0.25 + 0.2\ln 1.25 - \frac{1}{5}\ln j\right\} \\ &> 0.8 j^{-1/5}. \end{aligned}$$

That is, the convergence is extremely slow. For example, for $j = 10^9$, the error of the iterated solution is greater than 0.015. On the other hand, for the optimal stepsize factor of $\theta = 1/c = 5$, the optimal solution $x_* = 0$ is found in one iteration.

It could be added that the stepsizes $\gamma_j = \theta/j$ may become completely unacceptable when f loses strong convexity. For example, when $f(x) = x^4$, $X = [-1, 1]$, and there is no noise, these stepsizes result in a disastrously slow convergence: $|x_j| \geq O([\ln(j+1)]^{-1/2})$. The precise statement here is that with $\gamma_j = \theta/j$ and $0 < x_1 \leq \frac{1}{6\sqrt{\theta}}$, we have that $x_j \geq \frac{x_1}{\sqrt{1+32\theta x_1^2[1+\ln(j+1)]}}$ for $j = 1, 2, \ldots$.

We see that in order to make the SA "robust"—applicable to general convex objectives rather than to strongly convex ones—one should replace the classical stepsizes $\gamma_j = O(j^{-1})$, which can be too small to ensure a reasonable rate of convergence even in the "no noise" case, with "much larger" stepsizes. At the same time, a detailed analysis shows that "large" stepsizes poorly suppress noise. As early as in [15] it was realized that in order to resolve the arising difficulty, it makes sense to separate collecting information on the objective from generating approximate solutions. Specifically, we can use large stepsizes, say, $\gamma_j = O(j^{-1/2})$ in (2.1), thus avoiding too slow motion at the cost of making the trajectory "more noisy." In order to suppress, to some extent, this noisiness, we take, as approximate solutions, appropriate averages of the search points x_j rather than these points themselves.

2.2. Robust SA approach. Results of this section go back to Nemirovski and Yudin [15, 16]. Let us look again at the basic relations (2.2), (2.5), and (2.6). By convexity of $f(x)$, we have that $f(x) \geq f(x_t) + (x - x_t)^T \mathsf{g}(x_t)$ for any $x \in X$, and hence

$$\mathbb{E}\big[(x_t - x_*)^T \mathsf{g}(x_t)\big] \geq \mathbb{E}\big[f(x_t) - f(x_*)\big].$$

Together with (2.6), this implies (recall that $a_t = \mathbb{E}\left[\frac{1}{2}\|x_t - x_*\|_2^2\right]$)

$$\gamma_t \mathbb{E}\big[f(x_t) - f(x_*)\big] \leq a_t - a_{t+1} + \tfrac{1}{2}\gamma_t^2 M^2.$$

It follows that whenever $1 \leq i \leq j$, we have

$$\sum_{t=i}^{j} \gamma_t \mathbb{E}\big[f(x_t) - f(x_*)\big] \leq \sum_{t=i}^{j} [a_t - a_{t+1}] + \tfrac{1}{2}M^2 \sum_{t=i}^{j} \gamma_t^2 \leq a_i + \tfrac{1}{2}M^2 \sum_{t=i}^{j} \gamma_t^2, \tag{2.14}$$

and hence, setting $\nu_t = \frac{\gamma_t}{\sum_{\tau=i}^{j} \gamma_\tau}$,

$$\mathbb{E}\left[\sum_{t=i}^{j} \nu_t f(x_t) - f(x_*)\right] \leq \frac{a_i + \frac{1}{2}M^2 \sum_{t=i}^{j} \gamma_t^2}{\sum_{t=i}^{j} \gamma_t}. \tag{2.15}$$

Note that $\nu_t \geq 0$ and $\sum_{t=i}^{j} \nu_t = 1$. Consider the points

$$\tilde{x}_i^j = \sum_{t=i}^{j} \nu_t x_t, \tag{2.16}$$

and let

$$D_X = \max_{x \in X} \|x - x_1\|_2. \tag{2.17}$$

By convexity of X, we have $\tilde{x}_i^j \in X$, and, by convexity of f, we have $f(\tilde{x}_i^j) \leq \sum_{t=i}^j \nu_t f(x_t)$. Thus, by (2.15) and in view of $a_1 \leq D_X^2$ and $a_i \leq 4D_X^2$, $i > 1$, we get

$$\begin{aligned} &\text{(a)} \quad \mathbb{E}\left[f(\tilde{x}_1^j) - f(x_*)\right] \leq \frac{D_X^2 + M^2 \sum_{t=1}^j \gamma_t^2}{2 \sum_{t=1}^j \gamma_t} \quad \text{for} \quad 1 \leq j, \\ &\text{(b)} \quad \mathbb{E}\left[f(\tilde{x}_i^j) - f(x_*)\right] \leq \frac{4D_X^2 + M^2 \sum_{t=i}^j \gamma_t^2}{2 \sum_{t=i}^j \gamma_t} \quad \text{for} \quad 1 < i \leq j. \end{aligned} \tag{2.18}$$

Based on the resulting bounds on the expected inaccuracy of approximate solutions $\tilde{x}_i^j$, we can now develop "reasonable" stepsize policies along with the associated efficiency estimates.

Constant stepsizes and basic efficiency estimate. Assume that the number N of iterations of the method is fixed in advance and that $\gamma_t = \gamma$, $t = 1, \ldots, N$. Then it follows by (2.18(a)) that

$$\mathbb{E}\left[f\left(\tilde{x}_1^N\right) - f(x_*)\right] \leq \frac{D_X^2 + M^2 N \gamma^2}{2N\gamma}. \tag{2.19}$$

Minimizing the right-hand side of (2.19) over $\gamma > 0$, we arrive at the *constant* stepsize policy

$$\gamma_t = \frac{D_X}{M\sqrt{N}}, \quad t = 1, \ldots, N, \tag{2.20}$$

along with the associated efficiency estimate

$$\mathbb{E}\left[f\left(\tilde{x}_1^N\right) - f(x_*)\right] \leq \frac{D_X M}{\sqrt{N}}. \tag{2.21}$$

With the constant stepsize policy (2.20), we also have, for $1 \leq K \leq N$,

$$\mathbb{E}\left[f\left(\tilde{x}_K^N\right) - f(x_*)\right] \leq \frac{D_X M}{\sqrt{N}} \left[\frac{2N}{N - K + 1} + \frac{1}{2}\right]. \tag{2.22}$$

When $K/N \leq 1/2$, the right-hand side of (2.22) coincides, within an absolute constant factor, with the right-hand side of (2.21). Finally, for a constant $\theta > 0$, passing from the stepsizes (2.20) to the stepsizes

$$\gamma_t = \frac{\theta D_X}{M\sqrt{N}}, \quad t = 1, \ldots, N, \tag{2.23}$$

the efficiency estimate becomes

$$\mathbb{E}\left[f\left(\tilde{x}_K^N\right) - f(x_*)\right] \leq \max\left\{\theta, \theta^{-1}\right\} \frac{D_X M}{\sqrt{N}} \left[\frac{2N}{N - K + 1} + \frac{1}{2}\right], \ 1 \leq K \leq N. \tag{2.24}$$

Discussion. We conclude that the expected error in terms of the objective of *Robust SA* algorithm (2.1), (2.16), with constant stepsize policy (2.20), after N iterations is of order $O(N^{-1/2})$ in our setting. Of course, this is worse than the rate $O(N^{-1})$ for the classical SA algorithm as applied to a smooth strongly convex function attaining minimum at a point from the interior of the set X. However, the error bounds (2.21)

and (2.22) are guaranteed independently of any smoothness and/or strong convexity assumptions on f. All that matters is the convexity of f on the convex compact set X and the validity of (2.5). Moreover, scaling the stepsizes by positive constant θ affects the error bound (2.24) *linearly* in $\max\{\theta, \theta^{-1}\}$. This can be compared with a possibly disastrous effect of such scaling in the classical SA algorithm discussed in section 2.1. These observations, in particular the fact that there is no necessity in "fine tuning" the stepsizes to the objective function f, explain the adjective "robust" in the name of the method. Finally, it can be shown that without additional, as compared to convexity and (2.5), assumptions on f, the accuracy bound (2.21) within an absolute constant factor is the best one allowed by statistics (cf. [16]).

Varying stepsizes. When the number of steps is not fixed in advance, it makes sense to replace constant stepsizes with the stepsizes

$$\gamma_t = \frac{\theta D_X}{M\sqrt{t}}, \quad t = 1, 2, \ldots . \tag{2.25}$$

From (2.18(b)) it follows that with this stepsize policy, one has, for $1 \le K \le N$,

$$\mathbb{E}\left[f\left(\tilde{x}_K^N\right) - f(x_*)\right] \le \frac{D_X M}{\sqrt{N}} \left[\frac{2}{\theta}\left(\frac{N}{N-K+1}\right) + \frac{\theta}{2}\sqrt{\frac{N}{K}}\right]. \tag{2.26}$$

Choosing K as a fixed fraction of N, i.e., setting $K = \lceil rN \rceil$, with a fixed $r \in (0,1)$, we get the efficiency estimate

$$\mathbb{E}\left[f\left(\tilde{x}_K^N\right) - f(x_*)\right] \le C(r) \max\left\{\theta, \theta^{-1}\right\} \frac{D_X M}{\sqrt{N}}, \quad N = 1, 2, \ldots, \tag{2.27}$$

with an easily computable factor $C(r)$ depending solely on r. This bound, up to a factor depending solely on r and θ, coincides with the bound (2.21), with the advantage that our new stepsize policy should not be adjusted to a fixed-in-advance number of steps N.

2.3. Mirror descent SA method. On a close inspection, the robust SA algorithm from section 2.2 is intrinsically linked to the Euclidean structure of $\mathbb{R}^n$. This structure plays the central role in the very construction of the method (see (2.1)), the same as in the associated efficiency estimates, like (2.21) (since the quantities D_X, M participating in the estimates are defined in terms of the Euclidean norm, see (2.17) and (2.5)). By these reasons, from now on, we refer to the algorithm from section 2.2 as the (robust) *Euclidean SA* (E-SA). In this section we develop a substantial generalization of the E-SA approach allowing us to adjust, to some extent, the method to the geometry, not necessary Euclidean, of the problem in question. We shall see in the meantime that we can gain a lot, both theoretically and numerically, from such an adjustment. A rudimentary form of the generalization to follow can be found in Nemirovski and Yudin [16], from where the name "mirror descent" originates.

Let $\|\cdot\|$ be a (general) norm on $\mathbb{R}^n$ and $\|x\|_* = \sup_{\|y\|\le 1} y^T x$ be its dual norm. We say that a function $\omega : X \to \mathbb{R}$ is a *distance-generating function* modulus $\alpha > 0$ with respect to $\|\cdot\|$, if ω is convex and continuous on X, the set

$$X^o = \{x \in X : \partial\omega(x) \ne \emptyset\} \tag{2.28}$$

is convex (note that X^o always contains the relative interior of X) and restricted to X^o, ω is continuously differentiable and strongly convex with parameter α with

respect to $\|\cdot\|$, i.e.,

$$(x'-x)^T(\nabla\omega(x')-\nabla\omega(x)) \geq \alpha\|x'-x\|^2 \ \ \forall x', x \in X^o. \tag{2.29}$$

A simple example of a distance-generating function is $\omega(x) = \frac{1}{2}\|x\|_2^2$ (modulus 1 with respect to $\|\cdot\|_2$, $X^o = X$).

Let us define function $V : X^o \times X \to \mathbb{R}_+$ as follows:

$$V(x,z) = \omega(z) - \left[\omega(x) + \nabla\omega(x)^T(z-x)\right]. \tag{2.30}$$

In what follows we shall refer to $V(\cdot,\cdot)$ as *prox-function* associated with distance-generating function $\omega(x)$ (it is also called Bregman distance [4]). Note that $V(x,\cdot)$ is nonnegative and is a strongly convex modulus α with respect to the norm $\|\cdot\|$. Let us define *prox-mapping* $P_x : \mathbb{R}^n \to X^o$, associated with ω and a point $x \in X^o$, viewed as a parameter, as follows:

$$P_x(y) = \arg\min_{z\in X}\left\{y^T(z-x) + V(x,z)\right\}. \tag{2.31}$$

Observe that the minimum in the right-hand side of (2.31) is attained since ω is continuous on X and X is compact, and all the minimizers belong to X^o, whence the minimizer is unique, since $V(x,\cdot)$ is strongly convex on X^o. Thus, the prox-mapping is well defined.

For $\omega(x) = \frac{1}{2}\|x\|_2^2$, we have $P_x(y) = \Pi_X(x-y)$ so that (2.1) is the recurrence

$$x_{j+1} = P_{x_j}(\gamma_j \mathsf{G}(x_j,\xi_j)), \quad x_1 \in X^o. \tag{2.32}$$

Our goal is to demonstrate that the main properties of the recurrence (2.1) (which from now on we call the E-SA recurrence) are inherited by (2.32), *whatever be the underlying distance-generating function* $\omega(x)$.

The statement of the following lemma is a simple consequence of the optimality conditions of the right-hand side of (2.31) (proof of this lemma is given in the appendix).

LEMMA 2.1. *For every $u \in X, x \in X^o$, and $y \in \mathbb{R}^n$, one has*

$$V(P_x(y),u) \leq V(x,u) + y^T(u-x) + \frac{\|y\|_*^2}{2\alpha}. \tag{2.33}$$

Using (2.33) with $x = x_j$, $y = \gamma_j \mathsf{G}(x_j,\xi_j)$, and $u = x_*$, we get

$$\gamma_j(x_j - x_*)^T\mathsf{G}(x_j,\xi_j) \leq V(x_j,x_*) - V(x_{j+1},x_*) + \frac{\gamma_j^2}{2\alpha}\|\mathsf{G}(x_j,\xi_j)\|_*^2. \tag{2.34}$$

Note that with $\omega(x) = \frac{1}{2}\|x\|_2^2$, one has $V(x,z) = \frac{1}{2}\|x-z\|_2^2$, $\alpha = 1$, $\|\cdot\|_* = \|\cdot\|_2$. That is, (2.34) becomes nothing but the relation (2.6), which played a crucial role in all the developments related to the E-SA method. We are about to process, in a completely similar fashion, the relation (2.34) in the case of a general distance-generating function, thus arriving at the mirror descent SA. Specifically, setting

$$\Delta_j = \mathsf{G}(x_j,\xi_j) - \mathsf{g}(x_j), \tag{2.35}$$

we can rewrite (2.34), with j replaced by t, as

$$\gamma_t(x_t - x_*)^T\mathsf{g}(x_t) \leq V(x_t,x_*) - V(x_{t+1},x_*) - \gamma_t\Delta_t^T(x_t - x_*) + \frac{\gamma_t^2}{2\alpha}\|\mathsf{G}(x_t,\xi_t)\|_*^2. \tag{2.36}$$

Summing up over $t = 1, \ldots, j$, and taking into account that $V(x_{j+1}, u) \geq 0$, $u \in X$, we get

$$\sum_{t=1}^{j} \gamma_t (x_t - x_*)^T \mathsf{g}(x_t) \leq V(x_1, x_*) + \sum_{t=1}^{j} \frac{\gamma_t^2}{2\alpha} \|\mathsf{G}(x_t, \xi_t)\|_*^2 - \sum_{t=1}^{j} \gamma_t \Delta_t^T (x_t - x_*). \tag{2.37}$$

Setting $\nu_t = \frac{\gamma_t}{\sum_{i=1}^{j} \gamma_i}$, $t = 1, \ldots, j$, and

$$\tilde{x}_1^j = \sum_{t=1}^{j} \nu_t x_t \tag{2.38}$$

and invoking convexity of $f(\cdot)$, we have

$$\begin{aligned} \sum_{t=1}^{j} \gamma_t (x_t - x_*)^T \mathsf{g}(x_t) &\geq \sum_{t=1}^{j} \gamma_t \left[f(x_t) - f(x_*) \right] \\ &= \left(\sum_{t=1}^{j} \gamma_t \right) \left[\sum_{t=1}^{j} \nu_t f(x_t) - f(x_*) \right] \\ &\geq \left(\sum_{t=1}^{j} \gamma_t \right) \left[f(\tilde{x}_j) - f(x_*) \right], \end{aligned}$$

which combines with (2.37) to imply that

$$f(\tilde{x}_1^j) - f(x_*) \leq \frac{V(x_1, x_*) + \sum_{t=1}^{j} \frac{\gamma_t^2}{2\alpha} \|\mathsf{G}(x_t, \xi_t)\|_*^2 - \sum_{t=1}^{j} \gamma_t \Delta_t^T (x_t - x_*)}{\sum_{t=1}^{j} \gamma_t}. \tag{2.39}$$

Let us suppose, as in the previous section (cf. (2.5)), that we are given a positive number M_* such that

$$\mathbb{E}\left[\|\mathsf{G}(x, \xi)\|_*^2 \right] \leq M_*^2 \ \ \forall x \in X. \tag{2.40}$$

Taking expectations of both sides of (2.39) and noting that (i) x_t is a deterministic function of $\xi_{[t-1]} = (\xi_1, \ldots, \xi_{t-1})$, (ii) conditional on $\xi_{[t-1]}$, the expectation of Δ_t is 0, and (iii) the expectation of $\|\mathsf{G}(x_t, \xi_t)\|_*^2$ does not exceed M_*^2, we obtain

$$\mathbb{E}\left[f(\tilde{x}_1^j) - f(x_*) \right] \leq \frac{\max_{u \in X} V(x_1, u) + (2\alpha)^{-1} M_*^2 \sum_{t=1}^{j} \gamma_t^2}{\sum_{t=1}^{j} \gamma_t}. \tag{2.41}$$

Assume from now on that the method starts with the minimizer of ω:

$$x_1 = \operatorname{argmin}_X \omega(x).$$

Then, from (2.30), it follows that

$$\max_{z \in X} V(x_1, z) \leq D_{\omega, X}^2, \tag{2.42}$$

where

$$D_{\omega, X} := \left[\max_{z \in X} \omega(z) - \min_{z \in X} \omega(z) \right]^{1/2}. \tag{2.43}$$

Consequently, (2.41) implies that

$$\mathbb{E}\left[f(\tilde{x}_1^j) - f(x_*) \right] \leq \frac{D_{\omega, X}^2 + \frac{1}{2\alpha} M_*^2 \sum_{t=1}^{j} \gamma_t^2}{\sum_{t=1}^{j} \gamma_t}. \tag{2.44}$$

Constant stepsize policy. Assuming that the total number of steps N is given in advance and $\gamma_t = \gamma$, $t = 1, \ldots, N$, optimizing the right-hand side of (2.44) over $\gamma > 0$ we arrive at the constant stepsize policy

$$\gamma_t = \frac{\sqrt{2\alpha}D_{\omega,X}}{M_*\sqrt{N}}, \quad t = 1, \ldots, N \tag{2.45}$$

and the associated efficiency estimate

$$\mathbb{E}\left[f\left(\tilde{x}_1^N\right) - f(x_*)\right] \leq D_{\omega,X}M_*\sqrt{\frac{2}{\alpha N}} \tag{2.46}$$

(cf. (2.20), (2.21)). For a constant $\theta > 0$, passing from the stepsizes (2.45) to the stepsizes

$$\gamma_t = \frac{\theta\sqrt{2\alpha}D_{\omega,X}}{M_*\sqrt{N}}, \quad t = 1, \ldots, N, \tag{2.47}$$

the efficiency estimate becomes

$$\mathbb{E}\left[f\left(\tilde{x}_1^N\right) - f(x_*)\right] \leq \max\left\{\theta, \theta^{-1}\right\} D_{\omega,X}M_*\sqrt{\frac{2}{\alpha N}}. \tag{2.48}$$

We refer to the method (2.32), (2.38), and (2.47) as the (robust) *mirror descent SA* algorithm with constant stepsize policy.

Probabilities of large deviations. So far, all our efficiency estimates were upper bounds on the expected nonoptimality, in terms of the objective, of approximate solutions generated by the algorithms. Here we complement these results with bounds on probabilities of large deviations. Observe that by Markov inequality, (2.48) implies that

$$\mathrm{Prob}\left\{f\left(\tilde{x}_1^N\right) - f(x_*) > \varepsilon\right\} \leq \frac{\sqrt{2}\max\left\{\theta, \theta^{-1}\right\} D_{\omega,X}M_*}{\varepsilon\sqrt{\alpha N}} \quad \forall \varepsilon > 0. \tag{2.49}$$

It is possible, however, to obtain much finer bounds on deviation probabilities when imposing more restrictive assumptions on the distribution of $\mathsf{G}(x, \xi)$. Specifically, assume that

$$\mathbb{E}\left[\exp\left\{\|\mathsf{G}(x,\xi)\|_*^2/M_*^2\right\}\right] \leq \exp\{1\} \quad \forall x \in X. \tag{2.50}$$

Note that condition (2.50) is stronger than (2.40). Indeed, if a random variable Y satisfies $\mathbb{E}[\exp\{Y/a\}] \leq \exp\{1\}$ for some $a > 0$, then by Jensen inequality, $\exp\{\mathbb{E}[Y/a]\} \leq \mathbb{E}[\exp\{Y/a\}] \leq \exp\{1\}$, and therefore, $\mathbb{E}[Y] \leq a$. Of course, condition (2.50) holds if $\|\mathsf{G}(x,\xi)\|_* \leq M_*$ for all $(x, \xi) \in X \times \Xi$.

PROPOSITION 2.2. *In the case of* (2.50) *and for the constant stepsizes* (2.47)*, the following holds for any* $\Omega \geq 1$*:*

$$\mathrm{Prob}\left\{f\left(\tilde{x}_1^N\right) - f(x_*) > \frac{\sqrt{2}\max\left\{\theta, \theta^{-1}\right\} M_*D_{\omega,X}(12 + 2\Omega)}{\sqrt{\alpha N}}\right\} \leq 2\exp\{-\Omega\}. \tag{2.51}$$

Proof of this proposition is given in the appendix.

Varying stepsizes. Same as in the case of E-SA, we can modify the mirror descent SA algorithm to allow for time-varying stepsizes and "sliding averages" of the search points x_t in the role of approximate solutions, thus getting rid of the necessity to fix in advance the number of steps. Specifically, consider

$$\begin{aligned}\overline{D}_{\omega,X} &:= \sqrt{2}\sup_{x\in X^o, z\in X}\left[\omega(z)-\omega(x)-(z-x)^T\nabla\omega(x)\right]^{1/2}\\ &= \sup_{x\in X^o, z\in X}\sqrt{2V(x,z)}\end{aligned} \tag{2.52}$$

and assume that $\overline{D}_{\omega,X}$ is finite. This is definitely so when ω is continuously differentiable on the entire X. Note that for the E-SA, that is, with $\omega(x)=\frac{1}{2}\|x\|_2^2$, $\overline{D}_{\omega,X}$ is the Euclidean diameter of X.

In the case of (2.52), setting

$$\tilde{x}_i^j = \frac{\sum_{t=i}^j \gamma_t x_t}{\sum_{t=i}^j \gamma_t}, \tag{2.53}$$

summing up inequalities (2.34) over $K\le t\le N$, and acting exactly as when deriving (2.39), we get for $1\le K\le N$,

$$f\left(\tilde{x}_K^N\right)-f(x_*)\le \frac{V(x_K,x_*)+\sum_{t=K}^N \frac{\gamma_t^2}{2\alpha}\|\mathsf{G}(x_t,\xi_t)\|_*^2-\sum_{t=K}^N\gamma_t\Delta_t^T(x_t-x_*)}{\sum_{t=K}^N\gamma_t}.$$

Noting that $V(x_K,x_*)\le\frac{1}{2}\overline{D}_{\omega,X}^2$ and taking expectations, we arrive at

$$\mathbb{E}\left[f\left(\tilde{x}_K^N\right)-f(x_*)\right]\le\frac{\frac{1}{2}\overline{D}_{\omega,X}^2+\frac{1}{2\alpha}M_*^2\sum_{t=K}^N\gamma_t^2}{\sum_{t=K}^N\gamma_t} \tag{2.54}$$

(cf. (2.44)). It follows that with a decreasing stepsize policy

$$\gamma_t=\frac{\theta\overline{D}_{\omega,X}\sqrt{\alpha}}{M_*\sqrt{t}},\quad t=1,2,\ldots, \tag{2.55}$$

one has for $1\le K\le N$,

$$\mathbb{E}\left[f\left(\tilde{x}_K^N\right)-f(x_*)\right]\le\frac{\overline{D}_{\omega,X}M_*}{\sqrt{\alpha}\sqrt{N}}\left[\frac{2}{\theta}\frac{N}{N-K+1}+\frac{\theta}{2}\sqrt{\frac{N}{K}}\right] \tag{2.56}$$

(cf. (2.26)). In particular, with $K=\lceil rN\rceil$ for a fixed $r\in(0,1)$, we get an efficiency estimate

$$\mathbb{E}\left[f\left(\tilde{x}_K^N\right)-f(x_*)\right]\le C(r)\max\left\{\theta,\theta^{-1}\right\}\frac{\overline{D}_{\omega,X}M_*}{\sqrt{\alpha}\sqrt{N}}, \tag{2.57}$$

completely similar to the estimate (2.27) for the E-SA.

Discussion. Comparing (2.21) to (2.46) and (2.27) to (2.57), we see that for both the Euclidean and the mirror descent robust SA, the expected inaccuracy, in terms of the objective, of the approximate solution built in course of N steps is $O(N^{-1/2})$. A benefit of the mirror descent over the Euclidean algorithm is in the

potential possibility to reduce the constant factor hidden in $O(\cdot)$ by adjusting the norm $\|\cdot\|$ and the distance-generating function $\omega(\cdot)$ to the geometry of the problem.

Example. Let $X = \{x \in \mathbb{R}^n : \sum_{i=1}^n x_i = 1,\ x \geq 0\}$ be a standard simplex. Consider two setups for the mirror descent SA:

— *Euclidean setup*, where $\|\cdot\| = \|\cdot\|_2$ and $\omega(x) = \frac{1}{2}\|x\|_2^2$, and

— ℓ_1*-setup*, where $\|\cdot\| = \|\cdot\|_1$, with $\|\cdot\|_* = \|\cdot\|_\infty$ and ω is the *entropy* function

$$\omega(x) = \sum_{i=1}^n x_i \ln x_i. \tag{2.58}$$

The Euclidean setup leads to the Euclidean robust SA, which is easily implementable (computing the prox-mapping requires $O(n \ln n)$ operations) and guarantees that

$$\mathbb{E}\left[f\left(\tilde{x}_1^N\right) - f(x_*)\right] \leq O(1) \max\left\{\theta, \theta^{-1}\right\} M N^{-1/2}, \tag{2.59}$$

with $M^2 = \sup_{x \in X} \mathbb{E}\left[\|\mathsf{G}(x,\xi)\|_2^2\right]$, provided that the constant M is known and the stepsizes (2.23) are used (see (2.24), (2.17), and note that the Euclidean diameter of X is $\sqrt{2}$).

The ℓ_1-setup corresponds to $X^o = \{x \in X : x > 0\}$, $D_{\omega,X} = \sqrt{\ln n}$, $\alpha = 1$, and $x_1 = \operatorname{argmin}_X \omega = n^{-1}(1, \ldots, 1)^T$ (see appendix). The associated mirror descent SA is easily implementable: the prox-function here is

$$V(x,z) = \sum_{i=1}^n z_i \ln \frac{z_i}{x_i},$$

and the prox-mapping $P_x(y) = \operatorname{argmin}_{z \in X}\left[y^T(z-x) + V(x,z)\right]$ can be computed in $O(n)$ operations according to the explicit formula

$$[P_x(y)]_i = \frac{x_i e^{-y_i}}{\sum_{k=1}^n x_k e^{-y_k}},\ \ i = 1, \ldots, n.$$

The efficiency estimate guaranteed with the ℓ_1-setup is

$$\mathbb{E}\left[f\left(\tilde{x}_1^N\right) - f(x_*)\right] \leq O(1) \max\left\{\theta, \theta^{-1}\right\} \sqrt{\ln n} M_* N^{-1/2}, \tag{2.60}$$

with

$$M_*^2 = \sup_{x \in X} \mathbb{E}\left[\|\mathsf{G}(x,\xi)\|_\infty^2\right],$$

provided that the constant M_* is known and the constant stepsizes (2.47) are used (see (2.48) and (2.40)). To compare (2.60) and (2.59), observe that $M_* \leq M$, and the ratio M_*/M can be as small as $n^{-1/2}$. Thus, the efficiency estimate for the ℓ_1-setup never is much worse than the estimate for the Euclidean setup, and for large n, can be *far better* than the latter estimate:

$$\sqrt{\frac{1}{\ln n}} \leq \frac{M}{\sqrt{\ln n} M_*} \leq \sqrt{\frac{n}{\ln n}},\ N = 1, 2, \ldots,$$

both the upper and the lower bounds being achievable. Thus, when X is a standard simplex of large dimension, we have strong reasons to prefer the ℓ_1-setup to the usual Euclidean one.

Note that $\|\cdot\|_1$-norm can be coupled with "good" distance-generating functions different from the entropy one, e.g., with the function

$$\omega(x) = (\ln n)\sum_{i=1}^{n} |x_i|^{1+\frac{1}{\ln n}}, \quad n \geq 3. \tag{2.61}$$

Whenever $0 \in X$ and $\mathrm{Diam}_{\|\cdot\|_1}(X) \equiv \max_{x,y\in X} \|x-y\|_1$ equal to 1 (these conditions can always be ensured by scaling and shifting X), for the just-outlined setup, one has $\overline{D}_{\omega,X} = O(1)\sqrt{\ln n}$, $\alpha = O(1)$, so that the associated mirror descent robust SA guarantees that with $M_*^2 = \sup_{x\in X} \mathbb{E}\left[\|\mathsf{G}(x,\xi)\|_\infty^2\right]$ and $N \geq 1$,

$$\mathbb{E}\left[f\left(\tilde{x}^N_{\lceil rN\rceil}\right) - f(x_*)\right] \leq C(r)\frac{M_*\sqrt{\ln n}}{\sqrt{N}} \tag{2.62}$$

(see (2.57)), while the efficiency estimate for the Euclidean robust SA is

$$\mathbb{E}\left[f(\tilde{x}^N_{\lceil rN\rceil}) - f(x_*)\right] \leq C(r)\frac{M\,\mathrm{Diam}_{\|\cdot\|_2}(X)}{\sqrt{N}}, \tag{2.63}$$

with

$$M^2 = \sup_{x\in X} \mathbb{E}\left[\|\mathsf{G}(x,\xi)\|_2^2\right] \ \text{ and } \ \mathrm{Diam}_{\|\cdot\|_2}(X) = \max_{x,y\in X} \|x-y\|_2.$$

Ignoring logarithmic in n factors, the second estimate (2.63) can be much better than the first estimate (2.62) only when $\mathrm{Diam}_{\|\cdot\|_2}(X) \ll 1 = \mathrm{Diam}_{\|\cdot\|_1}(X)$, as it is the case, e.g., when X is an Euclidean ball. On the other hand, when X is an $\|\cdot\|_1$-ball or its nonnegative part (which is the simplex), so that the $\|\cdot\|_1$- and $\|\cdot\|_2$-diameters of X are of the same order, the first estimate (2.62) is much more attractive than the estimate (2.63) due to potentially much smaller constant M_*.

Comparison with the SAA approach. We compare now theoretical complexity estimates for the robust mirror descent SA and the SAA methods. Consider the case when (i) $X \subset \mathbb{R}^n$ is contained in the $\|\cdot\|_p$-ball of radius R, $p = 1, 2$, and the SA in question is either the E-SA ($p = 2$), or the SA associated with $\|\cdot\|_1$ and the distance-generating function[2] (2.61), (ii) in SA, the constant stepsize rule (2.45) is used, and (iii) the "light tail" assumption (2.50) takes place.

Given $\varepsilon > 0$, $\delta \in (0, 1/2)$, let us compare the number of steps $N = N_{\mathrm{SA}}$ of SA, which, with probability $\geq 1-\delta$, results in an approximate solution $\tilde{x}_1^N$ such that $f(\tilde{x}_1^N) - f(x_*) \leq \varepsilon$, with the sample size $N = N_{\mathrm{SAA}}$ for the SAA resulting in the same accuracy guarantees. According to Proposition 2.2 we have that $\mathrm{Prob}\left[f(\tilde{x}_1^N) - f(x_*) > \varepsilon\right] \leq \delta$ for

$$N_{\mathrm{SA}} = O(1)\varepsilon^{-2}D_{\omega,X}^2 M_*^2 \ln^2(1/\delta), \tag{2.64}$$

where M_* is the constant from (2.50) and $D_{\omega,X}$ is defined in (2.43). Note that the constant M_* depends on the chosen norm, $D_{\omega,X}^2 = O(1)R^2$ for $p = 2$, and $D_{\omega,X}^2 = O(1)\ln(n)R^2$ for $p = 1$.

This can be compared with the estimate of the sample size (cf. [25, 26])

$$N_{\mathrm{SAA}} = O(1)\varepsilon^{-2}R^2M_*^2\left[\ln(1/\delta) + n\ln\left(RM_*/\varepsilon\right)\right]. \tag{2.65}$$

[2]In the second case, we apply the SA after the variables are scaled to make X the unit $\|\cdot\|_1$-ball.

We see that both SA and SAA methods have logarithmic in δ and quadratic (or nearly so) in $1/\varepsilon$ complexity in terms of the corresponding sample sizes. It should be noted, however, that the SAA method requires solution of the corresponding (deterministic) problem, while the SA approach is based on simple calculations as long as stochastic subgradients could be easily computed.

3. Stochastic saddle point problem. We show in this section how the mirror descent SA algorithm can be modified to solve a convex-concave stochastic saddle point problem. Consider the following minimax (saddle point) problem:

$$\min_{x\in X}\max_{y\in Y}\big\{\phi(x,y)=\mathbb{E}[\Phi(x,y,\xi)]\big\}. \tag{3.1}$$

Here $X\subset\mathbb{R}^n$ and $Y\subset\mathbb{R}^m$ are nonempty bounded closed convex sets, ξ is a random vector whose probability distribution P is supported on set $\Xi\subset\mathbb{R}^d$, and $\Phi: X\times Y\times\Xi\to\mathbb{R}$. We assume that for every $(x,y)\in X\times Y$, the expectation

$$\mathbb{E}[\Phi(x,y,\xi)]=\int_{\Xi}\Phi(x,y,\xi)dP(\xi)$$

is well defined and finite valued and that the expected value function $\phi(x,y)$ is *convex* in $x\in X$ and *concave* in $y\in Y$. It follows that (3.1) is a *convex-concave saddle point* problem. In addition, we assume that $\phi(\cdot,\cdot)$ is *Lipschitz continuous* on $X\times Y$. It is well known that, in the above setting, (3.1) is solvable, i.e., the corresponding "primal" and "dual" optimization problems

$$\min_{x\in X}\left[\max_{y\in Y}\phi(x,y)\right]\quad\text{and}\quad\max_{y\in Y}\left[\min_{x\in X}\phi(x,y)\right],$$

respectively, are solvable with equal optimal values, denoted ϕ^*, and pairs (x^*,y^*) of optimal solutions to the respective problems form the set of saddle points of $\phi(x,y)$ on $X\times Y$.

As in the case of the minimization problem (1.1), we assume that neither the function $\phi(x,y)$ nor its sub/supergradients in x and y are available explicitly. However, we make the following assumption.

(A′2) We have at our disposal an oracle which, given an input of point $(x,y,\xi)\in X\times Y\times\Xi$, returns a *stochastic subgradient*, that is, $(n+m)$-dimensional vector $\mathsf{G}(x,y,\xi)=\begin{bmatrix}\mathsf{G}_x(x,y,\xi)\\ -\mathsf{G}_y(x,y,\xi)\end{bmatrix}$ such that vector

$$\mathsf{g}(x,y)=\begin{bmatrix}\mathsf{g}_x(x,y)\\ -\mathsf{g}_y(x,y)\end{bmatrix}=\begin{bmatrix}\mathbb{E}[\mathsf{G}_x(x,y,\xi)]\\ -\mathbb{E}[\mathsf{G}_y(x,y,\xi)]\end{bmatrix}$$

is well defined, $\mathsf{g}_x(x,y)\in\partial_x\phi(x,y)$, and $-\mathsf{g}_y(x,y)\in\partial_y(-\phi(x,y))$.

For example, if for every $\xi\in\Xi$ the function $\Phi(\cdot,\cdot,\xi)$ is *convex-concave* and the respective subdifferential and integral operators are interchangeable, we ensure (A′2) by setting

$$\mathsf{G}(x,y,\xi)=\begin{bmatrix}\mathsf{G}_x(x,y,\xi)\\ -\mathsf{G}_y(x,y,\xi)\end{bmatrix}\in\begin{bmatrix}\partial_x\Phi(x,y,\xi)\\ \partial_y(-\Phi(x,y,\xi))\end{bmatrix}.$$

Let $\|\cdot\|_x$ be a norm on $\mathbb{R}^n$ and $\|\cdot\|_y$ be a norm on $\mathbb{R}^m$, and let $\|\cdot\|_{*,x}$ and $\|\cdot\|_{*,y}$ stand for the corresponding dual norms. As in section 2.1, the basic assumption we make about the stochastic oracle (aside from its unbiasedness, which we have already postulated) is that we know positive constants $M_{*,x}^2$ and $M_{*,y}^2$ such that

$$\mathbb{E}\left[\|\mathsf{G}_x(u,v,\xi)\|_{*,x}^2\right]\le M_{*,x}^2\ \text{ and }\ \mathbb{E}\left[\|\mathsf{G}_y(u,v,\xi)\|_{*,y}^2\right]\le M_{*,y}^2\ \ \forall(u,v)\in X\times Y. \tag{3.2}$$

3.1. Mirror SA algorithm for saddle point problems. We equip X and Y with distance-generating functions $\omega_x : X \to \mathbb{R}$ modulus α_x with respect to $\|\cdot\|_x$, and $\omega_y : Y \to \mathbb{R}$ modulus α_y with respect to $\|\cdot\|_y$. Let $D_{\omega_x,X}$ and $D_{\omega_y,Y}$ be the respective constants (see definition (2.42)). We equip $\mathbb{R}^n \times \mathbb{R}^m$ with the norm

$$\|(x,y)\| = \sqrt{\tfrac{\alpha_x}{2D^2_{\omega_x,X}}\|x\|_x^2 + \tfrac{\alpha_y}{2D^2_{\omega_y,Y}}\|y\|_y^2}, \tag{3.3}$$

so that the dual norm is

$$\|(\zeta,\eta)\|_* = \sqrt{\tfrac{2D^2_{\omega_x,X}}{\alpha_x}\|\zeta\|_{*,x}^2 + \tfrac{2D^2_{\omega_y,Y}}{\alpha_y}\|\eta\|_{*,y}^2} \tag{3.4}$$

and set

$$M_*^2 = \frac{2D^2_{\omega_x,X}}{\alpha_x}M_{*,x}^2 + \frac{2D^2_{\omega_y,Y}}{\alpha_y}M_{*,y}^2. \tag{3.5}$$

It follows by (3.2) that

$$\mathbb{E}\big[\|\mathsf{G}(x,y,\xi)\|_*^2\big] \le M_*^2. \tag{3.6}$$

We use the notation $z = (x,y)$ and equip the set $Z = X \times Y$ with the distance-generating function

$$\omega(z) = \frac{\omega_x(x)}{2D^2_{\omega_x,X}} + \frac{\omega_y(y)}{2D^2_{\omega_y,Y}}.$$

It is immediately seen that ω indeed is a distance-generating function for Z modulus $\alpha = 1$ with respect to the norm $\|\cdot\|$ and that $Z^o = X^o \times Y^o$ and $D_{\omega,Z} = 1$. In what follows, $V(z,u) : Z^o \times Z \to \mathbb{R}$ and $P_z(\zeta) : \mathbb{R}^{n+m} \to Z^o$ are the prox-function and prox-mapping associated with ω and Z (see (2.30), (2.31)).

We are ready now to present the mirror SA algorithm for saddle point problems. This is the iterative procedure (compare with (2.32))

$$z_{j+1} = P_{z_j}(\gamma_j \mathsf{G}(z_j,\xi_j)), \tag{3.7}$$

where the initial point $z_1 \in Z$ is chosen to be the minimizer of $\omega(z)$ on Z. As before (compare with (2.38)), we define approximate solution $\tilde z_1^j = (\tilde x_1^j, \tilde y_1^j)$ of (3.1) after j iterations as

$$\tilde z_1^j = \frac{\sum_{t=1}^j \gamma_t z_t}{\sum_{t=1}^j \gamma_t}. \tag{3.8}$$

We refer to the procedure (3.7), (3.8) as the *saddle point mirror SA* algorithm.

Let us analyze convergence properties of the algorithm. We measure quality of an approximate solution $\tilde z = (\tilde x, \tilde y)$ by the error

$$\epsilon_\phi(\tilde z) := \left[\max_{y\in Y}\phi(\tilde x, y) - \phi_*\right] + \left[\phi_* - \min_{x\in X}\phi(x,\tilde y)\right] = \max_{y\in Y}\phi(\tilde x, y) - \min_{x\in X}\phi(x,\tilde y).$$

By convexity of $\phi(\cdot, y)$, we have

$$\phi(x_t,y_t) - \phi(x,y_t) \le (x_t - x)^T \mathsf{g}_x(x_t,y_t) \ \ \forall x \in X$$

and by concavity of $\phi(x,\cdot)$,

$$\phi(x_t,y)-\phi(x_t,y_t)\le (y-y_t)^T\mathtt{g}_y(x_t,y_t)\ \ \forall y\in Y$$

so that for all $z=(x,y)\in Z$,

$$\phi(x_t,y)-\phi(x,y_t)\le (x_t-x)^T\mathtt{g}_x(x_t,y_t)+(y-y_t)^T\mathtt{g}_y(x_t,y_t)=(z_t-z)^T\mathtt{g}(z_t).$$

Using once again the convexity-concavity of ϕ, we write

$$\begin{aligned}\epsilon_\phi(\tilde z_1^j)&=\max_{y\in Y}\phi\left(\tilde x_1^j,y\right)-\min_{x\in X}\phi\left(x,\tilde y_1^j\right)\\ &\le\left[\sum_{t=1}^j\gamma_t\right]^{-1}\left[\max_{y\in Y}\sum_{t=1}^j\gamma_t\phi(x_t,y)-\min_{x\in X}\sum_{t=1}^j\gamma_t\phi(x,y_t)\right]\\ &\le\left[\sum_{t=1}^j\gamma_t\right]^{-1}\max_{z\in Z}\sum_{t=1}^j\gamma_t(z_t-z)^T\mathtt{g}(z_t).\end{aligned}\tag{3.9}$$

To bound the right-hand side of (3.9), we use the result of the following lemma (its proof is given in the appendix).

LEMMA 3.1. *In the above setting, for any $j\ge 1$, the following inequality holds:*

$$\mathbb{E}\left[\max_{z\in Z}\sum_{t=1}^j\gamma_t(z_t-z)^T\mathtt{g}(z_t)\right]\le 2+\frac{5}{2}M_*^2\sum_{t=1}^j\gamma_t^2.\tag{3.10}$$

Now to get an error bound for the solution $\tilde z_1^j$, it suffices to substitute inequality (3.10) into (3.9) to obtain

$$\mathbb{E}\left[\epsilon_\phi\left(\tilde z_1^j\right)\right]\le\left[\sum_{t=1}^j\gamma_t\right]^{-1}\left[2+\frac{5}{2}M_*^2\sum_{t=1}^j\gamma_t^2\right].\tag{3.11}$$

Constant stepsizes and basic efficiency estimates. For a fixed number of steps N, with the constant stepsize policy

$$\gamma_t=\frac{2\theta}{M_*\sqrt{5N}},\ \ t=1,\dots,N,\tag{3.12}$$

condition (3.6) and estimate (3.11) imply that

$$\begin{aligned}\epsilon_\phi\left(\tilde z_1^N\right)&\le 2\max\left\{\theta,\theta^{-1}\right\}M_*\sqrt{\tfrac{5}{N}}\\ &=2\max\left\{\theta,\theta^{-1}\right\}\sqrt{\frac{10\left[\alpha_yD^2_{\omega_x,X}M^2_{*,x}+\alpha_xD^2_{\omega_y,Y}M^2_{*,y}\right]}{\alpha_x\alpha_yN}}.\end{aligned}\tag{3.13}$$

Variable stepsizes. Same as in the minimization case, assuming that

$$\begin{aligned}\overline{D}_{\omega,Z}&:=\sqrt{2}\sup_{z\in Z^o,w\in Z}\left[\omega(w)-\omega(z)-(w-z)^T\nabla\omega(z)\right]^{1/2}\\ &=\sqrt{2}\left[\sup_{z\in Z^o,w\in Z}V(z,w)\right]^{1/2}\end{aligned}\tag{3.14}$$

is finite, we can pass from constant stepsizes on a fixed "time horizon" to decreasing stepsize policy

$$\gamma_t = \frac{\theta \overline{D}_{\omega,Z}}{M_* \sqrt{t}}, \quad t = 1, 2, \ldots$$

(compare with (2.55) and take into account that we are in the situation of $\alpha = 1$), and from the averaging of all iterates to the "sliding averaging"

$$\tilde{z}_i^j = \frac{\sum_{t=i}^j \gamma_t z_t}{\sum_{t=i}^j \gamma_t},$$

arriving at the efficiency estimates (compare with (2.56) and (2.57))

$$\begin{aligned} \epsilon\left(\tilde{z}_K^N\right) &\le \frac{\overline{D}_{\omega,Z} M_*}{\sqrt{N}} \left[\frac{2}{\theta} \frac{N}{N-K+1} + \frac{5\theta}{2} \sqrt{\frac{N}{K}}\right], \quad 1 \le K \le N, \\ \epsilon\left(\tilde{z}_{\lceil rN \rceil}^N\right) &\le C(r) \max\left\{\theta, \theta^{-1}\right\} \frac{\overline{D}_{\omega,Z} M_*}{\sqrt{N}}, \qquad r \in (0,1). \end{aligned} \tag{3.15}$$

Probabilities of large deviations. Assume that instead of (3.2), the following stronger assumption holds:

$$\begin{aligned} \mathbb{E}\left[\exp\left\{\|\mathsf{G}_x(u,v,\xi)\|_{*,x}^2 / M_{*,x}^2\right\}\right] &\le \exp\{1\}, \\ \mathbb{E}\left[\exp\left\{\|\mathsf{G}_y(x,y,\xi)\|_{*,y}^2 / M_{*,y}^2\right\}\right] &\le \exp\{1\}. \end{aligned} \tag{3.16}$$

Proposition 3.2. *In the case of* (3.16)*, with the stepsizes given by* (3.12) and (3.6)*, one has, for any* $\Omega > 1$*,*

$$\text{Prob}\left\{\epsilon_\phi\left(\tilde{z}_1^N\right) > \frac{(8+2\Omega)\max\left\{\theta,\theta^{-1}\right\}\sqrt{5} M_*}{\sqrt{N}}\right\} \le 2\exp\{-\Omega\}. \tag{3.17}$$

Proof of this proposition is given in the appendix.

3.2. Application to minimax stochastic problems. Consider the following minimax stochastic problem:

$$\min_{x \in X} \max_{1 \le i \le m} \left\{f_i(x) = \mathbb{E}[F_i(x,\xi)]\right\}, \tag{3.18}$$

where $X \subset \mathbb{R}^n$ is a nonempty bounded closed convex set, ξ is a random vector whose probability distribution P is supported on set $\Xi \subset \mathbb{R}^d$, and $F_i : X \times \Xi \to \mathbb{R}$, $i = 1, \ldots, m$. We assume that the expected value functions $f_i(\cdot)$, $i = 1, \ldots, m$, are well defined, finite valued, *convex*, and *Lipschitz continuous* on X. Then the minimax problem (3.18) can be formulated as the following saddle point problem:

$$\min_{x \in X} \max_{y \in Y} \left\{\phi(x,y) = \sum_{i=1}^m y_i f_i(x)\right\}, \tag{3.19}$$

where $Y = \{y \in \mathbb{R}^m : \sum_{i=1}^m y_i = 1,\ y \ge 0\}$.

Assume that we are able to generate independent realizations $\xi_1, \ldots,$ of random vector ξ, and, for given $x \in X$ and $\xi \in \Xi$, we can compute $F_i(x,\xi)$ and its *stochastic subgradient* $\mathsf{G}_i(x,\xi)$ such that $\mathsf{g}_i(x) = \mathbb{E}[\mathsf{G}_i(x,\xi)]$ is well defined and $\mathsf{g}_i(x) \in \partial f_i(x)$,

$x \in X$, $i = 1, \ldots, m$. In other words, we have a stochastic oracle for the problem (3.19) such that assumption (A′2) holds, with

$$\mathsf{G}(x, y, \xi) = \begin{bmatrix} \sum_{i=1}^m y_i \mathsf{G}_i(x, \xi) \\ -\big(F_1(x, \xi), \ldots, F_m(x, \xi)\big) \end{bmatrix} \tag{3.20}$$

and

$$\mathsf{g}(x, y) = \mathbb{E}[\mathsf{G}(x, y, \xi)] = \begin{bmatrix} \sum_{i=1}^m y_i \mathsf{g}_i(x) \\ -\big(f_1(x), \ldots, f_m(x)\big) \end{bmatrix} \in \begin{bmatrix} \partial_x \phi(x, y) \\ -\partial_y \phi(x, y) \end{bmatrix}. \tag{3.21}$$

Suppose that the set X is equipped with norm $\|\cdot\|_x$, whose dual norm is $\|\cdot\|_{*,x}$, and a distance-generating function ω modulus α_x with respect to $\|\cdot\|_x$, and let $R_x^2 = \frac{D^2_{\omega_x, X}}{\alpha_x}$. We equip the set Y with the norm $\|\cdot\|_y = \|\cdot\|_1$, so that $\|\cdot\|_{*,y} = \|\cdot\|_\infty$, and with the distance-generating function

$$\omega_y(y) = \sum_{i=1}^m y_i \ln y_i$$

and set $R_y^2 = \frac{D^2_{\omega_y, Y}}{\alpha_y} = \ln m$. Next, following (3.3), we set

$$\|(x, y)\| = \sqrt{\frac{\|x\|_x^2}{2R_x^2} + \frac{\|y\|_1^2}{2R_y^2}},$$

and hence

$$\|(\zeta, \eta)\|_* = \sqrt{2R_x^2 \|\zeta\|_{*,x}^2 + 2R_y^2 \|\eta\|_\infty^2}.$$

Let us assume uniform bounds:

$$\max_{1 \le i \le m} \mathbb{E}\left[\|\mathsf{G}_i(x, \xi)\|_{*,x}^2\right] \le M_{*,x}^2, \quad \mathbb{E}\left[\max_{1 \le i \le m} |F_i(x, \xi)|^2\right] \le M_{*,y}^2, \quad i = 1, \ldots, m.$$

Note that

$$\mathbb{E}\left[\|\mathsf{G}(x, y, \xi)\|_*^2\right] = 2R_x^2 \, \mathbb{E}\left[\Big\|\sum_{i=1}^m y_i \mathsf{G}_i(x, \xi)\Big\|_{*,x}^2\right] + 2R_y^2 \, \mathbb{E}\left[\|F(x, \xi)\|_\infty^2\right],$$

and since $y \in Y$,

$$\left\|\sum_{i=1}^m y_i \mathsf{G}_i(x, \xi)\right\|_{*,x}^2 \le \left(\sum_{i=1}^m y_i \|\mathsf{G}_i(x, \xi)\|_{*,x}\right)^2 \le \sum_{i=1}^m y_i \|\mathsf{G}_i(x, \xi)\|_{*,x}^2.$$

It follows that

$$\mathbb{E}\left[\|\mathsf{G}(x, y, \xi)\|_*^2\right] \le M_*^2, \tag{3.22}$$

where

$$M_*^2 = 2R_x^2 M_{*,x}^2 + 2R_y^2 M_{*,y}^2 = 2R_x^2 M_{*,x}^2 + 2M_{*,y}^2 \ln m.$$

Let us now use the saddle point mirror SA algorithm (3.7)–(3.8) with the constant stepsize policy

$$\gamma_t = \frac{2}{M_*\sqrt{5N}}, \quad t = 1, 2, \ldots, N.$$

When substituting the value of M_*, we obtain the following from (3.13):

$$\begin{aligned} \mathbb{E}\left[\epsilon_\phi\left(\tilde{z}_1^N\right)\right] &= \mathbb{E}\left[\max_{y\in Y}\phi\left(\hat{x}_1^N, y\right) - \min_{x\in X}\phi\left(x, \hat{y}_1^N\right)\right] \\ &\le 2M_*\sqrt{\tfrac{5}{N}} \le 2\sqrt{\tfrac{10\left[R_x^2 M_{*,x}^2 + M_{*,x}^2 \ln m\right]}{N}}. \end{aligned} \tag{3.23}$$

Discussion. Looking at the bound (3.23), one can make the following important observation. The error of the saddle point mirror SA algorithm in this case is "almost independent" of the number m of constraints (it grows as $O(\sqrt{\ln m})$ as m increases). The interested reader can easily verify that if an E-SA algorithm were used in the same setting (i.e., the algorithm tuned to the norm $\|\cdot\|_y = \|\cdot\|_2$), the corresponding bound would grow with m much faster (in fact, our error bound would be $O(\sqrt{m})$ in that case).

Note that properties of the saddle point mirror SA can be used to reduce significantly the arithmetic cost of the algorithm implementation. To this end let us look at the definition (3.20) of the stochastic oracle: In order to obtain a realization $\mathsf{G}(x, y, \xi)$, one has to compute m random subgradients $\mathsf{G}_i(x, \xi)$, $i = 1, \ldots, m$, and then their convex combination $\sum_{i=1}^m y_i \mathsf{G}_i(x, \xi)$. Now let η be an independent of ξ and uniformly distributed on $[0, 1]$ random variable, and let $\imath(\eta, y) : [0, 1] \times Y \to \{1, \ldots, m\}$ equal to i when $\sum_{s=1}^{i-1} y_s < \eta \le \sum_{s=1}^{i} y_s$. That is, random variable $\hat{\imath} = \imath(\eta, y)$ takes values $1, \ldots, m$ with probabilities $y_1, \ldots, y_m$. Consider random vector

$$\mathsf{G}(x, y, (\xi, \eta)) = \begin{bmatrix} \mathsf{G}_{\imath(\eta,y)}(x, \xi) \\ -\big(F_1(x, \xi), \ldots, F_m(x, \xi)\big) \end{bmatrix}. \tag{3.24}$$

We refer to $\mathsf{G}(x, y, (\xi, \eta))$ as a *randomized oracle* for problem (3.19), the corresponding random parameter being (ξ, η). By construction, we still have $\mathbb{E}\big[\mathsf{G}(x, y, (\xi, \eta))\big] = \mathsf{g}(x, y)$, where g is defined in (3.21), and, moreover, the same bound (3.22) holds for $\mathbb{E}\big[\|\mathsf{G}(x, y, (\xi, \eta))\|_*^2\big]$. We conclude that the accuracy bound (3.23) holds for the error of the saddle point mirror SA algorithm with randomized oracle. On the other hand, in the latter procedure only one randomized subgradient $\mathsf{G}_{\hat{\imath}}(x, \xi)$ per iteration is computed. This simple idea is further developed in another interesting application of the saddle point mirror SA algorithm to bilinear matrix games, which we discuss next.

3.3. Application to bilinear matrix games. Consider the standard matrix game problem, that is, problem (3.1) with

$$\phi(x, y) = y^T A x + b^T x + c^T y,$$

where $A \in \mathbb{R}^{m\times n}$, and X and Y are the standard simplexes:

$$X = \left\{x \in \mathbb{R}^n : x \ge 0, \sum_{j=1}^n x_j = 1\right\}, \quad Y = \left\{y \in \mathbb{R}^m : y \ge 0, \sum_{i=1}^m y_i = 1\right\}.$$

In the case in question it is natural to equip X (respectively, Y) with the $\|\cdot\|_1$-norm on $\mathbb{R}^n$ (respectively, $\mathbb{R}^m$). We choose entropies as the corresponding distance-generating functions:

$$\omega_x(x) = \sum_{i=1}^n x_i \ln x_i, \quad \omega_y(x) = \sum_{i=1}^m y_i \ln y_i \qquad \left[\Rightarrow \tfrac{D^2_{\omega_x,X}}{\alpha_x} = \ln n, \tfrac{D^2_{\omega_y,Y}}{\alpha_y} = \ln m\right].$$

According to (3.3), we set

$$\|(x,y)\| = \sqrt{\frac{\|x\|_1^2}{2\ln n} + \frac{\|y\|_1^2}{2\ln m}} \Rightarrow \|(\zeta,\eta)\|_* = \sqrt{2\|\zeta\|_\infty^2 \ln n + 2\|\eta\|_\infty^2 \ln m}. \tag{3.25}$$

In order to compute the estimates $\mathsf{G}(x,y,\xi)$ of $\mathsf{g}(x,y) = (b + A^T y, -c - Ax)$, to be used in the saddle point mirror SA iterations (3.7), we use the *randomized oracle*

$$\mathsf{G}(x,y,\xi) = \begin{bmatrix} c + A^{\imath(\xi_1,y)} \\ -b - A_{\imath(\xi_2,x)} \end{bmatrix}, \tag{3.26}$$

where ξ_1 and ξ_2 are independent uniformly distributed on $[0,1]$ random variables and $\hat{j} = \imath(\xi_1, y)$, $\hat{\imath} = \imath(\xi_2, x)$ are defined as in (3.24) (i.e., $\hat{j}$ can take values $1,\dots,m$, with probabilities $y_1,\dots,y_m$ and $\hat{\imath}$ can take values $1,\dots,n$, with probabilities $x_1,\dots,x_n$), and A_j, $[A^i]^T$ are jth column and ith row in A, respectively.

Note that

$$\mathsf{g}(x,y) \equiv \mathbb{E}\left[\mathsf{G}\left(x,y,\left(\hat{j},\hat{\imath}\right)\right)\right] \in \begin{bmatrix} \partial_x \phi(x,y) \\ \partial_y(-\phi(x,y)) \end{bmatrix}. \tag{3.27}$$

Besides this,

$$\begin{array}{ll} |\mathsf{G}(x,y,\xi)_i| \le \max_{1\le j\le m} \|A^j + b\|_\infty, & 1 \le i \le n, \\ |\mathsf{G}(x,y,\xi)_i| \le \max_{1\le j\le n} \|A_j + c\|_\infty, & n+1 \le i \le n+m, \end{array}$$

whence, invoking (3.25), for any $x \in X$, $y \in Y$, and ξ,

$$\|\mathsf{G}(x,y,\xi)\|_*^2 \le M_*^2 = 2\ln n \max_{1\le j\le m} \|A^j + b\|_\infty^2 + 2\ln m \max_{1\le j\le n} \|A_j + c\|_\infty^2. \tag{3.28}$$

The bottom line is that our stochastic gradients along with the just-defined M_* satisfy both (A′2) and (3.16), and therefore with the constant stepsize policy (3.12), we have

$$\mathbb{E}\left[\epsilon_\phi\left(\tilde{z}_1^N\right)\right] = \mathbb{E}\left[\max_{y\in Y} \phi\left(\tilde{x}_1^N, y\right) - \min_{x\in X} \phi\left(x, \tilde{y}_1^N\right)\right] \le 2M_* \sqrt{\frac{5}{N}} \tag{3.29}$$

(cf. (3.13)). In our present situation, Proposition 3.2 in a slightly refined form (for proof, see the appendix) reads as follows.

PROPOSITION 3.3. *With the constant stepsize policy* (3.12)*, for the just-defined algorithm, one has for any* $\Omega \ge 1$*, that*

$$\mathrm{Prob}\left\{\epsilon_\phi\left(\tilde{z}_1^N\right) > 2M_* \sqrt{\tfrac{5}{N}} + \tfrac{4\overline{M}}{\sqrt{N}}\Omega\right\} \le \exp\left\{-\Omega^2/2\right\}, \tag{3.30}$$

where

$$\overline{M} = \max_{1\le j\le m} \|A^j + b\|_\infty + \max_{1\le j\le n} \|A_j + c\|_\infty. \tag{3.31}$$

Discussion. Consider a bilinear matrix game with $n \geq m$, $\ln(m) = O(1)\ln(n)$, and $b = c = 0$ (so that $M_* = O(1)\sqrt{\ln n}\,\overline{M}$ and $\overline{M} = \max_{i,j} |A_{ij}|$; see (3.28), (3.31)). Suppose that we are interested to solve it within a fixed relative accuracy ρ, that is, to ensure that the (perhaps random) approximate solution $\tilde{z}_1^N$, which we get after N iterations, satisfies the error bound

$$\epsilon_\phi(\tilde{z}_N) \leq \rho \max_{1 \leq i,\, j \leq n} |A_{ij}|$$

with probability at least $1 - \delta$. According to (3.30), to this end, one can use the randomized saddle point mirror SA algorithm (3.7), (3.8), (3.26) with stepsizes (3.12), (3.28) and with

$$N = O(1)\frac{\ln n + \ln(1/\delta)}{\rho^2}. \tag{3.32}$$

The computational cost of building $\tilde{z}_1^N$ with this approach is

$$\mathcal{C}(\rho) = O(1)\frac{[\ln n + \ln(1/\delta)]\,[\mathcal{R} + n]}{\rho^2}$$

arithmetic operations, where $\mathcal{R}$ is the arithmetic cost of extracting a column/row from A given the index of this column/row. The total number of rows and columns visited by the algorithm does not exceed the number of steps N as given in (3.32) so that the total number of entries in A used in the course of the entire computation does not exceed

$$M = O(1)\frac{n(\ln n + \ln(1/\delta))}{\rho^2}.$$

When ρ is fixed, $m = O(1)n$ and n is large, M is incomparably less than the total number mn of entries in A. Thus, our algorithm exhibits *sublinear-time behavior*: it produces reliable solutions of prescribed quality to large-scale matrix games by inspecting a negligible, as $n \to \infty$, part of randomly selected data. Note that randomization here is critical.[3] It can be seen that a deterministic algorithm, which is capable to find a solution with (deterministic) relative accuracy $\rho \leq 0.1$, has to "see" in the worst case *at least* $O(1)n$ rows/columns of A.

4. Numerical results. In this section, we report the results of our computational experiments where we compare the performance of the robust mirror descent SA method and the SAA method applied to three stochastic programming problems, namely: a stochastic utility problem, a stochastic max-flow problem, and a network planning problem with random demand. We also present a small simulation study of the performance of randomized mirror SA algorithm for bilinear matrix games.

The algorithms we were testing are the two variants of the robust mirror descent SA. The first variant, the E-SA, is as described in section 2.2; in terms of section 2.3, this is nothing but mirror descent robust SA with Euclidean setup; see the example in section 2.3. The second variant, referred to as the *non-Euclidean* SA (N-SA), is the mirror descent robust SA with ℓ_1-setup; see, the example in section 2.3.

[3] The possibility to solve matrix games in a sublinear-time fashion by a randomized algorithm was discovered by Grigoriadis and Khachiyan [9]. Their "ad hoc" algorithm is similar, although not completely identical to ours, and possesses the same complexity bounds.

TABLE 4.1
Selecting stepsize policy.

[method: N-SA, N:2,000, K:10,000, instance: L1]

	θ			
Policy	0.1	1	5	10
Variable	−7.4733	−7.8865	−7.8789	−7.8547
Constant	−6.9371	−7.8637	−7.9037	−7.8971

These two variants of the SA method are compared with the SAA approach in the following way: fixing an iid. sample (of size N) for the random variable ξ, we apply the three aforementioned methods to obtain approximate solutions for the test problem under consideration, and then the quality of the solutions yielded by these algorithms is evaluated using another iid. sample of size $K >> N$. It should be noted that SAA itself is not an algorithm, and in our experiments, it was coupled with the non-Euclidean restricted memory level (NERML) [2]—a powerful deterministic algorithm for solving the sample average problem (1.4).

4.1. Preliminaries.

Algorithmic schemes. Both E-SA and N-SA were implemented according to the description in section 2.3, the number of steps N being the parameter of a particular experiment. In such an experiment, we generated $\approx \log_2 N$ candidate solutions $\tilde{x}_i^N$, with $N-i+1 = \min[2^k, N]$, $k = 0, 1, \ldots, \lceil \log_2 N \rceil$. We then used an additional sample to estimate the objective at these candidate solutions in order to choose the best of these candidates, specifically, as follows: we used a relatively short sample to choose the two "most promising" of the candidate solutions, and then a large sample (of size $K \gg N$) to identify the best of these two candidates, thus getting the "final" solution. The computational effort required by this simple postprocessing is *not* reflected in the tables to follow.

The stepsizes. At the "pilot stage" of our experimentation, we made a decision on which stepsize policy—(2.47) or (2.55)—to choose and how to identify the underlying parameters M_* and θ. In all our experiments, M_* was estimated by taking the maxima of $\|\mathsf{G}(\cdot,\cdot)\|_*$ over a small (just 100) calls to the stochastic oracle at randomly generated feasible solutions. As about the value of θ and type of the stepsize policy ((2.47) or (2.55)), our choice was based on the results of experimentation with a single test problem (instance L1 of the utility problem, see below); some results of this experimentation are presented in Table 4.1. We have found that the constant stepsize policy (2.47) with $\theta = 0.1$ for the E-SA and $\theta = 5$ for the N-SA slightly outperforms other variants we have considered. This particular policy, combined with the aforementioned scheme for estimating M_*, was used in all subsequent experiments.

Format of test problems. All our test problems are of the form $\min_{x\in X} f(x)$, $f(x) = \mathbb{E}[F(x,\xi)]$, where the domain X either is a standard simplex $\{x \in \mathbb{R}^n : x \geq 0, \sum_i x_i = 1\}$ or can be converted into such a simplex by scaling of the original variables.

Notation in the tables. Below,

- n is the design dimension of an instance,
- N is the sample size (i.e., the number of steps in SA, and the size of the sample used to build the stochastic average in SAA),
- `Obj` is the empirical mean of the random variable $F(x,\xi)$, x being the approximate solution generated by the algorithm in question. The empirical means are taken over a large ($K = 10^4$ elements) dedicated sample,
- `CPU` is the CPU time in seconds.

TABLE 4.2
SA versus SAA on the stochastic utility problem.

-		L1: $n = 500$		L2: $n = 1{,}000$		L3: $n = 2{,}000$		L4: $n = 5{,}000$	
ALG.	N	Obj	CPU	Obj	CPU	Obj	CPU	Obj	CPU
N-SA	100	−7.7599	0	−5.8340	0	−7.1419	1	−5.4688	3
	1,000	−7.8781	2	−5.9152	2	−7.2312	6	−5.5716	13
	2,000	−7.8987	2	−5.9243	5	−7.2513	10	−5.5847	25
	4,000	−7.9075	5	−5.9365	12	−7.2595	20	−5.5935	49
E-SA	100	−7.6895	0	−5.7988	1	−7.0165	1	−4.9364	4
	1,000	−7.8559	2	−5.8919	4	−7.2029	7	−5.3895	20
	2,000	−7.8737	3	−5.9067	7	−7.2306	15	−5.4870	39
	4,000	−7.8948	7	−5.9193	13	−7.2441	29	−5.5354	77
SAA	100	−7.6571	7	−5.6346	8	−6.9748	19	−5.3360	44
	1,000	−7.8821	31	−5.9221	68	−7.2393	134	−5.5656	337
	2,000	−7.9100	72	−5.9313	128	−7.2583	261	−5.5878	656
	4,000	−7.9087	113	−5.9384	253	−7.2664	515	−5.5967	1,283

TABLE 4.3
The variability for the stochastic utility problem.

-		N-SA			E-SA			SAA		
		Obj		CPU	Obj		CPU	Obj		CPU
Inst	N	Mean	Dev	(Avg.)	Mean	Dev	(Avg.)	Mean	Dev	(Avg.)
L2	1,000	−5.9159	0.0025	2.63	−5.8925	0.0024	4.99	−5.9219	0.0047	67.31
L2	2,000	−5.9258	0.0022	5.03	−5.9063	0.0019	7.09	−5.9328	0.0028	131.25

4.2. A stochastic utility problem. Our first experiment was carried out with the utility model

$$\min_{x\in X}\left\{f(x)=\mathbb{E}\left[\phi\Big(\sum_{i=1}^{n}(i/n+\xi_i)x_i\Big)\right]\right\}, \tag{4.1}$$

where $X=\{x\in\mathbb{R}^n : x\geq 0, \sum_{i=1}^n x_i = 1\}$, $\xi_i \sim \mathcal{N}(0,1)$ are independent and $\phi(\cdot)$ is a piecewise linear convex function given by $\phi(t)=\max\{v_1+s_1t,\ldots,v_m+s_mt\}$, where v_k and s_k are certain constants. In our experiment, we used $m=10$ breakpoints, all located on $[0,1]$. The four instances L1, L2, L3, L4 we dealt with were of dimension varying from 500 to 2,000, each instance—with its own randomly generated function ϕ. All the algorithms were coded in ANSI C, and the experiments were conducted on an Intel PIV 1.6GHz machine with Microsoft windows XP professional.

We run each of the three aforementioned methods with various sample sizes on every one of the instances. The results are reported in Table 4.2.

In order to evaluate stability of the algorithms, we run each of them 100 times; the resulting statistics are shown in Table 4.3. In this relatively time-consuming experiment, we restrict ourselves with a single instance (L2) and just two sample sizes (N = 1,000 and 2,000). In Table 4.3, "Mean" and "Dev" are, respectively, the mean and the deviation, over 100 runs, of the objective value Obj at the resulting approximate solution.

The experiments demonstrate that as far as the quality of approximate solutions is concerned, N-SA outperforms E-SA and is almost as good as SAA. At the same time, the solution time for N-SA is significantly smaller than the one for SAA.

4.3. Stochastic max-flow problem. In the second experiment, we consider simple two-stage stochastic linear programming, namely, a stochastic max-flow problem. The problem is to optimize the capacity expansion of a stochastic network. Let

TABLE 4.4
SA versus SAA on the stochastic max-flow problem.

-		F1		F2		F3		F4	
(m,n)		(50,500)		(100, 1,000)		(100, 2,000)		(250, 5,000)	
ALG.	N	Obj	CPU	Obj	CPU	Obj	CPU	Obj	CPU
N-SA	100	0.1140	0	0.0637	0	0.1296	1	0.1278	3
	1,000	0.1254	1	0.0686	3	0.1305	6	0.1329	15
	2,000	0.1249	3	0.0697	6	0.1318	11	0.1338	29
	4,000	0.1246	5	0.0698	11	0.1331	21	0.1334	56
E-SA	100	0.0840	0	0.0618	1	0.1277	2	0.1153	7
	1,000	0.1253	3	0.0670	6	0.1281	16	0.1312	39
	2,000	0.1246	5	0.0695	13	0.1287	28	0.1312	72
	4,000	0.1247	9	0.0696	24	0.1303	53	0.1310	127
SAA	100	0.1212	5	0.0653	12	0.1310	20	0.1253	60
	1,000	0.1223	35	0.0694	84	0.1294	157	0.1291	466
	2,000	0.1223	70	0.0693	170	0.1304	311	0.1284	986
	4,000	0.1221	140	0.0693	323	0.1301	636	0.1293	1,885

$G = (N, A)$ be a diagraph with a source node s and a sink node t. Each arc $(i, j) \in A$ has an existing capacity $p_{ij} \geq 0$ and a random implementing/operating level ξ_{ij}. Moreover, there is a common random degrading factor η for all arcs in A. The goal is to determine how much capacity to add to the arcs, subject to a budget constraint, in order to maximize the expected maximum flow from s to t. Denoting by x_{ij} the capacity to be added to arc (i, j), the problem reads

$$\max_x \left\{ f(x) = \mathbb{E}[F(x;\xi,\eta)] : \sum_{(i,j)\in A} c_{ij}x_{ij} \leq b, x_{ij} \geq 0, \quad \forall (i,j) \in A \right\}, \tag{4.2}$$

where c_{ij} is the per unit cost for the capacity to be added, b is the total available budget, and $F(x;\xi,\eta)$ denotes the maximum $s-t$ flow in the network when the capacity of an arc (i, j) is $\eta\xi_{ij}(p_{ij} + x_{ij})$. Note that the above is a maximization rather than a minimization problem.

We assume that the random variables ξ_{ij}, θ are independent and uniformly distributed on $[0, 1]$ and $[0.5, 1]$, respectively, and consider the case of $p_{ij} = 0$, $c_{ij} = 1$ for all $(i, j) \in E$, and $b = 1$. We randomly generated 4 network instances (referred to as F1, F2, F3, and F4) using the network generator GRIDGEN available on DIMACS challenge. The push-relabel algorithm [8] was used to solve the second stage max-flow problem.

In the first test, each algorithm (N-SA, E-SA, SAA) was run once at each test instance; the results are reported in Table 4.4, where m, n stand for the number of nodes, respectively, arcs in G. Similar to the stochastic utility problem, we investigate the stability of the methods by running each of them 100 times. The resulting statistics is presented in Table 4.5, whose columns have exactly the same meaning as in Table 4.3.

This experiment fully supports the conclusions on the methods suggested by the experiments with the utility problem.

4.4. A network planning problem with random demand. In the last experiment, we consider the so-called SSN problem of Sen, Doverspike, and Cosares [24]. This problem arises in telecommunications network design where the owner of the network sells private-line services between pairs of nodes in the network, and the demands are treated as random variables based on the historical demand patterns.

TABLE 4.5
The variability for the stochastic max-flow problem.

-		N-SA			E-SA			SAA		
		Obj		Avg.	Obj		Avg.	Obj		Avg.
Inst	N	Mean	Dev	CPU	Mean	Dev	CPU	Mean	Dev	CPU
F2	1,000	0.0691	0.0004	3.11	0.0688	0.0006	4.62	0.0694	0.0003	90.15
F2	2,000	0.0694	0.0003	6.07	0.0692	0.0002	6.91	0.0695	0.0003	170.45

The optimization problem is to decide where to add capacity to the network to minimize the expected rate of unsatisfied demands. Since this problem has been studied by several authors (see, e.g., [12, 24]), it could be interesting to compare the results. Another purpose of this experiment is to investigate the behavior of the SA method when the Latin hyperplane sampling (LHS) variance reduction technique (introduced in [14]) is applied.

The problem has been formulated as a two-stage stochastic linear programming as follows:

$$\min_x \left\{ f(x) = \mathbb{E}[F(x,\xi)] : x \geq 0, \sum_i x_i = b \right\}, \tag{4.3}$$

where x is the vector of capacities to be added to the arcs of the network, b (the budget) is the total amount of capacity to be added, ξ denotes the random demand, and $F(x,\xi)$ represents the number of unserved requests, specifically,

$$F(x,\xi) = \min_{s,f} \left\{ \sum_i s_i : \begin{array}{l} \sum_i \sum_{r \in R(i)} A_r f_{ir} \leq x + c \\ \sum_{r \in R(i)} f_{ir} + s_i = \xi^i, \quad \forall i \\ f_{ir} \geq 0, s_i \geq 0, \quad \forall i, r \in R(i) \end{array} \right\}. \tag{4.4}$$

Here,

- $R(i)$ is the set of routes used for traffic i (traffic between the source-sink pair of nodes # i),
- ξ^i is the (random) demand for traffic i,
- A_r are the route-arc incidence vectors (so that jth component of A_r is 1 or 0 depending on whether arc j belongs to the route r),
- c is the vector of current capacities, f_{ir} is the fraction of traffic i transferred via route r, and s is the vector of unsatisfied demands.

In the SSN instance, there are $\dim x = 89$ arcs and $\dim \xi = 86$ source-sink pairs, and components of ξ are independent random variables with known discrete distributions (from 3 to 7 possible values per component), which result in $\approx 10^{70}$ possible demand scenarios.

In the first test with the SSN instance, each of our 3 algorithms was run once without and once with the LHS technique; the results are reported in Table 4.6. We then tested the stability of algorithms by running each of them 100 times; see statistics in Table 4.7. Note that experiments with the SSN problem were conducted on a more powerful computer: Intel Xeon 1.86GHz with Red Hat Enterprize Linux.

As far as comparison of our three algorithms is concerned, the conclusions are in full agreement with those for the utility and the max-flow problem. We also see that for our particular example, the LHS does not yield much of an improvement, especially when a larger sample size is applied. This result seems to be consistent with the observation in [12].

TABLE 4.6
SA versus SAA on the SSN problem.

-		Without LHS		With LHS	
Alg.	N	Obj	CPU	Obj	CPU
N-SA	100	11.0984	1	10.1024	1
	1,000	10.0821	6	10.0313	7
	2,000	9.9812	12	9.9936	12
	4,000	9.9151	23	9.9428	22
E-SA	100	10.9027	1	10.3860	1
	1,000	10.1268	6	10.0984	6
	2,000	10.0304	12	10.0552	12
	4,000	9.9662	23	9.9862	23
SAA	100	11.8915	24	11.0561	23
	1,000	10.0939	215	10.0488	216
	2,000	9.9769	431	9.9872	426
	4,000	9.8773	849	9.9051	853

TABLE 4.7
The variability for the SSN problem.

-		N-SA			E-SA			SAA		
		Obj		Avg.	Obj		Avg.	Obj		Avg.
N	LHS	Mean	Dev	CPU	Mean	Dev	CPU	Mean	Dev	CPU
1,000	no	10.0624	0.1867	6.03	10.1730	0.1826	6.12	10.1460	0.2825	215.06
1,000	yes	10.0573	0.1830	6.16	10.1237	0.1867	6.14	10.0135	0.2579	216.10
2,000	no	9.9965	0.2058	11.61	10.0853	0.1887	11.68	9.9943	0.2038	432.93
2,000	yes	9.9978	0.2579	11.71	10.0486	0.2066	11.74	9.9830	0.1872	436.94

4.5. N-SA versus E-SA. The data in Tables 4.3, 4.4, and 4.6 demonstrate that with the same sample size N, the N-SA somehow outperforms the E-SA in terms of both the quality of approximate solutions and the running time.[4] The difference in solutions' quality, at the first glance, seems slim, and one could think that adjusting the SA algorithm to the "geometry" of the problem in question (in our case, to minimization over a standard simplex) is of minor importance. We, however, do believe that such a conclusion would be wrong. In order to get a better insight, let us come back to the stochastic utility problem. This test problem has an important advantage—we can easily compute the value of the objective $f(x)$ at a given candidate solution x analytically.[5] Moreover, it is easy to minimize $f(x)$ over the simplex—on a closest inspection, this problem reduces to minimizing an easy-to-compute *univariate* convex function so that we can approximate the true optimal value f_* to high accuracy by bisection. Thus, in the case in question, we can compare solutions x generated by various algorithms in terms of their "true inaccuracy" $f(x) - f_*$, and this is the rationale behind our "Gaussian setup." We can now exploit this advantage of the stochastic utility problem for comparing properly N-SA and E-SA. In Table 4.8, we present the true values of the objective $f(\bar{x})$ at the approximate solutions $\bar{x}$ generated by N-SA and E-SA as applied to the instances L1 and L4 of the utility problem (cf. Table 4.3) along with the inaccuracies $f(\bar{x}) - f_*$ and the Monte Carlo estimates $\widehat{f}(\bar{x})$ of $f(\bar{x})$ obtained via 50,000-element samples. We see that the difference in

[4]The difference in running times can be easily explained: with X being a simplex, the prox-mapping for E-SA takes $O(n \ln n)$ operations versus $O(n)$ operations for N-SA.

[5]Indeed, $(\xi_1, \ldots, \xi_n) \sim \mathcal{N}(0, I_n)$, so that the random variable $\xi_x = \sum_i (a_i + \xi_i) x_i$ is normal with easily computable mean and variance, and since ϕ is piecewise linear, the expectation $f(x) = \mathbb{E}[\phi(\xi_x)]$ can be immediately expressed via the error function.

TABLE 4.8
N-SA versus E-SA.

Method	Problem	$\widehat{f}(\bar{x}), f(\bar{x})$	$f(\bar{x}) - f_*$	Time
N-SA, $N = 2{,}000$	L2: $n = 1{,}000$	$-5.9232/-5.9326$	0.0113	5.00
E-SA, $N = 2{,}000$	L2	$-5.8796/-5.8864$	0.0575	6.60
E-SA, $N = 10{,}000$	L2	$-5.9059/-5.9058$	0.0381	39.80
E-SA, $N = 20{,}000$	L2	$-5.9151/-5.9158$	0.0281	74.50
N-SA, $N = 2{,}000$	L4: $n = 5{,}000$	$-5.5855/-5.5867$	0.0199	25.00
E-SA, $N = 2{,}000$	L4	$-5.5467/-5.5469$	0.0597	44.60
E-SA, $N = 10{,}000$	L4	$-5.5810/-5.5812$	0.0254	165.10
E-SA, $N = 20{,}000$	L4	$-5.5901/-5.5902$	0.0164	382.00

the inaccuracy $f(\bar{x}) - f_*$ of the solutions produced by the algorithms is much more significant than is suggested by the data in Table 4.3 (where the actual inaccuracy is "obscured" by the estimation error and summation with f_*). Specifically, at the common for both algorithm sample sizes $N = 2{,}000$, the inaccuracy yielded by N-SA is 3–5 times less than the one for E-SA and in order to compensate for this difference, one should increase the sample size for E-SA (and hence the running time) by factor 5–10. It should be added that in light of theoretical complexity analysis carried out in Example 2.3, the outlined significant difference in performances of N-SA and E-SA is not surprising; the surprising fact is that E-SA works at all.

4.6. Bilinear matrix game. We consider here a bilinear matrix game

$$\min_{x\in X}\max_{y\in Y} y^T A x,$$

where both feasible sets are the standard simplexes in $\mathbb{R}^n$: $Y = X = \{x \in \mathbb{R}^n : \sum_{i=1}^n x_i = 1, x \geq 0\}$. We consider two versions of the randomized mirror SA algorithm (3.7), (3.8) for the saddle point problem. The first algorithm, the E-SA, uses $\frac{1}{2}\|x\|_2^2$ as ω_x, ω_y and $\|\cdot\|_2$ as $\|\cdot\|_x$, $\|\cdot\|_y$. The second algorithm, the N-SA, uses the entropy function (2.58) as ω_x, ω_y and the norm $\|\cdot\|_1$ as $\|\cdot\|_x$, $\|\cdot\|_y$. To compare the two procedures, we compute the corresponding approximate solutions $\tilde{z}_1^N$ and compute the exact values of the error:

$$\epsilon\left(\tilde{z}_1^N\right) = \max_{y\in Y} y^T A\tilde{x}_1^N - \min_{x\in X}\left[\tilde{y}_1^N\right]^T Ax, \quad i = 1, 2.$$

In our experiments we consider symmetric matrices A of two kinds. The matrices of the first family, parameterized by $\alpha > 0$, are given by

$$A_{ij} = \left(\frac{i+j-1}{2n-1}\right)^{\alpha}, \ \ 1 \leq i, j \leq n.$$

The second family of matrices, which are also parameterized by $\alpha > 0$, is given by

$$A_{ij} = \left(\frac{|i-j|+1}{2n-1}\right)^{\alpha}, \ \ 1 \leq i, j \leq n.$$

We use the notations $E_1(\alpha)$ and $E_2(\alpha)$ to refer to the experiments with the matrices of the first and second kind with parameter α. We present in Table 4.9 the results of experiments conducted for the matrices A of size $10^4 \times 10^4$. We made 100 simulation runs in each experiment and present the average error (column Mean), standard

TABLE 4.9
SA for bilinear matrix games.

	$E_2(2)$, $\epsilon(\tilde{z}_1) = 0.500$			$E_2(1)$, $\epsilon(\tilde{z}_1) = 0.500$			$E_2(0.5)$, $\epsilon(\tilde{z}_1) = 0.390$		
N-SA	$\epsilon\left(\tilde{z}_1^N\right)$		CPU	$\epsilon\left(\tilde{z}_1^N\right)$		CPU	$\epsilon\left(\tilde{z}_1^N\right)$		CPU
N	Mean	Dev	CPU	Mean	Dev	CPU	Mean	Dev	CPU
100	0.0121	3.9 e−4	0.58	0.0127	1.9 e−4	0.69	0.0122	4.3 e−4	0.81
1,000	0.00228	3.7 e−5	5.8	0.00257	2.2 e−5	7.3	0.00271	4.5 e−5	8.5
2,000	0.00145	2.1 e−5	11.6	0.00166	1.0 e−5	13.8	0.00179	2.7 e−5	16.4
E-SA	$\epsilon\left(\tilde{z}_1^N\right)$		CPU	$\epsilon\left(\tilde{z}_1^N\right)$		CPU	$\epsilon\left(\tilde{z}_1^N\right)$		CPU
N	Mean	DEV	(Avg.)	Mean	Dev	(Avg.)	Mean	Dev	(Avg.)
100	0.00952	1.0 e−4	1.27	0.0102	5.1 e−5	1.77	0.00891	1.1 e−4	1.94
1,000	0.00274	1.3 e−5	11.3	0.00328	7.8 e−6	17.6	0.00309	1.6 e−5	20.9
2,000	0.00210	7.4 e−6	39.7	0.00256	4.6 e−6	36.7	0.00245	7.8 e−6	39.2
	$E_1(2)$, $\epsilon(\tilde{z}_1) = 0.0625$			$E_1(1)$, $\epsilon(\tilde{z}_1) = 0.125$			$E_1(0.5)$, $\epsilon(\tilde{z}_1) = 0.138$		
N-SA	$\epsilon\left(\tilde{z}_1^N\right)$		CPU	$\epsilon\left(\tilde{z}_1^N\right)$		CPU	$\epsilon\left(\tilde{z}_1^N\right)$		CPU
N	Mean	Dev	(Avg.)	Mean	Dev	(Avg.)	Mean	Dev	(Avg.)
100	0.00817	0.0016	0.58	0.0368	0.0068	0.66	0.0529	0.0091	0.78
1,000	0.00130	2.7 e−4	6.2	0.0115	0.0024	6.5	0.0191	0.0033	7.6
2,000	0.00076	1.6 e−4	11.4	0.00840	0.0014	11.7	0.0136	0.0018	13.8
E-SA	$\epsilon\left(\tilde{z}_1^N\right)$		CPU	$\epsilon\left(\tilde{z}_1^N\right)$		CPU	$\epsilon\left(\tilde{z}_1^N\right)$		CPU
N	Mean	Dev	(Avg.)	Mean	Dev	(Avg.)	Mean	Dev	(Avg.)
100	0.00768	0.0012	1.75	0.0377	0.0062	2.05	0.0546	0.0064	2.74
1,000	0.00127	2.2 e−4	17.2	0.0125	0.0022	19.9	0.0207	0.0020	18.4
2,000	0.00079	1.6 e−4	35.0	0.00885	0.0015	36.3	0.0149	0.0020	36.7

deviation (column Dev) and the average running time (with excluded time to compute the error of the resulting solution). For comparison, we also present the error of the initial solution $\tilde{z}_1 = (x_1, y_1)$.

Our basic observation is as follows: Both N-SA and E-SA succeed to reduce the solution error reasonably fast. The N-SA implementation is preferable as it is more efficient in terms of running time. For comparison, it takes MATLAB from 10 (for the simplest problem) to 35 seconds (for the hardest one) to compute just one answer $\mathtt{g}(x, y) = \begin{bmatrix} A^T y \\ -Ax \end{bmatrix}$ of the deterministic oracle.

5. Conclusions. It is shown that for a certain class of convex stochastic optimization and saddle point problems, robust versions of the SA approach have similar theoretical estimates of computational complexity, in terms of the required sample size, to the SAA method. Numerical experiments, reported in section 4, confirm this conclusion. These results demonstrate that for considered problems, a properly implemented mirror descent SA algorithm produces solutions of comparable accuracy to the SAA method for the same sample size of generated random points. On the other hand, the implementation (computational) time of the SA method is significantly smaller with a factor of up to 30–40 for considered problems. Thus, both theoretical and numerical results suggest that the robust mirror descent SA is a viable alternative to the SAA approach, an alternative which at least deserves testing in particular applications. It is also shown that the robust mirror SA approach can be applied as a randomization algorithm to large-scale deterministic saddle point problems (in particular, to minimax optimization problems and bilinear matrix games) with encouraging results.

6. Appendix. *Proof of Lemma* 2.1. Let $x \in X^o$ and $v = P_x(y)$. Note that $v \in \operatorname{argmin}_{z \in X}[\omega(z) + p^T z]$, where $p = \nabla\omega(x) - y$. Thus, ω is differentiable at v and $v \in$

X^o. As $\nabla_v V(x,v) = \nabla\omega(v) - \nabla\omega(x)$, the optimality conditions for (2.31) imply that

$$(\nabla\omega(v) - \nabla\omega(x) + y)^T(v-u) \le 0 \quad \forall u \in X. \tag{6.1}$$

For $u \in X$, we therefore have

$$\begin{aligned} V(v,u) - V(x,u) &= \big[\omega(u) - \nabla\omega(v)^T(u-v) - \omega(v)\big] \\ &\quad - \big[\omega(u) - \nabla\omega(x)^T(u-x) - \omega(x)\big] \\ &= \nabla\omega(v) - \nabla\omega(x) + y)^T(v-u) + y^T(u-v) \\ &\le y^T(u-v) - V(x,v), \end{aligned}$$

where the last inequality is due to (6.1). By Young's inequality,[6] we have

$$y^T(x-v) \le \frac{\|y\|_*^2}{2\alpha} + \frac{\alpha}{2}\|x-v\|^2,$$

while $V(x,v) \ge \frac{\alpha}{2}\|x-v\|^2$, due to the strong convexity of $V(x,\cdot)$. We get

$$\begin{aligned} V(v,u) - V(x,u) &\le y^T(u-v) - V(x,v) = y^T(u-x) + y^T(x-v) - V(x,v) \\ &\le y^T(u-x) + \frac{\|y\|_*^2}{2\alpha}, \end{aligned}$$

as required in (2.33). □

Entropy as a distance-generating function on the standard simplex. The only property which is not immediately evident is that the entropy $w(x) = \sum_{i=1}^n x_i \ln x_i$ is strongly convex, modulus 1 with respect to $\|\cdot\|_1$-norm, on the standard simplex $X = \big\{x \in \mathbb{R}^n : x \ge 0,\ \sum_{i=1}^n x_i\big\}$. We are in the situation where $X^o = \{x \in X : x > 0\}$ and in order to establish the property in question, it suffices to verify that $h^T\nabla^2\omega(x)h \ge \|h\|_1^2$ for every $x \in X^o$. Here is the computation:

$$\left[\sum_i |h_i|\right]^2 = \left[\sum_i (x_i^{-1/2}|h_i|)x_i^{1/2}\right]^2 \le \left[\sum_i h_i^2 x_i^{-1}\right]\left[\sum_i x_i\right] = \sum_i h_i^2 x_i^{-1} = h^T\nabla^2\omega(x)h,$$

where the inequality follows by Cauchy inequality.

Proof of Lemma 3.1. By (2.33), we have, for any $u \in Z$, that

$$\gamma_t(z_t - u)^T\mathsf{G}(z_t,\xi_t) \le V(z_t,u) - V(z_{t+1},u) + \frac{\gamma_t^2}{2}\|\mathsf{G}(z_t,\xi_t)\|_*^2 \tag{6.2}$$

(recall that we are in the situation of $\alpha = 1$). This relation implies that for every $u \in Z$, one has

$$\gamma_t(z_t - u)^T\mathsf{g}(z_t) \le V(z_t,u) - V(z_{t+1},u) + \frac{\gamma_t^2}{2}\|\mathsf{G}(z_t,\xi_t)\|_*^2 - \gamma_t(z_t-u)^T\Delta_t, \tag{6.3}$$

where $\Delta_t = \mathsf{G}(z_t,\xi_t) - \mathsf{g}(z_t)$. Summing up these inequalities over $t = 1,\dots,j$, we get

$$\sum_{t=1}^j \gamma_t(z_t-u)^T\mathsf{g}(z_t) \le V(z_1,u) - V(z_{t+1},u) + \sum_{t=1}^j \frac{\gamma_t^2}{2}\|\mathsf{G}(z_t,\xi_t)\|_*^2 - \sum_{t=1}^j \gamma_t(z_t-u)^T\Delta_t.$$

Now we need the following simple lemma.

[6] For any $u, v \in \mathbb{R}^n$, we have, by the definition of the dual norm, that $\|u\|_*\|v\| \ge u^T v$, and hence $(\|u\|_*^2/\alpha + \alpha\|v\|^2)/2 \ge \|u\|_*\|v\| \ge u^T v$.

LEMMA 6.1. *Let $\zeta_1, \dots, \zeta_j$ be a sequence of elements of $\mathbb{R}^{n+m}$. Define the sequence v_t, $t = 1, 2, \dots$ in Z^o as follows: $v_1 \in Z^o$ and*

$$v_{t+1} = P_{v_t}(\zeta_t),\ 1 \le t \le j.$$

Then, for any $u \in Z$, the following holds:

$$\sum_{t=1}^{j} \zeta_t^T (v_t - u) \le V(v_1, u) + \tfrac{1}{2} \sum_{t=1}^{j} \|\zeta_t\|_*^2. \tag{6.4}$$

Proof. Using the bound (2.33) of Lemma 2.1 with $y = \zeta_t$ and $x = v_t$ (so that $v_{t+1} = P_{v_t}(\zeta_t)$) and recalling that we are in the situation of $\alpha = 1$, we obtain the following for any $u \in Z$:

$$V(v_{t+1}, u) \le V(v_t, u) + \zeta_t^T (u - v_t) + \tfrac{1}{2} \|\zeta_t\|_*^2.$$

Summing up from $t = 1$ to $t = j$, we conclude that

$$V(v_{j+1}, u) \le V(v_1, u) + \sum_{t=1}^{j} \zeta_t^T (u - v_t) + \tfrac{1}{2} \sum_{t=1}^{j} \|\zeta_t\|_*^2,$$

which implies (6.4) due to $V(v, u) \ge 0$ for any $v \in Z^o, u \in Z$. □

Applying Lemma 6.1 with $v_1 = z_1$, $\zeta_t = -\gamma_t \Delta_t$, we get

$$\sum_{t=1}^{j} \gamma_t \Delta_t^T (u - v_t) \le V(z_1, u) + \frac{1}{2} \sum_{t=1}^{j} \gamma_t^2 \|\Delta_t\|_*^2 \ \ \forall u \in Z. \tag{6.5}$$

Observe that

$$\mathbb{E}\|\Delta_t\|_*^2 \le 4\mathbb{E}\|\mathsf{G}(z_t, \xi_t)\|_*^2 \le 4 \left(\frac{2D_{\omega_x, X}^2}{\alpha_x} M_{*,x}^2 + \frac{2D_{\omega_y, Y}^2}{\alpha_y} M_{*,y}^2 \right) = 4M_*^2$$

so that when taking the expectation of both sides of (6.5), we get

$$\mathbb{E}\left[\sup_{u \in Z} \left\{ \sum_{t=1}^{j} \gamma_t \Delta_t^T (u - v_t) \right\} \right] \le 1 + 2M_*^2 \sum_{t=1}^{j} \gamma_t^2 \tag{6.6}$$

(recall that $V(z_1, \cdot)$ is bounded by 1 on Z). Now we proceed exactly as in section 2.2: we sum up (6.3) from $t = 1$ to j to obtain

$$\begin{aligned} &\sum_{t=1}^{j} \gamma_t (z_t - u)^T \mathsf{g}(z_t) \le V(z_1, u) + \sum_{t=1}^{j} \frac{\gamma_t^2}{2} \|\mathsf{G}(z_t, \xi_t)\|_*^2 - \sum_{t=1}^{j} \gamma_t (z_t - u)^T \Delta_t \\ &= V(z_1, u) + \sum_{t=1}^{j} \frac{\gamma_t^2}{2} \|\mathsf{G}(z_t, \xi_t)\|_*^2 - \sum_{t=1}^{j} \gamma_t (z_t - v_t)^T \Delta_t + \sum_{t=1}^{j} \gamma_t (u - v_t)^T \Delta_t. \end{aligned} \tag{6.7}$$

When taking into account that z_t and v_t are deterministic functions of $\xi_{[t-1]} = (\xi_1, \dots, \xi_{t-1})$ and that the conditional expectation of Δ_t, $\xi_{[t-1]}$ being given, vanishes, we conclude that $\mathbb{E}[(z_t - v_t)^T \Delta_t] = 0$. We take now suprema in $u \in Z$ and then

expectations on both sides of (6.7):

$$\begin{aligned}\mathbb{E}\left[\sup_{u\in Z}\sum_{t=1}^{j}\gamma_t(z_t-u)^T\mathsf{g}(z_t)\right] &\le \sup_{u\in Z}V(z_1,u)+\sum_{t=1}^{j}\frac{\gamma_t^2}{2}\mathbb{E}\|\mathsf{G}(z_t,\xi_t)\|_*^2\\ &\quad+\sup_{u\in Z}\sum_{t=1}^{j}\gamma_t(u-v_t)^T\Delta_t\\ \text{(by (6.6))}\quad &\le 1+\frac{M_*^2}{2}\sum_{t=1}^{j}\gamma_t^2+\left[1+2M_*^2\sum_{t=1}^{j}\gamma_t^2\right]\\ &= 2+\frac{5}{2}M_*^2\sum_{t=1}^{j}\gamma_t^2,\end{aligned}$$

and we arrive at (3.10).

Proof of Propositions 2.2 and 3.2. We provide here the proof of Proposition 3.2 only. The proof of Proposition 2.2 follows the same lines and can be easily reconstructed using the bound (2.39) instead of the relations (6.5) and (6.7) in the proof below.

First of all, with M_* given by (3.6), one has

$$\forall(z\in Z):\mathbb{E}\left[\exp\{\|\mathsf{G}(z,\xi)\|_*^2/M_*^2\}\right]\le\exp\{1\}. \tag{6.8}$$

Indeed, setting $p_x=\frac{2D^2_{\omega_x,X}M^2_{*,x}}{\alpha_x M_*^2}$, $p_y=\frac{2D^2_{\omega_y,Y}M^2_{*,y}}{\alpha_y M_*^2}$, we have $p_x+p_y=1$, whence, invoking (3.4),

$$\mathbb{E}\left[\exp\{\|\mathsf{G}(z,\xi)\|_*^2/M_*^2\}\right]=\mathbb{E}\left[\exp\{p_x\|\mathsf{G}_x(z,\xi)\|_{*,x}^2/M_{*,x}^2+p_y\|\mathsf{G}_y(z,\xi)\|_{*,y}^2/M_{*,y}^2\}\right],$$

and (6.8) follows from (3.16) by the Hölder inequality.

Setting $\Gamma_N=\sum_{t=1}^N\gamma_t$ and using the notation from the proof of Lemma 3.1, relations (3.9), (6.5), and (6.7) combined with the fact that $V(z_1,u)\le 1$ for $u\in Z$, imply that

$$\Gamma_N\epsilon_\phi(\tilde z_N)\le 2+\underbrace{\tfrac12\sum_{t=1}^{N}\gamma_t^2\left[\|\mathsf{G}(z_t,\xi_t)\|_*^2+\|\Delta_t\|_*^2\right]}_{\alpha_N}+\underbrace{\sum_{t=1}^{N}\gamma_t(v_t-z_t)^T\Delta_t}_{\beta_N}. \tag{6.9}$$

Now, from (6.8), it follows straightforwardly that

$$\mathbb{E}\left[\exp\left\{\|\Delta_t\|_*^2/(2M_*)^2\right\}\right]\le\exp\{1\},\quad \mathbb{E}\left[\exp\left\{\|\mathsf{G}(z_t,\xi_t)\|_*^2/M_*^2\right\}\right]\le\exp\{1\}, \tag{6.10}$$

which, in turn, implies that

$$\mathbb{E}[\exp\{\alpha_N/\sigma_\alpha\}]\le\exp\{1\},\quad \sigma_\alpha=\frac{5}{2}M_*^2\sum_{t=1}^{N}\gamma_t^2, \tag{6.11}$$

and therefore, by Markov inequality, for any $\Omega>0$,

$$\mathrm{Prob}\{\alpha_N\ge(1+\Omega)\sigma_\alpha\}\le\exp\{-\Omega\}. \tag{6.12}$$

Indeed, we have by (6.8)

$$\|\mathsf{g}(z_t)\|_*=\|\mathbb{E}[\mathsf{G}(z_t,\xi_t)|\xi_{[t-1]}]\|_*\le\sqrt{\mathbb{E}(\|\mathsf{G}(z_t,\xi_t)\|_*^2|\xi_{[t-1]})}\le M_*$$

and

$$\|\Delta_t\|_*^2 = \|\mathsf{G}(z_t,\xi_t) - \mathsf{g}(z_t)\|_*^2 \le (\|\mathsf{G}(z_t,\xi_t)\|_* + \|\mathsf{g}(z_t)\|_*)^2 \le 2\|\mathsf{G}(z_t,\xi_t)\|_*^2 + 2M_*^2,$$

which implies that

$$\alpha_N \le \sum_{t=1}^N \frac{\gamma_t^2}{2}\left[3\|\mathsf{G}(z_t,\xi_t)\|_*^2 + 2M_*^2\right].$$

Further, by the Hölder inequality, we have the following from (6.8):

$$\mathbb{E}\left[\exp\left\{\frac{\gamma_t^2\left[\frac{3}{2}\|\mathsf{G}(z_t,\xi_t)\|_*^2 + M_*^2\right]}{\frac{5}{2}\gamma_t^2 M_*^2}\right\}\right] \le \exp(1).$$

Observe that if $r_1,\dots,r_i$ are nonnegative random variables such that $\mathbb{E}[\exp\{r_t/\sigma_t\}] \le \exp\{1\}$ for some deterministic $\sigma_t > 0$, then, by convexity of the exponent, $w(s) = \exp\{s\}$ and

$$\mathbb{E}\left[\exp\left\{\frac{\sum_{t\le i} r_t}{\sum_{t\le i}\sigma_t}\right\}\right] \le \mathbb{E}\left[\sum_{t\le i}\frac{\sigma_t}{\sum_{\tau\le i}\sigma_\tau}\exp\{r_t/\sigma_t\}\right] \le \exp\{1\}. \tag{6.13}$$

Now applying (6.13) with $r_t = \gamma_t^2\left[\frac{3}{2}\|\mathsf{G}(z_t,\xi_t)\|_*^2 + M_*^2\right]$ and $\sigma_t = \frac{5}{2}\gamma_t^2 M_*^2$, we obtain (6.11).

Now let $\zeta_t = \gamma_t(v_t - z_t)^T\Delta_t$. Observing that v_t, z_t are deterministic functions of $\xi_{[t-1]}$, while $\mathbb{E}[\Delta_t|\xi_{[t-1]}] = 0$, we see that the sequence $\{\zeta_t\}_{t=1}^N$ of random real variables forms a martingale difference. Besides this, by strong convexity of ω with modulus 1 w.r.t. $\|\cdot\|$ and due to $D_{\omega,Z} \le 1$, we have

$$u \in Z \Rightarrow 1 \ge V(z_1,u) \ge \frac{1}{2}\|u - z_1\|^2,$$

whence the $\|\cdot\|$-diameter of Z does not exceed $2\sqrt{2}$ so that $|\zeta_t| \le 2\sqrt{2}\gamma_t\|\Delta_t\|_*$, and therefore

$$\mathbb{E}\left[\exp\left\{|\zeta_t|^2/\left(32\gamma_t^2M_*^2\right)\right\}|\xi_{[t-1]}\right] \le \exp\{1\}$$

by (6.10). Applying Cramer's deviation bound, we obtain, for any $\Omega > 0$,

$$\text{Prob}\left\{\beta_N > 4\Omega M_*\sqrt{\sum_{t=1}^N \gamma_t^2}\right\} \le \exp\{-\Omega^2/4\}. \tag{6.14}$$

Indeed, for $0 \le \gamma$, setting $\sigma_t = 4\sqrt{2}\gamma_t M_*$ and taking into account that ζ_t is a deterministic function of $\xi_{[t]}$, with $\mathbb{E}[\zeta_t|\xi_{[t-1]}] = 0$ and $\mathbb{E}[\exp\{\zeta_t^2/\sigma_t^2\}|\xi_{[t-1]}] \le \exp\{1\}$, we have

$$\begin{aligned}
&0 < \gamma\sigma_t \le 1 \Rightarrow \quad \left(\text{as } e^x \le x + e^{x^2}\right)\\
&\mathbb{E}[\exp\{\gamma\zeta_t\}|\xi_{[t-1]}] \le \mathbb{E}[\exp\left\{\gamma^2\zeta_t^2\right\}|\xi_{[t-1]}]\\
&\qquad\qquad \le \mathbb{E}\left[\left(\exp\left\{\zeta_t^2/\sigma_t^2\right\}\right)^{\gamma^2\sigma_t^2}|\xi_{[t-1]}\right] \le \exp\left\{\gamma^2\sigma_t^2\right\};\\
&\gamma\sigma_t > 1 \Rightarrow\\
&\mathbb{E}[\exp\{\gamma\zeta_t\}|\xi_{[t-1]}] \le \mathbb{E}\left[\exp\left\{\left[\tfrac{1}{2}\gamma^2\sigma_t^2 + \tfrac{1}{2}\zeta_t^2/\sigma_t^2\right\}|\xi_{[t-1]}\right]\\
&\qquad\qquad \le \exp\left\{\tfrac{1}{2}\gamma^2\sigma_t^2 + \tfrac{1}{2}\right\} \le \exp\left\{\gamma^2\sigma_t^2\right\},
\end{aligned}$$

that is, in both cases, $\mathbb{E}[\exp\{\gamma\zeta_t\}|\xi_{[t-1]}] \le \exp\{\gamma^2\sigma_t^2\}$. Therefore,

$$\mathbb{E}[\exp\{\gamma\beta_i\}] = \mathbb{E}\left[\exp\{\gamma\beta_{i-1}\}\mathbb{E}[\exp\{\gamma\zeta_i\}|\xi_{[i-1]}]\right] \le \exp\left\{\gamma^2\sigma_i^2\right\}\mathbb{E}[\exp\{\gamma\beta_{i-1}\}],$$

whence $\mathbb{E}[\exp\{\gamma\beta_N\}] \le \exp\{\gamma^2\sum_{t=1}^N\sigma_t^2\}$, and thus, by Markov inequality for every $\Omega > 0$, it holds

$$\mathrm{Prob}\left\{\beta_N > \Omega\sqrt{\sum_{t=1}^N\sigma_t^2}\right\} \le \exp\left\{\gamma^2\sum_{t=1}^N\sigma_t^2\right\}\exp\left\{-\gamma\Omega\sqrt{\sum_{t=1}^N\sigma_t^2}\right\}.$$

When choosing $\gamma = \frac{1}{2}\Omega\left(\sum_{t=1}^N\sigma_t^2\right)^{-1/2}$, we arrive at (6.14).

Combining (6.9), (6.10), and (6.14), we get the following for any positive Ω and Θ:

$$\mathrm{Prob}\left\{\Gamma_N\,\epsilon_\phi(\tilde z_t) > 2 + \frac{5}{2}(1+\Omega)M_*^2\sum_{t=1}^N\gamma_t^2 + 4\sqrt{2}\Theta M_*\sqrt{\sum_{t=1}^N\gamma_t^2}\right\} \le \exp\{-\Omega\} + \exp\left\{-\frac{1}{4}\Theta^2\right\}.$$

When setting $\Theta = 2\sqrt{\Omega}$ and substituting (3.12), we obtain (3.17). □

Proof of Proposition 3.3. As in the proof of Proposition 3.2, when setting $\Gamma_N = \sum_{t=1}^N\gamma_t$ and using the relations (3.9), (6.5), and (6.7), combined with the fact that $\|\mathsf{G}(z,\xi_y)\|_* \le M_*$, we obtain

$$\begin{aligned}\Gamma_N\,\epsilon_\phi(\tilde z_N) &\le 2 + \sum_{t=1}^N\frac{\gamma_t^2}{2}\left[\|\mathsf{G}(z_t,\xi_t)\|_*^2 + \|\Delta_t\|_*^2\right] + \sum_{t=1}^N\gamma_t(v_t - z_t)^T\Delta_t \\ &\le 2 + \frac{5}{2}M_*^2\sum_{t=1}^N\gamma_t^2 + \underbrace{\sum_{t=1}^N\gamma_t(v_t - z_t)^T\Delta_t}_{\alpha_N}. \qquad (6.15)\end{aligned}$$

Recall that by definition of Δ_t, $\|\Delta_t\|_* = \|\mathsf{G}(z_t,\xi_t) - \mathsf{g}(z_t)\|_* \le \|\mathsf{G}(z_t,\xi_t)\| + \|\mathsf{g}(z_t)\|_* \le 2M_*$.

Note that $\zeta_t = \gamma_t(v_t - z_t)^T\Delta_t$ is a bounded martingale difference, i.e., $\mathbb{E}(\zeta_t|\xi_{[t-1]}) = 0$ and $|\zeta_t| \le 4\gamma_t\overline{M}$ (here $\overline{M}$ is defined in (3.31)). Then, by Azuma–Hoeffding's inequality [1] for any $\Omega \ge 0$,

$$\mathrm{Prob}\left(\alpha_N > 4\Omega\overline{M}\sqrt{\sum_{t=1}^N\gamma_t^2}\right) \le e^{-\Omega^2/2}. \qquad (6.16)$$

Indeed, let us denote $v_t = (v_t^{(x)}, v_t^{(y)})$ and $\Delta_t = (\Delta_t^{(x)}, \Delta_t^{(y)})$. When taking into account that $\|v_t^{(x)}\|_1 \le 1$, $\|v_t^{(y)}\|_1 \le 1$ and $\|x_t\|_1 \le 1$, $\|y_t\|_1 \le 1$, we conclude that

$$\begin{aligned}|(v_t - z_t)^T\Delta_t| &\le \left|\left(v_t^{(x)} - x_t\right)^T\Delta_t^{(x)}\right| + \left|\left(v_t^{(y)} - y_t\right)^T\Delta_t^{(y)}\right| \\ &\le 2\left\|\Delta_t^{(x)}\right\|_\infty + 2\left\|\Delta_t^{(y)}\right\|_\infty \le 4\max_{1\le j\le m}\|A^j + b\|_\infty + 4\max_{1\le j\le n}\|A_j + c\|_\infty \\ &= 4\overline{M}.\end{aligned}$$

We conclude from (6.15) and (6.16) that

$$\text{Prob}\left(\Gamma_N\,\epsilon_\phi(\tilde{z}_N) > 2 + \frac{5}{2}M_*^2\sum_{t=1}^N\gamma_t^2 + 4\Omega\overline{M}\sqrt{\sum_{t=1}^N\gamma_t^2}\right) \le e^{-\Omega^2/2},$$

and the bound (3.30) of the proposition can be easily obtained by substituting the constant stepsizes γ_t as defined in (3.12). □

REFERENCES

[1] K. Azuma, *Weighted sums of certain dependent random variables*, Tökuku Math. J., 19 (1967), pp. 357–367.

[2] A. Ben-Tal and A. Nemirovski, *Non-Euclidean restricted memory level method for large-scale convex optimization*, Math. Program., 102 (2005), pp. 407–456.

[3] A. Benveniste, M. Métivier, and P. Priouret, *Algorithmes Adaptatifs et Approximations Stochastiques*, Masson, Paris, 1987 (in French). Adaptive Algorithms and Stochastic Approximations, Springer, New York, 1993 (in English).

[4] L.M. Bregman, *The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming*, Comput. Math. Math. Phys., 7 (1967), pp. 200–217.

[5] K.L. Chung, *On a stochastic approximation method*, Ann. Math. Statist., 25 (1954), pp. 463–483.

[6] Y. Ermoliev, *Stochastic quasigradient methods and their application to system optimization*, Stochastics, 9 (1983), pp. 1–36.

[7] A.A. Gaivoronski, *Nonstationary stochastic programming problems*, Kybernetika, 4 (1978), pp. 89–92.

[8] A.V. Goldberg and R.E. Tarjan, *A new approach to the maximum flow problem*, J. ACM, 35 (1988), pp. 921–940.

[9] M.D. Grigoriadis and L.G. Khachiyan, *A sublinear-time randomized approximation algorithm for matrix games*, Oper. Res. Lett., 18 (1995), pp. 53–58.

[10] A. Juditsky, A. Nazin, A. Tsybakov, and N. Vayatis, *Recursive aggregation of estimators by the mirror descent algorithm with averaging*, Probl. Inf. Transm., 41 (2005), pp. 368–384.

[11] A.J. Kleywegt, A. Shapiro, and T. Homem-de-Mello, *The sample average approximation method for stochastic discrete optimization*, SIAM J. Optim., 12 (2002), pp. 479–502.

[12] J. Linderoth, A. Shapiro, and S. Wright, *The empirical behavior of sampling methods for stochastic programming*, Ann. Oper. Res., 142 (2006), pp. 215–241.

[13] W.K. Mak, D.P. Morton, and R.K. Wood, *Monte Carlo bounding techniques for determining solution quality in stochastic programs*, Oper. Res. Lett., 24 (1999), pp. 47–56.

[14] M.D. McKay, R.J. Beckman, and W.J. Conover, *A comparison of three methods for selecting values of input variables in the analysis of output from a computer code*, Technometrics, 21 (1979), pp. 239–245.

[15] A. Nemirovski and D. Yudin, *On Cezari's convergence of the steepest descent method for approximating saddle point of convex-concave functions*, Dokl. Akad. Nauk SSSR, 239 (1978) (in Russian). Soviet Math. Dokl., 19 (1978) (in English).

[16] A. Nemirovski and D. Yudin, *Problem Complexity and Method Efficiency in Optimization*, Wiley-Intersci. Ser. Discrete Math. 15, John Wiley, New York, 1983.

[17] Y. Nesterov, *Primal-dual subgradient methods for convex problems*, Math. Program., Ser. B, http://www.springerlink.com content-b441795t5254m533.

[18] B.T. Polyak, *New stochastic approximation type procedures*, Automat. i Telemekh., 7 (1990), pp. 98–107.

[19] B.T. Polyak and A.B. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM J. Control Optim., 30 (1992), pp. 838–855.

[20] G.C. Pflug, *Optimization of Stochastic Models*, The Interface Between Simulation and Optimization, Kluwer, Boston, 1996.

[21] H. Robbins and S. Monro, *A stochastic approximation method*, Ann. Math. Stat., 22 (1951), pp. 400–407.

[22] A. Ruszczyński and W. Syski, *A method of aggregate stochastic subgradients with on-line stepsize rules for convex stochastic programming problems*, Math. Prog. Stud., 28 (1986), pp. 113–131.

[23] J. Sacks, *Asymptotic distribution of stochastic approximation*, Ann. Math. Stat., 29 (1958), pp. 373–409.
[24] S. Sen, R.D. Doverspike, and S. Cosares, *Network planning with random demand*, Telecomm. Syst., 3 (1994), pp. 11–30.
[25] A. Shapiro, *Monte Carlo sampling methods*, in Stochastic Programming, Handbook in OR & MS, Vol. 10, A. Ruszczyński and A. Shapiro, eds., North-Holland, Amsterdam, 2003.
[26] A. Shapiro and A. Nemirovski, *On complexity of stochastic programming problems*, in Continuous Optimization: Current Trends and Applications, V. Jeyakumar and A.M. Rubinov, eds., Springer, New York, 2005, pp. 111–144.
[27] J.C. Spall, *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*, John Wiley, Hoboken, NJ, 2003.
[28] V. Strassen, *The existence of probability measures with given marginals*, Ann. Math. Statist., 38 (1965), pp. 423–439.
[29] B. Verweij, S. Ahmed, A.J. Kleywegt, G. Nemhauser, and A. Shapiro, *The sample average approximation method applied to stochastic routing problems: A computational study*, Comput. Optim. Appl., 24 (2003), pp. 289–333.

SIAM J. OPTIM.
Vol. 19, No. 4, pp. 1610–1632

SHAPE OPTIMIZATION UNDER UNCERTAINTY—A STOCHASTIC PROGRAMMING PERSPECTIVE*

SERGIO CONTI†, HARALD HELD†, MARTIN PACH†, MARTIN RUMPF‡, AND RÜDIGER SCHULTZ†

Abstract. We present an algorithm for shape optimization under stochastic loading and representative numerical results. Our strategy builds upon a combination of techniques from two-stage stochastic programming and level-set-based shape optimization. In particular, usage of linear elasticity and quadratic objective functions permits us to obtain a computational cost which scales linearly in the number of *linearly independent* applied forces, which often is much smaller than the number of different realizations of the stochastic forces. Numerical computations are performed using a level set method with composite finite elements both in two and in three spatial dimensions.

Key words. two-stage stochastic programming, shape optimization in elasticity, level set method

AMS subject classifications. 49N30, 74P05

DOI. 10.1137/070702059

1. Introduction.

Uncertainty is a prevailing issue in many, if not most, practical shape optimization problems. In the optimization of elastic structures, one usually deals with volume and in particular surface loadings which are not fixed but vary stochastically over time. Decisions on the shape have to be made before the stochastic forcing is applied. Thus, an optimal structure for the expectation of the stochastic loading does not properly reflect the actual stochastic optimization set up. Indeed, one observes a striking similarity with two-stage stochastic programming. Our work received inspiration from this field, and this paper is intended to work out this analogy in the case of shape optimization for linear elastic material laws and stochastic volume and surface loadings.

Optimization under uncertainty depends on information available on the uncertain problem components. At the one end, there are worst-case approaches, as in online or robust optimization [3,14]. These approaches assume that only the ranges of the uncertain parameters are known, without distributional information. At the other end, stochastic optimization deals with models where uncertainty can be captured by a probability distribution. Stochastic optimization has been analyzed in continuous time, as, for example, in stochastic dynamic programming or stochastic control [18,27]. In particular, there exists a rich theory and methodology to treat stochastic uncertainty in (mostly finite-dimensional) mathematical programming models, mainly linear [49], less often linear mixed-integer or nonlinear programming models [12, 46, 55]. In two-stage stochastic programming [16, 33, 47], first-stage decisions must be taken without knowing the realizations of the random data, and then, after observation of

*Received by the editors September 5, 2007; accepted for publication (in revised form) August 26, 2008; published electronically January 21, 2009. This work was supported by the Deutsche Forschungsgemeinschaft through the Schwerpunktprogramm 1253 *Optimization with Partial Differential Equations*.

http://www.siam.org/journals/siopt/19-4/70205.html

†Department of Mathematics, University of Duisburg-Essen, Lotharstr. 65, D-47048 Duisburg, Germany (sergio.conti@uni-bonn.de, held@math.uni-duisburg.de, pach@math.uni-duisburg.de, schultz@math.uni-duisburg.de).

‡Institute for Numerical Simulation, Rheinische Friedrich-Wilhelms-Universität Bonn, Nussallee 15, 53115 Bonn, Germany.

the random data, a second-stage (or recourse) decision is taken. The requirement that the first-stage decision must not depend on the future observation is referred to as nonanticipativity. This notion extends accordingly if the two-stage scheme of alternating decision and observation is expanded into a (finite) multistage scheme. For a recent comprehensive overview we refer to [58]. Related work on nonlinear models can be found in optimal design of structural systems under uncertainty; see [40] and references therein. The essential difference from the present work is that design decisions in these contributions vary in Euclidean spaces, while our design decisions are shapes (open sets) in suitable working domains.

Shape optimization under deterministic loading is a well-developed field, which can be seen as an instance of PDE-constrained infinite-dimensional optimization; see, e.g., the books [4, 15, 54]; a brief review of the points relevant for us is presented below. We are not aware of two- or multistage stochastic programming approaches in shape optimization, or more generally in PDE-constrained optimization. There are, however, recent approaches in shape optimization which generalize the single load assumption. In so-called multiload approaches a fixed (usually small) number of different loading configurations is considered, and optimization refers to this set of configurations; see, e.g., [7, 29, 59] and references therein, as well as [11] for an one-dimensional (1D) model. In these approaches each evaluation of the objective functional requires a separate computation for each of the possible stochastic forces, which renders them infeasible if the set of possible forces is large, as, for example, is the case when one aims at approximating a continuous distribution of forces. A more efficient method was derived for a truss model in [10], where it is shown that optimization of the expected compliance is equivalent to a convex problem and hence efficiently solvable. This, however, is based on additional geometrical assumptions, namely, on considering a fixed *ground structure*, and leaving only the thickness of the bars to be optimized. A robust probabilistic approach for the optimization of beam models is discussed in [1], whereas in [41] structural reliability is discussed for beam geometries with uncertain load magnitude. Worst-case situations in a multiload context have also been considered; see, e.g., [13].

The paper is organized as follows. In section 1.1 we formulate the stochastic shape optimization problem considered in this paper. Then in section 1.2 we review deterministic shape optimization based on a level formulation. Then in section 1.3 we recall finite-dimensional, two-stage stochastic optimization to underline the close similarity of the approach to shape optimization to be discussed here. In section 2 the two-stage shape optimization with stochastic volume and surface loads is introduced the primal and dual stochastic state equations are investigated in section 2.1, and a representation of the stochastic shape gradient is given in section 2.2. A finite element discretization for elastic domains described via level sets is discussed in section 3. In section 3.1 we introduce composite finite elements and suitable multigrid methods to apply them for the efficient solution of the discrete primal and dual problem in section 3.2, whereas in section 3.3 the actual numerical algorithm based on a regularized gradient descent is presented. Finally, in section 4 we discuss various applications in two and three space dimensions and show corresponding numerical results.

1.1. Setup of the shape optimization problem. In shape optimization one seeks the shape $\mathcal{O}$ of a body which optimizes certain response properties. We shall focus here on optimality criteria which depend on the linear elastic response to applied forces. Therefore we start by describing the elastic problem. Given an admissible shape $\mathcal{O} \subset \mathbb{R}^d$ $(d = 2, 3)$ representing the elastic body, the displacement $u : \mathcal{O} \to \mathbb{R}^d$

is determined as the solution of the following system of linear partial differential equations:

$$\begin{aligned} -\operatorname{div}(Ae(u)) &= f(\omega) \quad \text{in } \mathcal{O}\,, \\ u &= 0 \quad \text{on } \Gamma_D\,, \\ (Ae(u))n &= g(\omega) \quad \text{on } \Gamma_N\,, \\ (Ae(u))n &= 0 \quad \text{on } \partial\mathcal{O} \setminus \Gamma_N \setminus \Gamma_D\,. \end{aligned} \tag{1.1}$$

Here, $e(u) = \frac{1}{2}(\nabla u + \nabla u^\top)$ is the linearized strain tensor and $A = (A_{ijkl})_{ijkl}$ the elasticity tensor. We shall for simplicity focus on isotropic materials, where $A_{ijkl} = 2\mu\delta_{ik}\delta_{jl} + \lambda\delta_{ij}\delta_{kl}$, where δ_{ij} denotes the Kronecker symbol and μ, λ the positive Lamé constants of the material. We consider only admissible shapes $\mathcal{O}$ which are subsets of a fixed, bounded working domain $D \subset \mathbb{R}^d$. On $\Gamma_D \subset \partial\mathcal{O}$ we assume homogeneous Dirichlet boundary conditions $u = 0$, and on $\Gamma_N \subset \partial\mathcal{O}$ we assume inhomogeneous Neumann boundary conditions, with $\Gamma_D \cap \Gamma_N = \emptyset$. Both parts of the boundary are kept fixed during the optimization. Precisely, we shall fix a certain open set $\mathcal{O}_* \subset D$, restrict the class of admissible shapes to $\mathcal{O}$ such that $\mathcal{O}_* \subset \mathcal{O} \subset D$, and assume that $\Gamma_D, \Gamma_N \subset \partial\mathcal{O}_* \cap \partial D$. Then necessarily $\Gamma_D, \Gamma_N \subset \partial\mathcal{O}$. Finally, $f(\omega) \in L^2(D;\mathbb{R}^d)$ and $g(\omega) \in L^2(\Gamma_N;\mathbb{R}^d)$ are random volume forces and surface loads, respectively, and ω is a realization on a probability space Ω. Standard results show that for any connected open set $\mathcal{O}$ with Lipschitz boundary and any fixed realization ω, the elasticity problem (1.1) has a unique weak solution $u = u(\mathcal{O},\omega) \in H^1(\mathcal{O};\mathbb{R}^d)$ [19, 39].

The unique solution to (1.1) can be equivalently characterized as the unique minimizer of a corresponding quadratic variational problem. In fact, $u(\mathcal{O},\omega)$ minimizes

$$E(\mathcal{O},u,\omega) := \frac{1}{2}A(\mathcal{O},u,u) - l(\mathcal{O},u,\omega) \quad \text{with} \tag{1.2}$$

$$A(\mathcal{O},\psi,\vartheta) := \int_{\mathcal{O}} A_{ijkl}e_{ij}(\psi)e_{kl}(\vartheta)\,\mathrm{d}x\,, \tag{1.3}$$

$$l(\mathcal{O},\vartheta,\omega) := \int_{\mathcal{O}} f_i(\omega)\vartheta_i\,\mathrm{d}x + \int_{\partial\mathcal{O}} g_i(\omega)\vartheta_i\,\mathrm{d}\mathcal{H}^{d-1} \tag{1.4}$$

among all u in $H^1_{\Gamma_D}(\mathcal{O};\mathbb{R}^d) := \{u \in H^1(\mathcal{O};\mathbb{R}^d) \,|\, u = 0 \text{ on } \Gamma_D \text{ in the sense of traces}\}$; see [19, 26, 39] for details. Here and below, we implicitly sum over repeated Cartesian indices.

As an objective functional $\mathbf{J}$ we consider

$$\mathbf{J}(\mathcal{O},\omega) = J(\mathcal{O},u(\mathcal{O},\omega)) := \int_{\mathcal{O}} j(u(\mathcal{O},\omega))\,dx + \gamma\int_{\partial\mathcal{O}} \mathrm{d}\mathcal{H}^{d-1}, \tag{1.5}$$

where γ is a nonnegative control parameter. The second term measuring surface area serves as a regularization. We assume that $j(\cdot)$ is linear or quadratic and does not depend explicitly on the realization ω.

A shape optimization problem under uncertainty is then formulated as

$$\text{minimize} \quad \{\mathbb{E}_\omega(\mathbf{J}(\mathcal{O},\omega)) \,:\, \mathcal{O} \in \mathcal{U}_{ad}\}, \tag{1.6}$$

where $\mathcal{U}_{ad}$ is the set of admissible shapes, e.g., $\mathcal{U}_{ad} := \{\mathcal{O} \subset D \,:\, \mathcal{O} \text{ open and of finite perimeter } \mathcal{O}_* \subset \mathcal{O}, \operatorname{Per}(\mathcal{O}) < \infty\}$, where $\operatorname{Per}(\mathcal{O}) = \mathrm{d}\mathcal{H}^{d-1}(\partial\mathcal{O})$ and for notational simplicity we write ∂) for the reduced boundary. Here and below, $\mathbb{E}_\omega(\dots)$ represents the expected value with respect to the probability distribution of the random variables $f(\omega)$, $g(\omega)$.

We emphasize that we solve the elasticity problem only in the physical domain $\mathcal{O}$. This differs from common practice in shape optimization, which is based on solving the elasticity problem on D with very small (but still positive) values of the elasticity constants λ and μ on $D \setminus \mathcal{O}$. For existence results in this context we refer to [9] and references therein. Our approach is closer to physical reality but brings some technical difficulties. The surface area term in the definition of the cost functional (1.5) ensures rectifiability of the domain boundary for configurations with finite energy but is not expected to guarantee existence of an optimal design. From a theoretical viewpoint, we are unaware of any result for the existence of solutions for the presently considered shape optimization problem. From a numerical viewpoint, this requires robust techniques to solve elasticity problems on badly shaped domains, which are discussed below. Furthermore, numerically different regularization strategies can be considered.

In the optimization problem (1.6) there is a natural information constraint stating that first, and independently of the realizations of $f(\omega), g(\omega)$, the shape $\mathcal{O}$ has to be selected. Then, after observation of $f(\omega), g(\omega)$, (1.1) determines the displacement field $u = u(\mathcal{O}, \omega)$, leading to the objective value $\mathbf{J}(\mathcal{O}, \omega)$. This manifests the interpretation of (1.6) as a two-stage random optimization problem: In the outer optimization, or first stage, the nonanticipative decision on $\mathcal{O}$ has to be taken. After observation of $f(\omega), g(\omega)$ the second-stage optimization problem is the mentioned variational problem, given $\mathcal{O}$ and ω. This second-stage optimization process is neither associated with further stochastic parameters nor with the optimization of additional material properties. In fact, it consists of the determination of the elastic displacement, which in turn is required for the computation of the elastic energy and the cost functional. Even though there is no additional decision making involved, the variational structure of the elasticity problem we are solving gives an obvious analogy to the second-stage problem in stochastic programming.

1.2. Deterministic level-set-based shape optimization. For the readers' convenience and to introduce notation we here briefly sketch the general procedure in deterministic shape optimization, where the volume and surface forces do not depend on a stochastic realization ω. Furthermore, we give an outline of our level set approach.

To get started, we consider variations $\mathcal{O}_v = (\mathrm{Id} + v)(\mathcal{O})$ of a smooth elastic domain $\mathcal{O}$ for a smooth vector field v defined on the working domain D. The shape derivative [22] of the objective functional $\mathbf{J}$ in the direction v takes the form

$$\begin{aligned}(1.7)\quad \mathbf{J}'(\mathcal{O})(v) &= J_{,\mathcal{O}}(\mathcal{O}, u(\mathcal{O}))(v) + J_{,u}(\mathcal{O}, u(\mathcal{O}))(u'(\mathcal{O})(v)) \\ &= \int_{\partial\mathcal{O}} (v \cdot n)\,(j(u(\mathcal{O})) + \gamma\, h)\ \mathrm{d}\mathcal{H}^{d-1} + \int_{\mathcal{O}} j_{,u}(u(\mathcal{O}))\,(u'(\mathcal{O})(v))\ \mathrm{d}x\,.\end{aligned}$$

Here, h denotes the mean curvature on $\partial\mathcal{O}$, defined as the sum of the principal curvatures, and $u'(\mathcal{O})(v)$ denotes the shape derivative of the elastic displacement defined by $u'(\mathcal{O})(v) = \lim_{t\to 0}[u((\mathrm{Id} + tv)\mathcal{O}) - u(\mathcal{O})]/t$.

In order to avoid the need of a separate evaluation of $u'(\mathcal{O})(v)$ for any infinitesimal domain displacement v, we seek a simpler expression for the $J_{,u}$ term. This is obtained by determining the variation of u with v implicitly, through its definition. Precisely, $u(\mathcal{O})$ was defined as the weak solution of (1.1), i.e.,

$$(1.8)\qquad A(\mathcal{O}, u(\mathcal{O}), \vartheta) = l(\mathcal{O}, \vartheta)$$

for all $\vartheta \in H^1_{\Gamma_D}(\mathcal{O}; \mathbb{R}^d)$. Differentiating this with respect to the variation v of the domain $\mathcal{O}$ (which in this entire discussion is assumed to be sufficiently smooth), we

get

$$A(\mathcal{O}, u'(\mathcal{O})(v), \vartheta) = l_{,\mathcal{O}}(\mathcal{O}, \vartheta)(v) - A_{,\mathcal{O}}(\mathcal{O}, u(\mathcal{O}), \vartheta)(v) \quad \text{with} \tag{1.9}$$

$$A_{,\mathcal{O}}(\mathcal{O}, \psi, \vartheta)(v) = \int_{\partial\mathcal{O}} (v \cdot n) A_{ijkl} e_{ij}(\psi) e_{kl}(\vartheta) \, \mathrm{d}\mathcal{H}^{d-1} , \tag{1.10}$$

$$l_{,\mathcal{O}}(\mathcal{O}, \vartheta)(v) = \int_{\partial\mathcal{O}} (v \cdot n) \, (f_i + g_i \, h + \partial_n g_i) \, \vartheta_i \, \mathrm{d}\mathcal{H}^{d-1}. \tag{1.11}$$

We observe that $J_{,u}(\mathcal{O}, u(\mathcal{O}))(\cdot)$ is a linear bounded functional on $L^2(D; \mathbb{R}^d)$. Therefore we can consider the dual problem and define $p(\mathcal{O}) \in H^1_{\Gamma_D}(D; \mathbb{R}^d)$ to be the solution of

$$A(\mathcal{O}, \vartheta, p(\mathcal{O})) = -J_{,u}(\mathcal{O}, u(\mathcal{O}))(\vartheta) \tag{1.12}$$

for all ϑ in $H^1_{\Gamma_D}(\mathcal{O}; \mathbb{R}^d)$. For the purpose of later reference let us also give a variational interpretation of this dual approach. Equation (1.12) corresponds to the fact that $p(\mathcal{O}) \in H^1_{\Gamma_D}(\mathcal{O}; \mathbb{R}^d)$ minimizes the quadratic functional

$$F(q) = \frac{1}{2} A(\mathcal{O}, q, q) + J_{,u}(\mathcal{O}, u(\mathcal{O}))(q) \tag{1.13}$$

among all $q \in H^1_{\Gamma_D}(\mathcal{O}; \mathbb{R}^d)$. In the strong formulation, we thus ask for a solution p of the system of partial differential equations $-\mathrm{div}\,(Ae(p(\mathcal{O}))) = -j_{,u}(u(\mathcal{O}))$, with $p(\mathcal{O}) = 0$ on Γ_D and $Ae(p(\mathcal{O})) \cdot n = 0$ on $\partial\mathcal{O} \setminus \Gamma_D$. Choosing $\vartheta = u'(\mathcal{O})(v)$ in (1.12) and recalling (1.19), one finally rewrites the shape derivative (1.7) of the objective functional as follows:

$$\begin{aligned} \mathbf{J}'(\mathcal{O})(v) &= J_{,\mathcal{O}}(\mathcal{O}, u(\mathcal{O}))(v) - A(\mathcal{O}, u'(\mathcal{O})(v), p(\mathcal{O})) \\ &= J_{,\mathcal{O}}(\mathcal{O}, u(\mathcal{O}))(v) - l_{,\mathcal{O}}(\mathcal{O}, p(\mathcal{O}))(v) + A_{,\mathcal{O}}(\mathcal{O}, u(\mathcal{O}), p(\mathcal{O}))(v) \\ &= \int_{\partial\mathcal{O}} (v \cdot n) \big[j(u(\mathcal{O})) + \gamma \, h - (f_i + g_i \, h + \partial_n g_i) \, p_i(\mathcal{O}) \\ &\qquad + A_{ijkl} e_{ij}(u(\mathcal{O})) e_{kl}(p(\mathcal{O})) \big] \, \mathrm{d}\mathcal{H}^{d-1} . \end{aligned} \tag{1.14}$$

In order to permit the topology of the domain $\mathcal{O}$ to change, we consider an implicit description of shapes in terms of a level set function $\phi : D \to \mathbb{R}$. In particular, the elastic body is represented by $\mathcal{O} = \{\phi < 0\} := \{x \in D \,|\, \phi(x) < 0\}$, and its boundary $\partial\mathcal{O}$ corresponds to the zero level set of ϕ, i.e., $D \cap \partial\mathcal{O} = \{\phi = 0\}$. Shape optimization and shape analysis for elastic solids via level set methods has been investigated by various authors [9, 23, 36, 52]. In particular, Allaire and coworkers [4–6, 9] have extensively studied a level set modeling of shapes in 2D and 3D structural optimization and compared and combined this approach with homogenization methods. In [8] they recently investigated topological optimization in the context of minimizing the expected elastic stress.

Interface propagation based on level sets was first introduced by Osher and Sethian [43] and since then attracted very much attention due to their enormous flexibility. For a general overview we refer to [42,51]. If a domain boundary $\partial\mathcal{O}$ propagates with speed v, the evolution of the corresponding level set function ϕ is given by the level set equation $\partial_t \phi + |\nabla\phi| \, v \cdot n = 0$, where $n = \frac{\nabla\phi}{|\nabla\phi|}$ is the field of outer normals on the level sets. In fact, the level set equation identifies variations $s = \partial\phi$ of the level set function with variations $v \cdot n$ of the level sets in the direction of the normal

n. Even though hypersurfaces are described in the level set context by functions on the whole domain, suitable implementations lead to efficient numerical algorithms as well [2, 34, 57]. Fairly general shapes can be effectively described and modeled with level sets [38]. Shape sensitive analysis as introduced by Sokolowski and Zolésio [54] can be phrased elegantly in terms of level sets. Let us rewrite the objective functional $\mathbf{J}(\mathcal{O})$ in terms of a level set function ϕ and define

$$\mathcal{J}(\phi) := \mathbf{J}(\{\phi < 0\}). \tag{1.15}$$

Due to the above identification we obtain for the shape derivative of $\mathcal{J}(\phi)$ with respect to a variation s of ϕ (again, working for the moment on smooth domains and away from degeneracies and topological changes)

$$\mathcal{J}'(\phi)(s) = \mathbf{J}'(\{\phi < 0\}) \left(-s\,|\nabla\phi|^{-1} n\right) . \tag{1.16}$$

For the relaxation of the shape functional we now consider a gradient descent

$$\partial_t \phi(t) = -\mathrm{grad}_{\mathcal{G}} \mathcal{J}(\phi)$$

with respect to a metric $\mathcal{G}$ on the space of variations of the level set function ϕ (cf. [45]). This metric ensures smoothness of the descent path and is expected to approximate a regular minimizer from the set of all minimizers. For an overview on optimal design based on level sets and suitable energy descent methods we refer to a recent survey by Burger and Osher [17]. From (1.14) we learn that the support of $\mathbf{J}'(\mathcal{O})(\cdot)$ is contained in $\partial\mathcal{O} \setminus \Gamma_D$. Thus, we take into account a regularized gradient descent based on the metric

$$\mathcal{G}(\zeta, \xi) = \int_D \zeta\xi + \frac{\rho^2}{2} \nabla\zeta \cdot \nabla\xi \,\mathrm{d}x\,, \tag{1.17}$$

which is related to a Gaussian filter with width ρ. For the time discretization, we consider Armijo rule as a step size control, and starting with an initial level set function ϕ^0 we iteratively compute a sequence of level set functions $(\phi^k)_{k=1,\dots}$ given by

$$\mathcal{G}(\phi^{k+1} - \phi^k, \xi) = -\tau \mathcal{J}'(\phi^k)(\xi) \tag{1.18}$$

for all test functions ξ and a sequence of time steps $(\tau^k)_{k=1,\dots}$. In each time step a linear elliptic problem of the type $(\mathrm{Id} - \frac{\rho^2}{2}\Delta)\phi = r$ has to be solved. Alternatively, one might consider a relaxation of shapes described via an evolution of signed distance functions [21,28]. For the spatial discretization we consider piecewise affine continuous finite element functions on the working domain D. Shape relaxations tend to create fine scale structures and complicated domains $\mathcal{O}$. To evaluate the objective functional itself and the shape derivative, the elastic displacement u on $\mathcal{O}$ has to be computed solving the Euler Lagrange equations (1.1) of the inner, elastic minimization subproblem. Here, we apply multilevel composite finite elements introduced by Hackbusch and Sauter [31, 50]. They incorporate the characteristic behavior of the solution on fine scales into the coarse scale shape functions without, necessarily, adding degrees of freedom.

1.3. Two-stage stochastic programming revisited. Before we apply two-stage stochastic programming to our shape optimization problem, let us recall the basic concepts from finite-dimensional stochastic optimization. Consider the random linear program

$$\min \left\{ c^\top x + q^\top y \;:\; Tx + Wy \,=\, z(\omega),\;\; x \in X,\; y \in Y \right\} \tag{1.19}$$

for finite-dimensional polyhedra X and Y in Euclidean space together with the information constraint

$$\text{decide } x \;\mapsto\; \text{observe } \omega \;\mapsto\; \text{decide } y = y(x,\omega).$$

We assume that the minimum exists; possibly making the spaces larger, we can also without loss of generality replace the condition $y \in Y$ by $y \geq 0$ (that is, $y_i \geq 0$ for all i). We also remark that given x and $z(\omega)$ there are multiple solutions y from which we have to select one.

Let us emphasize the two-stage characteristic of this optimization problem. Indeed, rewriting (1.19) yields

$$\begin{aligned} &\min_x \left\{ c^\top x + \min_y \{ q^\top y \,:\, Wy \,=\, z(\omega) - Tx,\; y \in Y \} \;:\; x \in X \right\} \\ &= \min \left\{ c^\top x + \Phi(z(\omega) - Tx) \;:\; x \in X \right\}, \end{aligned} \tag{1.20}$$

where $\Phi(v) := \min\{q^\top y \,:\, Wy = v,\; y \in Y\}$ is the value function of a linear program with parameters on the right-hand side. The cost functional we aim to minimize is $\mathsf{j}(x,\omega) := c^\top x + \Phi(z(\omega) - Tx)$. The representation (1.20) gives rise to understanding the search for a "best" nonanticipative decision x in the initial random optimization problem as the search for a "minimal" member in the family of random variables $\{\mathsf{j}(x,\omega) \;:\; x \in X\}$, where x is seen as an "index" varying in the set X. In a risk-neutral setting, these random variables are ranked by their expectations, leading to the (nonlinear) optimization problem

$$\min\{Q_{\mathbb{E}}(x) \,:=\, \mathbb{E}_\omega\left(\mathsf{j}(x,\omega)\right) \;:\; x \in X\}. \tag{1.21}$$

The straightforward but crucial idea is to detect structural properties and algorithmic possibilities in (1.21) by resorting to the dual of the linear program with value function $\Phi(\cdot)$. Indeed, one observes

$$\begin{aligned} \Phi(v) &= \min\left\{ q^\top y \,:\, Wy = v,\; y \geq 0 \right\} \\ &= \max\left\{ v^\top y \,:\, W^\top y \leq q \right\} = \max_{l=1,\dots,L} d_l^\top v\,, \end{aligned} \tag{1.22}$$

where $\{d_l\}_{l=1,\dots,L}$ denotes the set of vertices of the dual polyhedron $\{y \,:\, W^\top y \leq q\}$, which is assumed compact, and $v = z(\omega) - Tx$. Recalling the cost functional $\mathsf{j}(x,\omega)$, we can rewrite (1.21) and obtain

$$\min\left\{ c^\top x + \sum_{\sigma=1}^{S} \pi_\sigma \max_{l=1,\dots,L} d_l^\top (z_\sigma - Tx) \;\;:\;\; x \in X \right\} \tag{1.23}$$

in the case of a discrete probability distribution with realizations z_σ and probabilities π_σ for $\sigma = 1,\dots,S$. Here, S is the total number of scenarios. Thus, minimizing $Q_{\mathbb{E}}$ amounts to minimizing a piecewise linear convex function over a polyhedron. Let us emphasize that in our concrete setup, the functional to be minimized in (1.23) depends linearly on the random variable z, which can be exploited further in the actual numerical minimization.

Algorithmically, two aspects are important: By its very definition, computing $Q_{\mathbb{E}}(x)$ in (1.21) would amount to solving $\min\{q^\top y \,:\, Wy = z_\sigma - Tx,\; y \geq 0\}$ for all scenarios z_σ with $\sigma = 1,\dots,S$ and this again at any new iteration point x. In

(1.23) this is prevented by using dual information. Here, the situation is particularly comfortable since cutting planes generated in adaptations of bundle methods, see, e.g., [48,53], capture (at least approximately) information on the objective also locally around iteration points. The second aspect is that (sub)gradient information on $Q_{\mathbb{E}}$ is made available by the help of the dual, cf. (1.23).

The facts reviewed above form our guideline for treating shape optimization under uncertainty: Departing from the outlined two-stage model with shape decisions in the first stage and displacements in the second, we will formulate an (infinite-dimensional) counterpart to the expectation problem (1.21). The variational formulation of the elasticity system will provide an inner optimization problem in the spirit of (1.20). As in (1.22) a duality argument will provide information for the shape derivative. In what follows, the domain $\mathcal{O}$ replaces the variable x, the elastic deformation $u(\mathcal{O},\omega)$ the optimal solution y being a minimizer of the above $\Phi(v)$, where v depends on x and $z(\omega)$. Finally, as a counterpart to the cost functional $\mathsf{j}(x,\omega)$ we consider the objective functional $\mathbf{J}(\mathcal{O},\omega)$. Moreover, as above, in each iteration of a descent method linearity of the elasticity PDE will avoid the solution of as many related PDEs as there are scenarios.

2. Two-stage stochastic programming formulation of shape optimization. We now present our stochastic shape optimization scheme, which incorporates the techniques from deterministic shape optimization discussed in section 1.2 and the two-stage stochastic programming reviewed in section 1.3. In our setting, the second stage optimization problem is the variational problem of linearized elasticity, where for a fixed elastic domain $\mathcal{O}$ and random state ω one seeks a displacement u which minimizes the energy $E(\mathcal{O},u,\omega)$ defined in (1.2). In turn, the objective functional can be computed from the domain $\mathcal{O}$ and the displacement u and hence can be seen as a function of $\mathcal{O}$ and the random state ω. We observe the following information constraints:

$$\text{decide } \mathcal{O} \;\mapsto\; \text{observe } \omega \;\mapsto\; \text{compute } u = u(\mathcal{O},\omega)\,.$$

In other words, one first selects a domain $\mathcal{O}$ (like in section 1.3 one decided for some x), then random volume and boundary forces $f(\omega)$ and $g(\omega)$ are applied (the counterpart of the right-hand side $z(\omega)$ in (1.19)), and only at this point the elastic displacement u (the counterpart of the degree of freedom y in (1.19)) and hence the objective functional can be computed. Thus, in analogy to (1.20) we can reformulate the random shape optimization problem in a two-stage optimization manner as follows:

$$\min\left\{\mathbf{J}(\mathcal{O},\omega)\;:\; u(\mathcal{O},\omega) = \operatorname{argmin}_{u\in H^1_{\Gamma_D}(\mathcal{O};\mathbb{R}^d)} E(\mathcal{O},u,\omega)\right\}.$$

As mentioned above, $\mathcal{O}$ has the role of the first-stage and $u(\mathcal{O},\omega)$ of the second-stage decisions. Finally, the stochastic program

$$\min\left\{Q_{\mathbb{E}}(\mathcal{O}) := \mathbb{E}_\omega\left(\mathbf{J}(\mathcal{O},\omega)\right)\;:\; \mathcal{O}\in\mathcal{U}_{ad}\right\} \tag{2.1}$$

arises as the "natural" counterpart to (1.21). Replacing the variational problem in (2.1) by its Euler equation enables us to introduce the (dual or) adjoint system needed to effectively compute gradients of the stochastic objective functional.

2.1. Stochastic primal and dual problem. We start from the analysis of the second-stage problem. As illustrated in section 1.3, in order to determine the shape derivative of the objective function it is convenient to solve both the primal and the

dual elastic problem as a counterpart of (1.22) in two-stage stochastic programming. Precisely, given $\mathcal{O}$ and ω we seek a primal solution $u(\mathcal{O},\omega) \in H^1_{\Gamma_D}(D;\mathbb{R}^d)$ and a dual solution $p(\mathcal{O},\omega) \in H^1_{\Gamma_D}(D;\mathbb{R}^d)$ such that

$$A(\mathcal{O}, u(\mathcal{O},\omega), \vartheta) = l(\mathcal{O},\vartheta,\omega)\,, \tag{2.2}$$

$$A(\mathcal{O}, \vartheta, p(\mathcal{O},\omega)) = -J_{,u}(\mathcal{O}, u(\mathcal{O},\omega))(\vartheta) \tag{2.3}$$

for all $\vartheta \in H^1_{\Gamma_D}(D;\mathbb{R}^d)$. The function $u(\mathcal{O},\omega)$ entering the dual problem is the solution to the primal problem. Let us emphasize that as in (1.22) both the primal and the dual state solve variational problems, in fact (1.2) and (1.13), respectively.

A key simplification in the solution of these equations arises from the general fact that the solution of a linear problem depends linearly on the data. We phrase this fact first in general terms and then discuss the implications in our setting. Let $A_{\mathcal{O}} : H^1_{\Gamma_D} \to H^{-1}_{\Gamma_D}(D;\mathbb{R}^d)$ be the elliptic operator induced by the quadratic form $A(\mathcal{O},\cdot,\cdot)$, in fact $A_{\mathcal{O}}(u)(\vartheta) = A(\mathcal{O},u,\vartheta)$. By the positivity of the elastic coefficients, for any Lipschitz, connected domain $\mathcal{O}$, and under the assumption that $\Gamma_D \subset \partial\mathcal{O}$ has positive $(d-1)$-dimensional measure, the operator $A_{\mathcal{O}}$ is bounded and coercive on the Hilbert space $H^1_{\Gamma_D}(D;\mathbb{R}^d)$ and therefore invertible. This implies that for any $l \in H^{-1}_{\Gamma_D}(D;\mathbb{R}^d)$ one can find a unique solution u to $A_{\mathcal{O}}(u,\vartheta) = l(\vartheta)$, namely, $u = A^{-1}_{\mathcal{O}} l$. Therefore both (2.2) and (2.3) have a unique solution, which depends linearly on the right-hand side.

We now consider the specific case of interest here, namely, the dependence of u and p on ω. The crucial point is that the left-hand side of both equations, i.e., the quadratic form $A(\mathcal{O},\cdot,\cdot)$, does not depend on ω. The right-hand side depends on ω only through f, g, and u, and this dependence is linear. Here, it is important that the integrand j entering the objective function is linear or quadratic. We shall now exploit this fact in order to obtain an efficient algorithm, which does not require us to solve (2.2) and (2.3) for every ω but only for a representative subset (a "basis").

We start from the primal problem (2.2). Since the right-hand side is linear in the forces f and g, and A does not depend on ω, the solution u depends linearly on the forces f and g. In order to make this more explicit, assume that f and g are random combinations of finitely many forces $f^1,\dots,f^K \in L^2(D;\mathbb{R}^d)$ and $g^1,\dots,g^M \in H^1(D;\mathbb{R}^d)$, respectively, i.e.,

$$f(\omega) = \sum_{k=1}^{K} \alpha_k(\omega) f^k, \qquad g(\omega) = \sum_{m=1}^{M} \beta_m(\omega) g^m.$$

Here, the $\alpha_k(\omega)$ and $\beta_m(\omega)$ are stochastic coefficients. For later convenience we assume that $\sum_{k=1}^{K} \alpha_k(\omega) = \sum_{m=1}^{M} \beta_k(\omega) = 1$. (This can always be achieved by a rescaling of the coefficients of f^k's and g^m's such that their corresponding sums are both smaller than 1. Adding two virtual loads equal to zero, we easily can ensure the equality sign.) We assume that ω follows a discrete distribution with scenarios ω_σ and probabilities π_σ, with $\sigma = 1,\dots,S$ ($\sum_{\sigma=1}^{S} \pi_\sigma = 1$); continuous distributions can be recovered in the limit $S \to \infty$. For any pair $(k,m) \in \{1,\dots,K\} \times \{1,\dots,M\}$ let $u^{km}(\mathcal{O})$ be the solution to the elasticity system $(1.1)_{km}$, which is (1.1) with right-hand sides f^k, g^m. Then, for any $\sigma = 1,\dots,S$,

$$\bar{u}(\mathcal{O},\omega_\sigma) \;:=\; \sum_{k=1}^{K}\sum_{m=1}^{M} \alpha_k(\omega_\sigma)\beta_m(\omega_\sigma) u^{km}(\mathcal{O}) \tag{2.4}$$

solves (1.1) for $\omega = \omega_\sigma$. In the numerical implementation, we confine to $K + M$ displacements, where each of them corresponds either to a single volume force or a single boundary force and vanishing components with respect to all other forces. To simplify the presentation, we consider here an overdetermined system of $K\,M$ spanning displacements. This is a substantial algorithmic shortcut, in the case that the discretization parameter of the probability measure S is larger than the sum of the $K + M$ effective base forces.

An analogous argument applies to the dual problem (2.3). We first determine, for each pair $(k, m) \in \{1, \dots, K\} \times \{1, \dots, M\}$, the solution $p^{km}(\mathcal{O})$ of the basis problem

$$A\left(\mathcal{O}, \vartheta, p^{km}(\mathcal{O})\right) = -J_{,u}\left(\mathcal{O}, u^{km}(\mathcal{O})\right)(\vartheta) \text{ for all } \vartheta \in H^1_{\Gamma_D}\left(D; \mathbb{R}^d\right). \tag{2.5}$$

Since j depends linearly or quadratically on u, the dependence of $j_{,u}$ on u is linear (possibly trivial). Therefore (2.4) and the above-introduced normalization implies

$$J_{,u}(\mathcal{O}, \bar{u}(\mathcal{O}, \omega_\sigma))(\vartheta) = \sum_{k=1}^{K} \sum_{m=1}^{M} \alpha_k(\omega_\sigma)\beta_m(\omega_\sigma) J_{,u}(\mathcal{O}, u^{km}(\mathcal{O}))(\vartheta),$$

and linearity of the inverse operator $A^{-1}_{\mathcal{O}}$ gives

$$\bar{p}(\mathcal{O}, \omega_\sigma) = \sum_{k=1}^{K} \sum_{m=1}^{M} \alpha_k(\omega_\sigma)\beta_m(\omega_\sigma) p^{km}(\mathcal{O}). \tag{2.6}$$

Obviously, $\bar{p}(\mathcal{O}, \omega_\sigma)$ is the weak solution $\bar{p}$ of $-\mathrm{div}\,(Ae(\bar{p})) = -j'(\bar{u}(\mathcal{O}, \omega_\sigma))$ on the domain $\mathcal{O}$ with $\bar{p} = 0$ on Γ_D and $Ae(\bar{p}) \cdot n = 0$ on $\partial\mathcal{O} \setminus \Gamma_D$.

2.2. Shape gradient in the stochastic optimization problem. Now, with the primal solution $\bar{u}(\mathcal{O}, \omega_\sigma)$ for a particular realization ω_σ at hand, the stochastic program (2.1) can be rewritten as follows:

$$\begin{aligned} \min \Bigg\{ & \gamma \int_{\partial\mathcal{O}} \mathrm{d}\mathcal{H}^{d-1} + \sum_{\sigma=1}^{S} \pi_\sigma \int_{\mathcal{O}} j(\bar{u}(\mathcal{O}, \omega_\sigma))\, dx \ : \\ & \bar{u}(\mathcal{O}, \omega_\sigma) := \sum_{k=1}^{K} \sum_{m=1}^{M} \alpha_k(\omega_\sigma)\beta_m(\omega_\sigma) u^{km}(\mathcal{O}),\ \sigma = 1, \dots, S \Bigg\}. \end{aligned} \tag{2.7}$$

Using the primal solution for the elastic deformation $\bar{u}(\mathcal{O}, \omega_\sigma)$ and the dual solution $\bar{p}(\mathcal{O}, \omega_\sigma)$ for any realization ω_σ, we deduce the stochastic shape derivative (1.7) of the objective functional $\mathbf{J}(\mathcal{O}, \omega_\sigma)$ and achieve, from (1.14),

$$\begin{aligned} \mathbf{J}'(\mathcal{O}, \omega_\sigma)(v) &= J_{,\mathcal{O}}(\mathcal{O}, \bar{u}(\mathcal{O}, \omega_\sigma))(v) - l_{,\mathcal{O}}(\mathcal{O}, \bar{p}(\mathcal{O}, \omega_\sigma))(v) \\ &\quad + A_{,\mathcal{O}}(\mathcal{O}, \bar{u}(\mathcal{O}, \omega_\sigma), \bar{p}(\mathcal{O}, \omega_\sigma))(v) \\ &= \int_{\partial\mathcal{O}} (v \cdot n)\big(j(\bar{u}(\mathcal{O}, \omega_\sigma)) + \gamma\, h - (f_i(\omega_\sigma) + g_i(\omega_\sigma)\, h + \partial_n g_i)\, \bar{p}_i(\mathcal{O}, \omega_\sigma) \\ &\qquad + A_{ijkl} e_{ij}(\bar{u}(\mathcal{O}, \omega_\sigma)) e_{kl}(\bar{p}(\mathcal{O}, \omega_\sigma))\big)\ \mathrm{d}\mathcal{H}^{d-1}. \end{aligned} \tag{2.8}$$

Finally, the shape derivative of our actual stochastic cost functional, namely, of the expectation of the cost $Q_{\mathbb{E}}(\mathcal{O})$ in the case of S scenarios $(\omega_\sigma)_{\sigma=1,\dots,S}$, is given by

$$Q'_{\mathbb{E}}(\mathcal{O})(v) = \mathbb{E}_\omega\left(\mathbf{J}'(\mathcal{O}, \omega)(v)\right) = \sum_{\sigma=1}^{S} \pi_\sigma J'(\mathcal{O}; \omega_\sigma)(v). \tag{2.9}$$

In the algorithm this shape derivative can be used as a descent direction. Thereby, first the $K\,M$ primal and dual base states are computed. These allow for the efficient evaluation of the effective deformations $\bar{u}(\mathcal{O}, \omega_\sigma)$ and the effective dual states $\bar{p}(\mathcal{O}, \omega_\sigma)$ for a set of S scenarios ω_σ with S usually much larger than $K\,M$.

3. Multiscale finite element implementation. In this section we detail the concrete numerical algorithm and consider a finite element approach for the representation of the level set function ϕ on the working domain D, which implicitly describes the discrete elastic domain $\mathcal{O}$ as the sublevel set of the discrete level set function. The elastic state equations for u^{km} and the corresponding set of dual problems for p^{km} are discretized as well with finite elements. Here, we pick up the composite finite element approach originally proposed by [31] and investigated in the level set context for complicated 3D geometries in [35]. Finally, we will discuss the time step control used in our descent scheme.

3.1. Finite element spaces. Without any restriction we suppose our working domain D to be a hexahedron ($d = 3$) or a rectangle ($d = 2$), respectively. In a first step, a hierarchical grid is generated based on successive subdivision of hexahedrons (resp., rectangles) into 8 (resp., 4) equally sized child hexahedrons (resp., rectangles). Next, each cell of the resulting fine grid is split into 6 tetrahedra (2 triangles) such that a regular simplicial grid $\mathcal{T}_h$ of the domain D is obtained. We denote the simplicial elements of this grid by $T \in \mathcal{T}_h$ and the set of nodes by $\mathcal{N}_h = \{X_i\}_{i \in I_h}$ with a corresponding index set I_h. Let us emphasize that we do not represent this simplicial grid explicitly. Instead access to element data is implicitly encoded in look up tables. Here, h indicates the grid size. Let $\mathcal{V}_h$ be the space of continuous, piecewise affine functions on $\mathcal{T}_h$ with the canonical basis $\{\Phi_i\}_{i \in I_h}$, given by $\Theta_i(X_j) = \delta_{ij}$. In what follows, discrete variables will always be capitalized, whereas continuous ones will be lowercase. Now, we consider a discrete level set function $\Phi(x) = \sum_{i \in I_h} \Phi_i \Theta_i(x)$. As a consequence, the discrete domain $\mathcal{O}_h = \{x \in D \,|\, \Phi(x) < 0\}$ is polygonal. This algorithmic advantage justifies the use of a tetrahedral grid. A solution of the state equation (2.2) and the dual problem (2.3) is defined on the elastic domain only. Here, we explicitly work with a void phase $D \setminus \mathcal{O}$, and, at variance with [6, 9], we do not consider a softer elastic material outside of actual elastic body $\mathcal{O}$ to be optimized. Thus, we have to define suitable finite element spaces on the discrete elastic domain $\mathcal{O}_h$ implicitly described by a level set function $\Phi \in \mathcal{V}_h$. A straightforward mesh generation based on a marching cube-type algorithm [37, 56] leads to badly shaped tetrahedra with a significant impact on the condition number of the linear systems to be solved. Explicit grid generation would require a regular remeshing of the boundary $\partial\mathcal{O}_h$ followed by the actual meshing algorithm in $\mathcal{O}_h \subset \mathbb{R}^d$ [24, 44]. Both steps are fairly complicated in the case of general elastic domains and result in nonhierarchical, unstructured meshes which do not allow for a multilevel algorithm for the discrete PDE problems. To avoid these drawbacks we construct a suitable composite finite element space. In contrast to explicit meshing approaches, the geometry is encoded in the design of the basis functions, which still correspond to grid nodes of the regular underlying grid. In fact, given a basis function $\Theta_i \in \mathcal{V}_h$ whose support intersects the discrete elastic domain $\mathcal{O}_h$, we define the corresponding composite finite element basis function $\Theta_i^{\text{cfe}}(x) = \chi_{\mathcal{O}_h}(x)\Theta_i(x)$, selecting the part of the old basis function contained in the elastic domain [31]. Here, $\chi_{\mathcal{O}_h}$ denotes the characteristic function of the discrete domain $\mathcal{O}_h$. Let us remark that there are also degrees of freedom at nodes outside the actual domain as long as the support of the corresponding basis function interests $\mathcal{O}_h$. Collecting all of these basis functions we obtain the composite

finite element space

$$\mathcal{V}_h^{\mathrm{cfe}} := \{\Theta_i^{\mathrm{cfe}}(x) = \chi_{\mathcal{O}_h}(x)\Theta_i(x) \,|\, \operatorname{supp}\Theta_i \cap \mathcal{O}_h \neq \emptyset\},$$

and the resulting nodal index set I_h^{cfe} is a subset of the index set I_h. Hence, far from the domain boundary the basis functions coincide with the standard basis functions, whereas in the vicinity of the boundary, the standard basis is modified to resolve the domain geometry. Finally, let us incorporate boundary data and define $\mathcal{V}_{h,\Gamma_D}^{\mathrm{cfe}} = (\mathcal{V}_h^{\mathrm{cfe}})^3 \cap H^1_{\Gamma_D}(D;\mathbb{R}^d)$ as the space of discrete vector valued functions which vanish on the Dirichlet boundary Γ_D. For the sake of simplicity, we assume here Γ_D to be resolved on the underlying regular grid. Thus, no special treatment of the Dirichlet boundary condition in the construction of the composite finite elements [32] is required. Indeed, to conserve the Dirichlet boundary condition we furthermore freeze the level set function ϕ in a small neighborhood of the Dirichlet boundary Γ_D and the Neumann boundary Γ_N on which the surface load is applied. Hence, in this region the body still behaves elastic but does not undergo any optimization. As basis functions for the vector valued problem we consider $\Theta_i^{\mathrm{cfe}} e_j$ with $i \in I_h^{\mathrm{cfe}}$ and $1 \leq j \leq d$.

3.2. Discrete primal and dual solutions. Given the composite finite element space $\mathcal{V}_{h,\Gamma_D}^{\mathrm{cfe}}$ we can solve the primal and the dual problem numerically. Explicitly, the discrete primal solutions are defined as the finite element functions $U^{km} \in \mathcal{V}_{h,\Gamma_D}^{\mathrm{cfe}}$ solving

$$A\left(\mathcal{O}_h, U^{km}, \Theta\right) = l^{km}(\mathcal{O}_h, \Theta) \tag{3.1}$$

for all $\Theta \in \mathcal{V}_{h,\Gamma_D}^{\mathrm{cfe}}$, where $l^{km}(\mathcal{O}_h,\Theta) := \int_{\mathcal{O}_h} f_i^k \Theta_i \,\mathrm{d}x + \int_{\partial\mathcal{O}_h} g_i^m \Theta_i \,\mathrm{d}\mathcal{H}^{d-1}$ for $1 \leq k \leq K$ and $1 \leq m \leq M$. The corresponding set of dual solutions are those functions $P^{km} \in \mathcal{V}_{h,\Gamma_D}^{\mathrm{cfe}}$ for which

$$A\left(\mathcal{O}_h, \Theta, P^{km}\right) = -J_{,u}\left(\mathcal{O}_h, U^{km}\right)(\Theta) \tag{3.2}$$

for all $\Theta \in \mathcal{V}_{h,\Gamma_D}^{\mathrm{cfe}}$. For the variation of the cost functional J with respect to the discrete elastic displacement U we obtain $J_{,u}(\mathcal{O}_h, U^{km})(\Theta) = \int_{\mathcal{O}_h} j_{,u}(U^{km})(\Theta)\,\mathrm{d}x$. Due to the assumption that $j(\cdot)$ is a linear or quadratic polynomial, the resulting integrant is at most quadratic and can be integrated exactly using a Gauss quadrature rule. In the case of the compliance cost functional $J(\mathcal{O}, u(\mathcal{O},\omega)) = l^{km}(\mathcal{O}, u^{km}, \omega) + \gamma \int_{\partial\mathcal{O}} \mathrm{d}\mathcal{H}^{d-1}$, we derive as usual from (3.1) the representation $J_{,u}(\mathcal{O}_h, U^{km})(\Theta) = A(\mathcal{O}_h, U^{km}, \Theta) = \int_{\mathcal{O}_h} A_{ijkl} e_{ij}(U^{km}) e_{kl}(\Theta)\,\mathrm{d}x$. The numerical solution of (3.1) and (3.2) both require numerical quadrature for the assembly of the stiffness matrix $(A(\mathcal{O}_h, \Theta_i e_j, \Theta_r e_s))_{i,r\in I_h^{\mathrm{cfe}}, 1\leq j,s\leq d}$ and the right-hand side vectors $(l^{km}(\mathcal{O}_h, \Theta_r e_s))_{r\in I_h^{\mathrm{cfe}}, 1\leq s\leq d}$ and $(-J_{,u}(\mathcal{O}_h, U^{km})(\Theta_r e_s))_{r\in I_h^{\mathrm{cfe}}, 1\leq s\leq d}$, respectively. For this purpose, on simplices of the original mesh which are intersected by the domain boundary $\partial\mathcal{O}_h$ a local, virtual grid is generated. Based on a look up table the cells generated by the marching cube-type method in the construction of the composite finite elements are subdivided into simplices. On these simplices and on the simplices within $\mathcal{O}_h$ not intersected by the domain boundary a one point, center of mass quadrature rule is applied. The evaluation of the boundary integral $\int_{\partial\mathcal{O}_h} g_s^m \Theta_r \,\mathrm{d}\mathcal{H}^{d-1}$ is treated analogously.

As long as the discrete domain $\mathcal{O}_h$ is connected in the following discrete sense, namely, for every node X_i with $i \in I_h^{\mathrm{cfe}}$ there is a chain of nodes $(X_j)_{j=0,\ldots,n}$ with $j \in I_h^{\mathrm{cfe}}$ such that $[X_j, X_{j+1}]$ is an edge of $\mathcal{T}_h$, $X_0 = X_i$, and X_n is a node on Γ_D,

we easily verify that there exist unique solutions U^{km} and P^{km} of (3.1) and (3.2), respectively. The resulting symmetric linear systems of equations are solved with a conjugate gradient method for $d = 2$ and with a multigrid method for $d = 3$. In general a still high condition number for the corresponding linear system of equations on the finest grid level will reflect the badly shaped support of single composite basis functions. Here, in particular the multigrid method leads to convergence rates which are independent of the grid size h and—for a wide range of problems—the geometric complexity of the domain. For the multigrid solver, we first recursively construct coarse grid matrices and right-hand sides. Here, the underlying hierarchical grid induces a canonical projection operator for any grid level to the next finer one generated by the cell subdivision. Let us emphasize that this applies not only for the hierarchical hexahedral grid but analogously for the associated simplicial mesh as well. Based on the projection operator a standard Galerkin projection [30] is applied both for the matrices and the right-hand sides. We then use a multigrid method with V cycles and symmetric Block–Gauß–Seidel iterations as a smoother. Thereby, we gather the 3 spatial components of the solution at a grid node and apply the Gauß–Seidel iterations on the resulting 3×3 blocks. In the applications considered here, 3 pre- and post-smoothing steps in the V cycle turned out to be a reasonable choice. For details on the composite finite element approach and the multigrid method we refer to [35].

3.3. Discrete gradient descent algorithm. The numerical relaxation of the shape functional is based on the time discretized, regularized gradient descent scheme given in (1.18) and applied to the spatially discrete stochastic shape functional

$$Q_{\mathbb{E},h}(\mathcal{O}_h) := \mathbb{E}_\omega\left(\mathcal{J}(\Phi,\omega)\right) = \sum_{\sigma=1}^{S} \pi_\sigma \mathcal{J}(\Phi,\omega_\sigma), \tag{3.3}$$

where the shape functional $\mathcal{J}$ for a discrete level set function Φ is defined in a straightforward way by $\mathcal{J}(\Phi,\omega_\sigma) := \mathbf{J}(\{\Phi < 0\},\omega_\sigma)$ for any realization ω_σ. Here, for the ease of presentation we notationally do not distinguish continuous and discrete shape functionals; in fact, in what follows discrete shape functionals always involve the corresponding discrete solution of the state equation. For an initial level set function $\Phi^0 \in \mathcal{V}_h$ we iteratively compute a sequence of level set functions $(\Phi^k)_{k=1,\dots}$ given by

$$\mathcal{G}\left(\Phi^{k+1} - \Phi^k, \Xi\right) = -\tau \mathbb{E}_\omega\left(\mathcal{J}'\left(\Phi^k,\omega\right)(\Xi)\right) \tag{3.4}$$

for all $\Xi \in \mathcal{V}_h$. Hence, in every time step the vector $(\mathbb{E}_\omega(\mathcal{J}'(\Phi^k,\omega)(\Psi_i)))_{i\in I_h}$ of variations of the expectation of the objective functional $\mathcal{J}$ in all basis directions Ψ_i for $i \in I_h$ has to be evaluated. Furthermore, one has to solve the linear system of equations resulting from a standard finite element discretization of $\mathcal{G}$. As already discussed the time step τ is chosen according to a simple variant of the Armijo step size control. Indeed, given a constant $\beta \in (0,1)$ we accept a time step τ if the condition

$$\mathbb{E}_\omega\left(\mathcal{J}\left(\Phi^{k+1},\omega\right)\right) - \mathbb{E}_\omega\left(\mathcal{J}\left(\Phi^k,\omega\right)\right) \le -\beta\mathcal{G}\left(\Phi^{k+1} - \Phi^k, \Phi^{k+1} - \Phi^k\right)$$

is satisfied; otherwise the timestep is reduced.

Let us now detail the evaluation of $\mathcal{J}'(\Phi)(\Xi)$ in the spatially discrete setting. For any scenario of the stochastic loading ω_σ with $\sigma = 1,\dots,S$ we obtain a discrete effective displacement $\bar{U}(\mathcal{O}_h,\omega_\sigma) \in \mathcal{V}^{\mathrm{cfe}}_{h,\Gamma_D}$ and an effective dual solution $\bar{P}(\mathcal{O}_h,\omega_\sigma) \in \mathcal{V}^{\mathrm{cfe}}_{h,\Gamma_D}$

as the following linear combinations of U^{km} and P^{km} (cf. (2.4), (2.6)), respectively:

$$\bar{U}(\mathcal{O}_h,\omega_\sigma) = \sum_{k=1}^{K}\sum_{m=1}^{M} \alpha_k(\omega_\sigma)\beta_m(\omega_\sigma) U^{km}(\mathcal{O}_h)\,,$$
$$\bar{P}(\mathcal{O}_h,\omega_\sigma) = \sum_{k=1}^{K}\sum_{m=1}^{M} \alpha_k(\omega_\sigma)\beta_m(\omega_\sigma) P^{km}(\mathcal{O}_h)\,.$$

Given the discrete primal solution $\bar{U}(\mathcal{O}_h,\omega_\sigma)$ the variation of the objective functional (cf. (1.5))

$$\mathcal{J}(\Phi,\omega_\sigma) = \int_{\mathcal{O}_h} j\left(\bar{U}(\mathcal{O}_h,\omega_\sigma)\right)\,dx + \gamma\int_{\partial\mathcal{O}_h} \mathrm{d}\mathcal{H}^{d-1}$$

for a particular realization ω_σ of the stochastic loading and a shape domain $\mathcal{O}_h$ implicitly defined by the discrete level set function Φ (that is, $\mathcal{O}_h = \{\Phi < 0\}$) can be computed as follows (cf. (1.16) and (2.8)):

$$\begin{aligned}\mathcal{J}'(\Phi,\omega_\sigma)(\Xi) &= \mathbf{J}'(\mathcal{O}_h,\omega_\sigma)\left(-\Xi\,|\nabla\Phi|^{-1}N\right)\\ &= \int_{\partial\mathcal{O}_h}\left(-\Xi|\nabla\Phi|^{-1}\right)\Big(j\left(\bar{U}(\mathcal{O}_h,\omega_\sigma)\right) + \gamma\,H\\ &\qquad - \left(f_i(\omega_\sigma) + g_i(\omega_\sigma)\,H + \partial_N g_i(\omega_\sigma)\right)\bar{P}_i(\mathcal{O}_h,\omega_\sigma)\\ &\qquad + A_{ijkl}e_{ij}\left(\bar{U}(\mathcal{O}_h,\omega_\sigma)\right)e_{kl}\left(\bar{P}(\mathcal{O}_h,\omega_\sigma)\right)\Big)\,\mathrm{d}\mathcal{H}^{d-1}\,.\end{aligned} \tag{3.5}$$

Here, N denotes the outer normal on $\partial\mathcal{O}_h$ and H a discrete mean curvature function on $\partial\mathcal{O}_h$. As a suitable approximation we consider N and H to be piecewise affine on $\partial\mathcal{O}_h$. The discrete mean curvature vector $H\,N$ is defined on each vertex X on $\partial\mathcal{O}_h$ as the gradient vector of the area functional with respect to the position of the vertex (cf. [25] for the resulting formula and the relation to the continuous mean curvature). For the numerical integration we apply a Gauss quadrature of degree 4. Hence, the integration is exact as long as f and g are (piecewise) affine functions on $\mathbb{R}^d$. Finally, the discrete counterpart of the shape derivative of our actual stochastic cost functional, namely, the expectation of the discrete cost functional $Q_{\mathbb{E},h}(\mathcal{O}_h)$ in the case of S scenarios $(\omega_\sigma)_{\sigma=1,\dots,S}$ (cf. (3.3)), is given by

$$Q'_{\mathbb{E},h}(\mathcal{O}_h)(V) = \mathbb{E}_\omega\left(\mathcal{J}'(\Phi,\omega)(\Xi)\right) = \sum_{\sigma=1}^{S}\pi_\sigma\mathcal{J}'(\Phi,\omega_\sigma)(\Xi)\,, \tag{3.6}$$

where $V = -\Xi|\nabla\Phi|^{-1}N$ is the normal variation corresponding to the variation Ξ of the level set function Φ. In the algorithm this shape derivative can be used as a descent direction. In any step of the considered time-discrete gradient descent (3.4) one has to compute for the current discrete domain $\mathcal{O}_h$ once the $K\,M$ discrete primal base deformations $U^{km}(\mathcal{O}_h)$ and the corresponding discrete dual base states $P^{km}(\mathcal{O}_h)$. From these, we can efficiently compute the effective deformations $\bar{U}(\mathcal{O},\omega_\sigma)$ and the effective dual states $\bar{P}(\mathcal{O},\omega_\sigma)$ for a possibly very large set of scenarios $\{\omega_\sigma\,|\,\sigma = 1,\dots,S\}$, and using (3.5) and (3.6) we then evaluate the stochastic descent direction.

4. Computational results. As discussed above the major characteristic of two-stage stochastic shape optimization investigated here is that one first decides the domain $\mathcal{O}$, and then the stochastic loading is observed. Hence, we expect the resulting

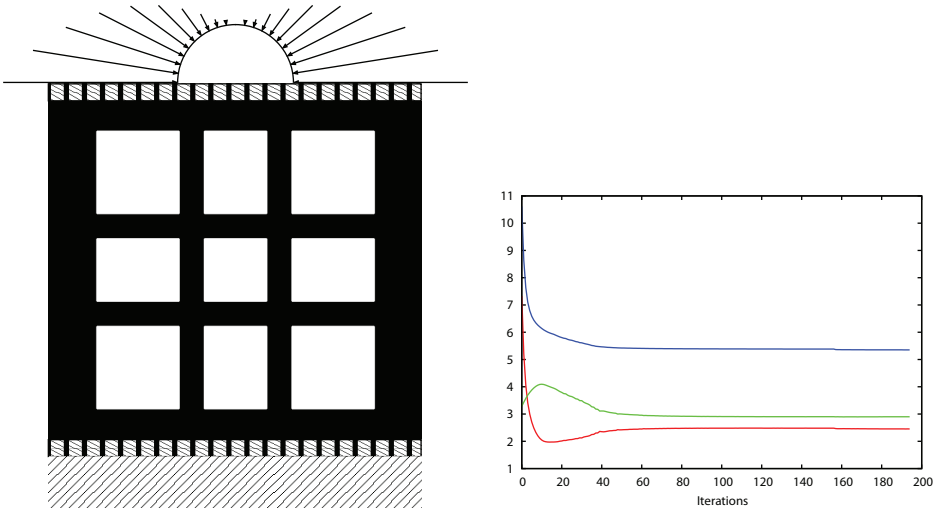


FIG. 4.1. *The initial domain considered in the computation of the optimal shapes in Figures* 4.2 *and* 4.3 *is depicted on the left. On the right the different contributions to the objective function are plotted over the number of iterations. The upper curve shows the robust decay of the objective functional, whereas the lower curve and the middle curve display the evolution and the interplay of the compliance functional and the enclosed volume term, respectively.*

optimal shapes to differ significantly from those obtained in the case of an optimization for the load straightforwardly computed as the expected value of the stochastic loads. In what follows we consider shape optimization applications in two and three dimensions which in particular reflect this consideration. Let us assume a vanishing volume load $f(\omega)$ and Neumann boundary conditions $g(\omega)$ with support Γ_N. As explained above, we assume neither Γ_D nor Γ'_N not to be modified in the actual shape optimization. Indeed, we choose $7h$ as the size of this neighborhood of Γ_D and Γ_N, where the level set function is kept fixed. As the objective function, we take into account a sum of the expectation of the compliance load $\int_{\Gamma_N} g(\omega) \cdot u(\mathcal{O}, \omega)\, \mathrm{d}\mathcal{H}^{d-1}$ and the weighted volume $\eta \int_{\partial\mathcal{O}} \mathrm{d}\mathcal{H}^{d-1}$ of the structure, where η is a positive constant.

4.1. 2D carrier plate. The first application in 2D is a carrier plate, where we optimize the shape of the carrier construction between a floor slap, whose lower boundary is assumed to be the Dirichlet boundary, and the upper plate, on which the loading is applied. Figure 4.1 depicts the initial shape and a sketch of a particular instance of the stochastic loading on the upper plate. Figures 4.2, 4.3, and 4.4 show results obtained by the stochastic optimization algorithm presented here. Each realization of the stochastic load is spatially uniform on the upper plate; realizations only differ by the direction of the force. Hence, two base loads g^1 and g^2 are required to span a load space containing all realizations of the stochastic load. Hence, $m = 2$, whereas S ranges from 2 in Figure 4.2 to 20 in Figure 4.3 to 21 in Figure 4.4. In Figure 4.3 a slightly nonsymmetric set of stochastic scenarios is taken into account, whereas the stochastic load configuration in Figure 4.4 is symmetric. The resulting optimal domains reflect this break of symmetry. Both figures show on the single stochastically optimal shape the von Mises stress distribution for different load scenarios. For the stochastic optimization result in Figure 4.2, we have evaluated the relative error in the stress when refining the underlying grid once. Explicitly, we

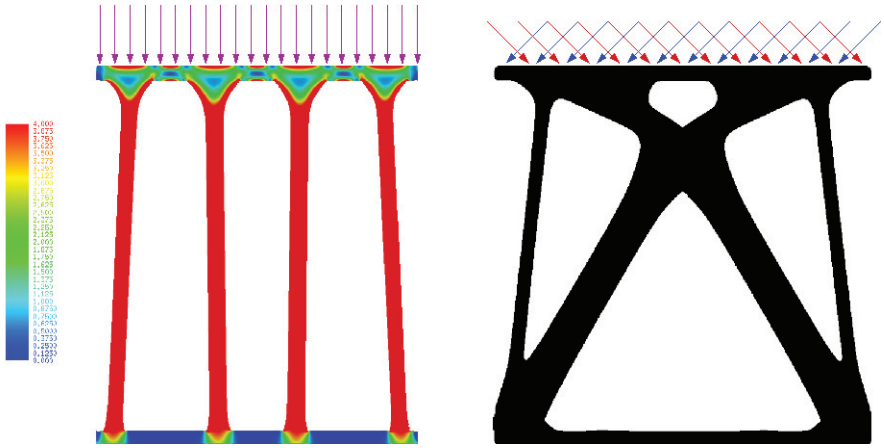


FIG. 4.2. *A direct comparision of two-scale stochastic optimization and deterministic optimization for an averaged load is shown. On the right, a stochastically optimal shape is rendered together with the two underlying load scenarios ω_1 and ω_2 on the upper plate, with surface loads $g(\omega_1)$ and $g(\omega_2)$ both with probability $\frac{1}{2}$. On the left the optimal shape colorcoded with the von Mises stress is drawn for a deterministic load $\frac{1}{2}g(\omega_1) + \frac{1}{2}g(\omega_2)$.*

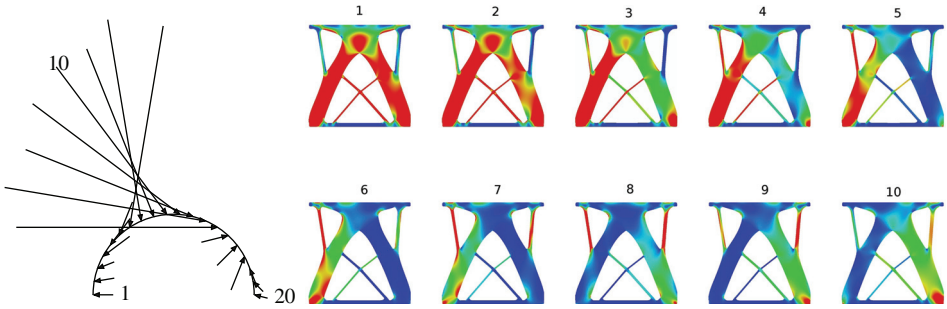


FIG. 4.3. *Stochastic shape optimization based on* 20 *scenarios is depicted. On the left the different loads $g(\omega_\sigma)$ with probabilities π_σ are sketched. Each arrow represents one scenario where the arrow length is determined by the corresponding force intensity weighted with the probability π_σ of the corresponding scenario. On the right the von Mises stress distribution is color coded on the optimal shape for* 10 *out of the* 20 *realizations of the stochastic loading. Due to the nonsymmetric loading configuration the resulting shape is nonsymmetric as well. In particular, the right carrier is significantly thicker than the left one, whereas the connecting diagonal stray pointing up right is thinner than the one point down left.*

obtain a relative error $\int_{\mathcal{O}}[A_{ijkl}e_{ij}(\bar{U}_h(\mathcal{O}) - \bar{U}_{\frac{h}{2}}(\mathcal{O}))]^2\,\mathrm{d}x / \int_{\mathcal{O}}[A_{ijkl}e_{ij}(\bar{U}_h(\mathcal{O}))]^2\,\mathrm{d}x$ of about 0.25 percent. Here, $\bar{U}_h(\mathcal{O})$ is the solution for the grid size $h = 2^{-8}$ and $\bar{U}_{\frac{h}{2}}(\mathcal{O})$ the corresponding solution on grid size $h = 2^{-9}$ for the same discrete domain $\mathcal{O}$.

4.2. 2D cantilever. The second application deals with shape optimization of a 2D cantilever. The initial domain and the optimal shape in the case of deterministic loading are shown in Figure 4.5. Here, the cantilever is fixed on the left side, and a downward pointing force is applied on the right. In Figure 4.6 the dependence of the computed optimal shape on the initial domain is depicted. In Figure 4.7 we focus on the coefficient in front of the volume penalty term and its impact on the optimal shape. A stochastic counterpart is presented in Figure 4.8 with 21 different scenarios

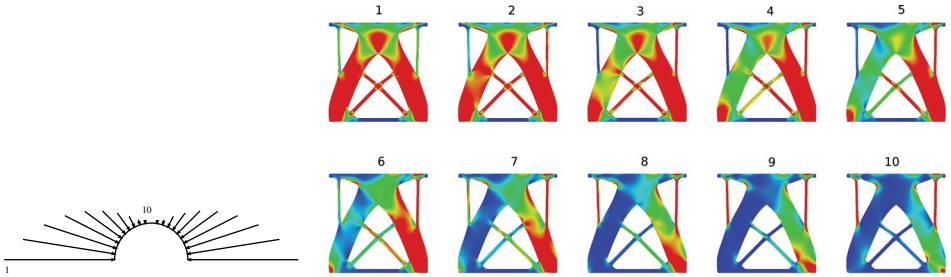


FIG. 4.4. *Results for a symmetric load configuration with* 21 *scenarios, to be contrasted with those reported with a nonsymmetric configuration in Figure* 4.3. *Again on the left the configuration is sketched, and on the right the von Mises stress distribution in plotted in the case of the first* 10 *scenarios.*

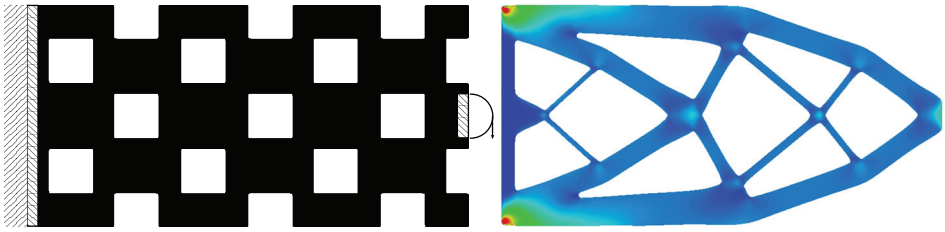


FIG. 4.5. *The initial domain for the computation in the case of a cantilever geometry is rendered on the left. The left boundary is a Dirichlet boundary where the cantilever is attached to a vertical wall. The center part of the right boundary is the support* Γ_N *of the boundary force, which is a deterministic downward-pointing force in this sketch. The resulting optimal shape computed by the proposed level set algorithm is plotted on the right and color coded with the von Mises stress. The corresponding stochastic case is reported in Figure* 4.8.

pulling in different directions. Again the realizations of the stochastic load on the smaller plate on the right are spatially uniform. Thus, the space of realization is 2D, and we can choose $m = 2$.

The subset of the domain Ω, which does not undergo an optimization but is still treated as elastic material, is indicated by the hatched box texture in Figures 4.1 and 4.5. The diameter of the initial domain is 0.9, and the Lamé coefficients in all instances are $\lambda = 40$ and $\mu = 40$. For the parameters in the objective functional we choose $\eta = 8$ in the application in Figures 4.2, 4.3, and 4.4, whereas $\eta = 0.3$ in the case of Figure 4.5. Here, instead of a regularizing surface area term we consider an iterative regularization strategy based on a weaker morphological operator applied during the gradient descent. In all 2D computations the underlying grid is a uniform grid with 257×257 nodes; the discrete primal and dual state equations are solved using a conjugate gradient approach. Furthermore, we take into account $\beta = 0.2$ for the parameter in the Armijo rule and reduce halve the step size as required. Finally, we set $\rho = 6h$ for the computations in Figures 4.3 and 4.4 and $\rho = 4h$ in Figures 4.2 and 4.5, where ρ is the filter parameter in the regularized gradient descent. As mentioned above, we regularize the discrete shape boundary after a couple of iterations applying the morphological operator $D(s)E(2s)D(s)$, where $D(\cdot)$ and $E(\cdot)$ are discrete dilation and erosion operators, respectively. These operators are implemented via a fast marching method [51]. We set $s = 0.5h$ for the width parameter of these operators. Starting from the initial configuration, the decay of the

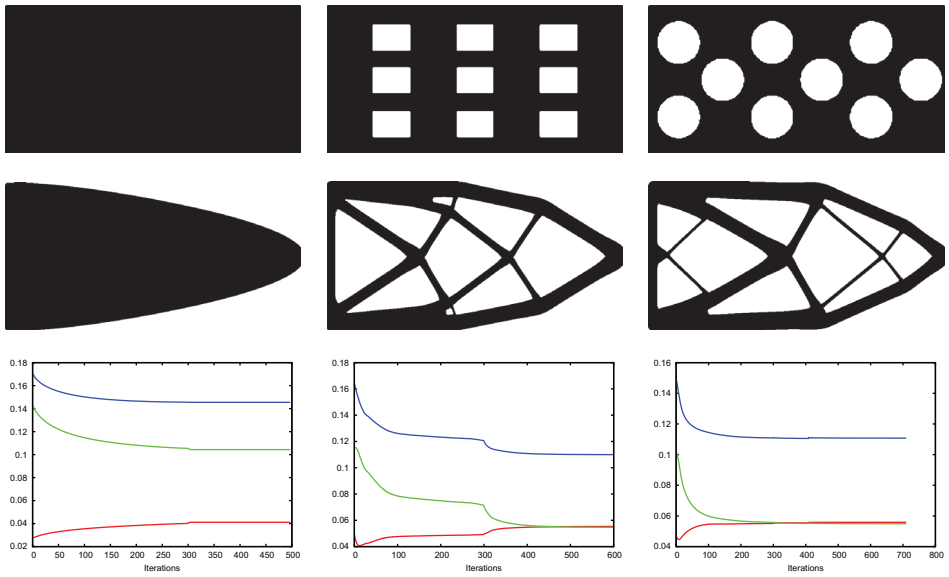


FIG. 4.6. *Results for different initial shapes for the deterministic cantilever computation (see Figure* 4.5*). The top row shows the intial guess. The corresponding optimal shapes and energy plots are depicted in the second and third rows, respectively. In all cases,* η *is fixed to* 0.3*. The middle and right simulation results are obviously local minima with values of the cost functional that are fairly close, as indicated by the error plot.*

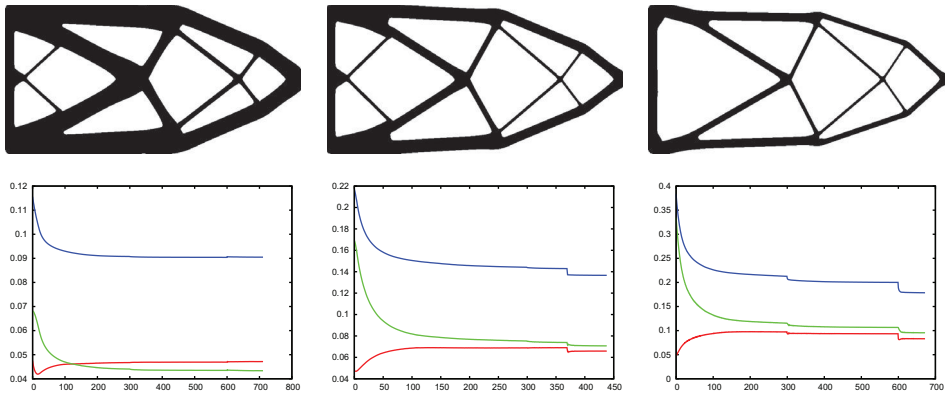


FIG. 4.7. *Results for variations of the volume penalization parameter* η*. In all shown test runs, the initial shape shown in Figure* 4.6 *on the right was used. From left to right, the optimal solutions correspond to the choices* $\eta = 0.2$*,* $\eta = 0.5$*, and* $\eta = 1$*.*

different energy contributions is plotted already on the right-hand side in Figure 4.1. The underlying stochastic scenario is shown in Figure 4.4.

4.3. VSS and EVPI. As stochastic programs are known to be computationally hard to solve, the question arises whether the additional effort pays off compared to solving simpler deterministic problems. There are two common concepts to measure the quality of the stochastic solution: the *value of the stochastic solution* (VSS) and the *expected value of perfect information* (EVPI) (see [16] for details). We computed these two values for the instance shown in Figure 4.2. The optimal objective

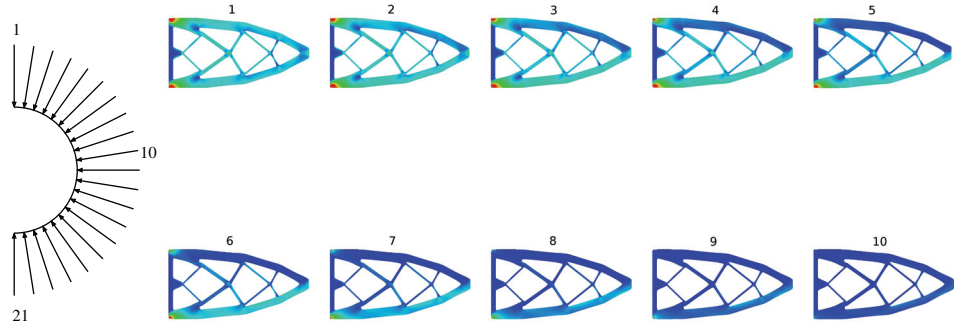


FIG. 4.8. *Stochastic shape optimization in the cantilever case with* 21 *scenarios. The different loads* $g(\omega_\sigma)$ *with probabilities* π_σ *are sketched on the left. The von Mises stress distribution is color coded on the stochastically optimal shape for* 10 *out of the* 21 *scenarios.*

TABLE 4.1

Let $\mathcal{O}_1$ *denote the optimal shape from Figure* 4.3 *and* $\mathcal{O}_2$ *the one from Figure* 4.4. *The table shows the cost functionals arising from the different stochastic loadings shown in Figures* 4.3 *and* 4.4, *respectively, evaluated at* $\mathcal{O}_1$ *and* $\mathcal{O}_2$.

	$\mathcal{O}_1$	$\mathcal{O}_2$
objective from Figure 4.3	4.32398	4.4342
objective from Figure 4.4	5.54182	5.35328

value of the *recourse problem* (2.1) is denoted by RP, and we consider the following deterministic program, which is called the *expected value problem*:

$$\mathrm{EV} := \min \{\mathbf{J}(\mathcal{O}, \bar{\omega}) \ : \ \mathcal{O} \in \mathcal{U}_{ad}\},$$

where $\bar{\omega}$ indicates that all occuring random variables are substituted by their expectations. Let $\mathcal{O}_{\mathrm{EV}} \in \arg\min\{\mathbf{J}(\mathcal{O}, \bar{\omega}) \ : \ \mathcal{O} \in \mathcal{U}_{ad}\}$. Note that in our example, $\mathcal{O}_{\mathrm{EV}}$ is shown in Figure 4.2 on the left. Next, we can define the *expected result of using the EV solution* as $\mathrm{EEV} := \sum_{\sigma=1}^{S} \pi_\sigma \mathbf{J}(\mathcal{O}_{\mathrm{EV}}, \omega_\sigma)$, which finally leads to the VSS given by $\mathrm{VSS} = \mathrm{EEV} - \mathrm{RP}$. For our particular instance, we have $\mathrm{VSS} = 53.68$, or 94 % of the EEV.

To compute the EVPI, we have to compute the so-called *wait-and-see* solution (WS). If $\mathcal{O}_\sigma$ for $\sigma = 1, \ldots, S$ denote the solutions to the many problems

$$\min \{\mathbf{J}(\mathcal{O}, \omega_\sigma) \ : \ \mathcal{O} \in \mathcal{U}_{ad}\}, \quad \sigma = 1, \ldots, S$$

(and there are as many of those as scenarios), then WS is defined to be $\mathrm{WS} := \sum_{\sigma=1}^{S} \pi_\sigma \mathbf{J}(\mathcal{O}_\sigma, \omega_\sigma)$, and $\mathrm{EVPI} := \mathrm{RP} - \mathrm{WS}$. For our instance, we obtained $\mathrm{EVPI} = 0.24$. Finally, we directly compared in Table 4.1 the values of the cost functional on the two different optimal shapes computed by our algorithm shown in Figures 4.3 and 4.4 for both stochastic load scenarios. Even though the shapes visually do not differ too much, a clear preference is demonstrated for the shape optimized with respect to a particular stochastic load configuration.

4.4. 3D cantilever. Finally, we consider a 3D cantilever as a generalization of the problem considered in the first example in 2D. On one side, a disk-shaped plate is fixed on a wall prescribing zero Dirichlet conditions. On the other side, a small rectangular plate opposite to the center of the disk is considered as Neumann boundary loaded with different deterministic and stochastic boundary forces. Figure 4.9 shows

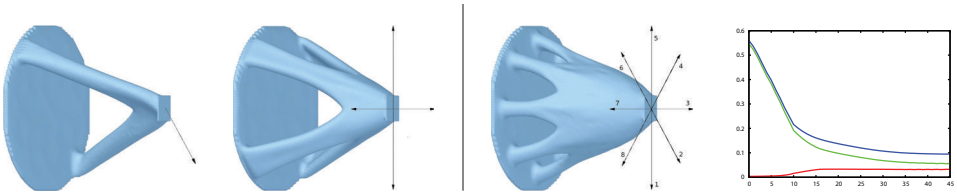


FIG. 4.9. *From left to right the optimal shapes in the deterministic approach and the stochastic optimization approach for one, four, and eight scenarios are shown. The arrows represent the different involved loads $g(\omega_\sigma)$ for varying scenario indices σ. On the right the energy decay is shown for the eight scenario configuration. Again the upper curve represents the total value of the objective functional, the middle one the enclosed volume term, and the lower one the compliance functional.*

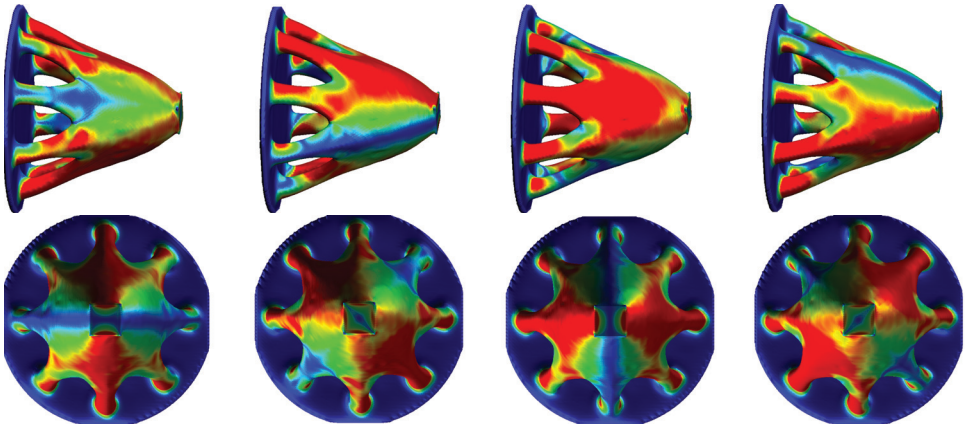


FIG. 4.10. *The optimal design in the case of stochastic shape optimization for the cantilever problem with eight scenerios is depicted. From left to right four scenerios are color coded with the von Mises stress in a consecutive clockwise ordering with respect to the sketch of the loads in Figure* 4.9. *The upper and the lower row show the shape geometry under different perspectives.*

the optimal designs in the case of a single deterministic load and for four and eight stochastic loading scenarios. Furthermore, the energy decay during the numerical relaxation of the shape functional is depicted. As the initial shape we have considered a 3D version of the initial 2D shape shown in Figure 4.1. Figure 4.10 displays a color coding of the von Mises stress distribution on the optimal shape in the stochastic setting with eight equally probable and equally distributed load scenarios.

Here, we choose $\eta = 1$ for the volume penalization parameter, and the elastic behavior is described by the Lame coefficients $\lambda = 40$ and $\mu = 40$ for a structure diameter of the order 1. The parameters involved in the Armijo step control are the same as those in the 2D applications. The underlying grid is a regular grid with 128^3 nodes. The shape optimization is first performed on a 64^3 grid. Then the level set function is prolongated to the next finer grid level. Before the gradient descent of the shape functional is released a morphological smoothing operator $D(s)E(2s)D(s)$ is applied. Here, as in the 2D case $D(s)$ and $E(s)$ represent discrete dilation and erosion operators, implemented based in a fast marching algorithm in 3D. As the width parameter we select $s = 0.45h$. The filter parameter in the regularized gradient descent is $\rho = 2.5h$. A multigrid method for the numerical solution of the discrete primal and dual problem is applied.

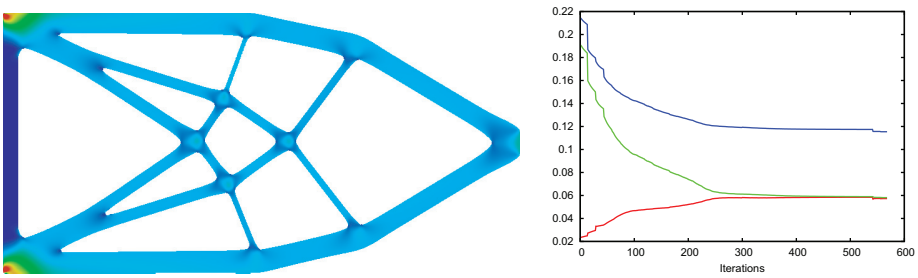


FIG. 4.11. *The optimal shape for a cantilever problem with deterministic loading is computed based on a combined level-set and topogical derivative approach. A relaxation step with respect to the topological derivative is considered in every* 15*th step of a general shape gradient descent method (left). Furthermore, the corresponding energies, i.e., the total value of the objective function, the enclosed volume, and the compliance functional are plotted on the right.*

Outlook. Our approach of using two-stage stochastic optimization with the Lamé equation as a variational problem does not link only stochastic shape optimization to two-stage stochastic programming. Indeed, it offers the flexibility to go beyond expected value optimization and addresses risk aversion. To this end, the expectation in (2.1) is replaced by suitable risk measures such as *expected excess* and *excess probability*. The resulting optimal shapes will significantly depend on the chosen risk measure. Furthermore, we included results on a combination of level-set and topological shape optimization which can be extended to the case of uncertain loadings (cf. Figure 4.11). For a detailed discussion of these issues we refer to a forthcoming publication [20].

REFERENCES

[1] S. ADALI, JR., J. C. BRUCH, I. S. SADEK, AND J. M. SLOSS, *Robust shape control of beams with load uncertainties by optimally placed piezo actuators*, Struct. Multidiscip. Optim., 19 (2000), pp. 274–281.

[2] D. ADALSTEINSSON AND J. A. SETHIAN, *A fast level set method for propagating interfaces*, J. Comput. Phys., 118 (1995), pp. 269–277.

[3] S. ALBERS, *Online algorithms: a survey*, Math. Program., 97 (2003), pp. 3–26.

[4] G. ALLAIRE, *Shape Optimization by the Homogenization Method*, Appl. Math. Sci. 146, Springer, New York, 2002.

[5] G. ALLAIRE, E. BONNETIER, G. FRANCFORT, AND F. JOUVE, *Shape optimization by the homogenization method*, Numer. Math., 76 (1997), pp. 27–68.

[6] G. ALLAIRE, DE F. GOURNAY, F. JOUVE, AND A.-M. TOADER, *Structural optimization using topological and shape sensitivity via a level set method*, Control and Cybernetics, 34 (2005), pp. 59–80.

[7] G. ALLAIRE AND F. JOUVE, *A level-set method for vibration and multiple loads structural optimization*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 3269–3290.

[8] G. ALLAIRE, F. JOUVE, AND H. MAILLOT, *Topology optimization for minimum stress design with the homogenization method*, Struct. Multidiscip. Optim., 28 (2004), pp. 87–98.

[9] G. ALLAIRE, F. JOUVE, AND A.-M. TOADER, *Structural optimization using sensitivity analysis and a level-set method*, J. Comput. Phys., 194 (2004), pp. 363–393.

[10] F. ALVAREZ AND M. CARRASCO, *Minimization of the expected compliance as an alternative approach to multiload truss optimization*, Struct. Multidiscip. Optim., 29 (2005), pp. 470–476.

[11] P. ATWAL, *Kontinuumstheorie für eine diskrete Mikrostruktur zwecks Formoptimierung*, Diplomarbeit, Universität Duisburg-Essen, Duisburg, Germany, 2006.

[12] F. BASTIN, C. CIRILLO, AND P. L. TOINT, *Convergence theory for nonconvex stochastic programming with an application to mixed logit*, Math. Program., 108 (2006), pp. 207–234.

[13] A. BEN-TAL, M. KOČVARA, A. NEMIROVSKI, AND J. ZOWE, *Free material design via semidefinite programming: The multiload case with contact conditions*, SIAM J. Optim., 9 (1999), pp. 813–832.

[14] A. Ben-Tal and A. Nemirovski, *Robust Optimization - methodology and applications*, Math. Program., 92 (2002), pp. 453–480.
[15] M. P. Bendsøe, *Optimization of Structural Topology, Shape, and Material*, Springer, Berlin, 1995.
[16] J. R. Birge and F. Loveaux, *Introduction to Stochastic Programming*, Springer, New York, 1997.
[17] M. Burger and S. J. Osher, *A survey on level set methods for inverse problems and optimal design*, European J. Appl. Math., 16 (2005), pp. 263–301.
[18] F. R. Chang, *Stochastic Optimization in Continuous Time*, Cambridge University Press, Cambridge, 2004.
[19] P. G. Ciarlet, *Mathematical Elasticity, Volume* I: *Three-Dimensional Elasticity*, Stud. Math. Appl. 20, North–Holland, Amsterdam, 1988.
[20] S. Conti, H. Held, M. Pach, M. Rumpf, and R. Schultz, *Risk averse shape optimization*, in preparation.
[21] M. C. Delfour, *Oriented distance funtion and its evolution equation for initial sets with thin boundary*, SIAM J. Control Optim., 42 (2004), pp. 2286–2304.
[22] M. C. Delfour and J. P. Zolésio, *Shapes and Geometries: Analysis, Differential Calculus, and Optimization*, Adv. Des. Control 4, SIAM, Philadelphia, 2001.
[23] G. P. Dias, J. Herskovits, and F. A. Rochinha, *Simultaneous shape optimization and nonlinear analysis of elastic solids*, in Proceedings of the Fourth World Congress on Computational Mechanics, S. Idelsohn, E. Oñate, and E. Dvorkin, eds., Buenos Aires, Argentina, 1998, CIME, Barcelona, Spain, 1998.
[24] Q. Du and D. Wang, *Tetrahedral mesh generation and optimization based on centroidal Voroni tesselations*, Internat. J. Numer. Methods Engrg., 56 (2003), pp. 1355–1373.
[25] G. Dziuk, *An algorithm for evolutionary surfaces*, Numer. Math., 58 (1991), pp. 603–611.
[26] L. C. Evans, *Partial Differential Equations*, Grad. Stud. Math. 19, American Mathematical Society, Providence, RI, 1998.
[27] W. H. Fleming and R. W. Rishel, *Deterministic and Stochastic Optimal Control*, Springer, New York, 1975.
[28] J. Gomes and O. Faugeras, *Reconciling distance functions and level sets*, in Scale-Space Theories in Computer Vision, Second International Conference Proceedings, Scale-Space '99, Corfu, Greece, 1999, Lecture Notes in Comput. Sci. 1682, M. Nielsen, P. Johansen, O. F. Olsen, and J. Weickert, eds., Springer, Berlin, 1999, pp. 70–81.
[29] J. M. Guedes, H. C. Rodrigues, and M. P. Bendsøe, *A material optimization model to approximate energy bounds for cellular materials under multiload conditions*, Struct. Multidiscip. Optim., 25 (2003), pp. 446–452.
[30] W. Hackbusch, *Multi-grid Methods and Applications*, Springer Ser. Comput. Math. 4, Springer, New York, 1985.
[31] W. Hackbusch and S. Sauter, *Composite finite elements for the approximation of PDEs on domains with complicated micro-structures*, Numer. Math., 75 (1997), pp. 447–472.
[32] W. Hackbusch and S. A. Sauter, *Composite Finite Elements for Problems with Complicated Boundary. Part* III: *Essential Boundary Conditions*, Technical report, Universität Kiel, Kiel, Germany, 1997.
[33] P. Kall and S. W. Wallace, *Stochastic Programming*, Wiley, Chichester, 1994.
[34] C.-Y. Kao, S. Osher, and Y.-H. Tsai, *Fast sweeping methods for static Hamilton–Jacobi equations*, SIAM J. Numer. Anal., 42 (2005), pp. 2612–2632.
[35] F. Liehr, T. Preusser, M. Rumpf, S. Sauter, and L. O. Schwen, *Composite finite elements for* 3*D image based computing*, Comput. Vis. Sci.. submitted.
[36] Z. Liu, J. G. Korvink, and R. Huang, *Structure topology optimization: Fully coupled level set method via femlab*, Struct. Multidiscip. Optim., 29 (2005), pp. 407–417.
[37] W. E. Lorensen and H. E. Cline, *Marching cubes: A high resolution* 3*D surface construction algorithm*, Computer Graph., 21 (1987), p. 163.
[38] R. Malladi and J. A. Sethian, *An O(N log N) algorithm for shape modeling*, in Proc. Natl. Acad. Sci. USA 93 (1996), pp. 9389–9392.
[39] J. E. Marsden and T. J. R. Hughes, *Mathematical Foundations of Elasticity*, Prentice–Hall, Englewood Cliffs, NJ, 1983.
[40] K. Marti, *Stochastic Optimization Methods*, Springer, Berlin, 2005.
[41] R. E. Melchers, *Optimality-criteria-based probabilistic structural design*, Struct. Multidiscip. Optim., 23 (2001), pp. 34–39.
[42] S. Osher and R. Fedkiw, *Level Set Methods and Dynamic Implicit Surfaces*, Appl. Math. Sci. 153, Springer, New York, 2003.
[43] S. J. Osher and J. A. Sethian, *Fronts propagating with curvature dependent speed: Algorithms based on Hamilton–Jacobi formulations*, J. Comput. Phys., 79 (1988), pp. 12–49.

[44] S. J. Owen, *A survey of unstructured mesh generation technology*, in Proceedings of the 7th International Meshing Roundtable, Dearborn, MI, Sandia National Laboratories, 1998, pp. 239–267.
[45] M. Pach, *Levelsetverfahren in der Shapeoptimierung*, Diploma thesis, University Duisburg, Duisburg, Germany, 2005.
[46] T. Pennanen, *Epi-convergent discretizations of multistage stochastic programs*, Math. Oper. Res., 30 (2005), pp. 245–256.
[47] A. Prékopa, *Stochastic Programming*, Kluwer, Dordrecht, the Netherlands, 1995.
[48] A. Ruszczyński, *Some advances in decomposition methods for stochastic linear programming*, Ann. Oper. Res., 85 (1999), pp. 153–172.
[49] A. Ruszczyński and A. Shapiro, eds., *Stochastic Programming*, Handbooks Oper. Res. Management Sci. 10, Elsevier, Amsterdam, 2003.
[50] S. Sauter, *Vergröberung von Finite-Elemente-Räumen*, Habilitationsschrift, Universität Kiel, Kiel, Germany, 1997.
[51] J. A. Sethian, *Level Set Methods and Fast Marching Methods*, Cambridge Monogr. Appl. Comput. Math. 3, Cambridge University Press, Cambridge, 1999.
[52] J. A. Sethian and A. Wiegmann, *Structural boundary design via level set and immersed interface methods*, J. Comput. Phys., 163 (2000), pp. 489–528.
[53] R. M. Van Slyke and R. S. Wets, *L-shaped linear programs with application to optimal control and stochastic programming*, SIAM J. Appl. Math., 17 (1969), pp. 638–663.
[54] J. Sokolowski and J.-P. Zolésio, *Introduction to Shape Optimization Shape Sensitivity Analysis*, Springer Ser. Comput. Math., Springer, New York, 1992.
[55] M. C. Steinbach, *Tree-sparse convex programs*, Math. Methods Oper. Res., 56 (2002), pp. 347–376.
[56] G. M. Treece, R. W. Prager, and A. H. Gee, *Regularized marching tetrahedra: Improved iso-surface extraction*, Comput. Graph., 23 (1999), pp. 583–598.
[57] J. Tsitsiklis, *Efficient algorithms for globally optimal trajectories*, IEEE Trans. Automat. Control, 40 (1995), pp. 1528–1538.
[58] S. W. Wallace and W. T. Ziemba, *Applications of Stochastic Programming*, MPS SIAM Ser. Optim., SIAM, Philadelphia, 2005.
[59] C. Zhuang, Z. Xiong, and H. Ding, *A level set method for topology optimization of heat conduction problem under multiple load cases*, Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 1074–1084.

SIAM J. OPTIM.
Vol. 19, No. 4, pp. 1633–1647

ERROR BOUNDS FOR CONVEX POLYNOMIALS*

W. H. YANG†

Abstract. The purpose of this paper is to investigate error bounds for convex polynomials. We prove that for a convex polynomial f in n variables which is not everywhere positive and which is not constant on any affine subspace, either f is a sum of a convex polynomial in fewer variables and a linear form with negative coefficients or the negativity set of f is compact. As an application, we deduce various types of error bounds for unconstrained and polyhedral-constrained convex polynomials.

Key words. error bound, convex polynomial, kernel, recession cone

AMS subject classifications. 90C25, 90C31

DOI. 10.1137/070689838

1. Introduction.

Multivariate polynomial minimization has received much attention over the years thanks to its application in engineering and economics. It has been studied extensively, and there are many works devoted to this subject. The reader is referred to [6, 7, 11, 15, 17] and also the references therein. A popular method for solving the polynomial minimization problem is the sum-of-squares (SOS) method, which was first introduced by Shor [17] and further developed by Nesterov [11], Lasserre [7], and Parrilo [14]. The SOS method is based on semidefinite programming, for which efficient algorithms are now available. Another important approach for polynomial minimization is based on tensor analysis, which involves various computational topics of higher-order tensors, such as tensor decomposition, computation of tensor rank, computation of tensor eigenvalues, and so on. For the interested reader, see [15] for a survey on the state-of-the-art knowledge on this topic.

In this paper, we will study the properties of convex multivariate polynomials (convex polynomials for short), since it is the simplest type of nonlinear and nonquadratic polynomial. As far as we know, Belousov was the first one who used convex analysis to study polynomial minimization. In 1977, Belousov derived some basic properties of convex polynomials in his book [5]. Following a way similar to that proposed by Belousov, Bank and Mandel [1] extended the results in [5] to quasi-convex polynomials. In [4], Belousov and Klatte generalized the Frank–Wolfe-type theorem to convex polynomial systems. The impetus of this manuscript came from Lemmas 1 and 2 in [4]. We will present some further results on convex polynomials using convex analysis. As an application, we derive various error bounds for unconstrained and polyhedral-constrained convex polynomials.

Recently, error bounds have found important applications in various areas in mathematical programming such as sensitivity analysis, convergence analysis of algorithms, and asymptotic analysis. There is plenty of literature on this subject. The reader is referred to the papers [13, 18, 20] and references therein for the theory and applications of error bounds. When considering the polynomial inequality systems, Hölderian error bounds have been demonstrated for these systems in [8, 9, 10, 12, 19].

*Received by the editors April 28, 2007; accepted for publication (in revised form) September 15, 2008; published electronically January 21, 2009.
http://www.siam.org/journals/siopt/19-4/68983.html
†School of Mathematical Sciences, Fudan University, Shanghai 200433, People's Republic of China (whyang@fudan.edu.cn).

It is proved in this paper that for an mth-order convex polynomial f, if the negativity set of f is nonempty, then f has a linear error bound. Otherwise, f has a local Hölderian error bound of order $1/m$.

The rest of the paper is organized as follows. In section 2, we introduce some notations and definitions on convex analysis. Some preliminary results on convex polynomials are presented. In section 3, we show that for a convex polynomial f which is not constant on any affine subspace, if the lower level set of f is unbounded, then f can be represented as a sum of a convex polynomial in fewer variables and a linear form with negative coefficients. We also investigate some properties of the kernel of a homogeneous convex polynomial. Finally, in section 4, we identify exactly and establish various types of error bounds for unconstrained and polyhedral-constrained convex polynomials.

2. Definitions and preliminary results. In this section, we give the notations, definitions, and preliminary results which will be used throughout the paper. For a positive integer n, we use $[n]$ to denote the set $\{1,\dots,n\}$. For $x \in \mathbb{R}^n$, x_i ($i \in [n]$) is the ith component of x as usual. The transpose of x is denoted by x^T. For a subspace $L \subset \mathbb{R}^n$, $L^\perp$ denotes the orthogonal complement of L in $\mathbb{R}^n$. The dimension of L is denoted by $\dim(L)$.

For a function $\varphi : \mathbb{R}^n \to \mathbb{R}$, the lower level set of φ is defined by $S_\varphi := \{x \in \mathbb{R}^n : \varphi(x) \le 0\}$. We say that φ has an error bound of order γ if there exists $\tau > 0$ such that

$$d(x, S_\varphi) \le \tau [\varphi(x)]_+^\gamma \quad \forall x \in \mathbb{R}^n,$$

where $[\varphi(x)]_+ = \max\{\varphi(x), 0\}$ and $d(x, S_\varphi)$ denotes the distance from x to S_φ. If φ has an error bound of order 1, we also say φ has a linear error bound. The optimal solution set of φ is denoted by $\operatorname{argmin}_{x\in\mathbb{R}^n}\varphi(x)$.

Let $C \subset \mathbb{R}^n$ be a closed convex set. We use $\mathrm{bd}(C)$ to denote the boundary of C. For $x \in C$, the normal cone of C at x is defined by

$$N_C(x) := \left\{ z \in \mathbb{R}^n : z^T(y - x) \le 0 \ \ \forall y \in C \right\}.$$

Let $N_C^1(x) := \{h \in N_C(x) : \|h\| = 1\}$ denote the set of all unit vectors in $N_C(x)$. The recession cone C^∞ [16] of C is defined by

$$C^\infty := \{d \in \mathbb{R}^n : x + td \in C, \ \ \forall t \ge 0, \ \forall x \in C\}.$$

It is easy to see that C^∞ is a convex cone. For a convex function ψ, by [16, Theorem 8.7], all the nonempty level sets of the form $\{x : \psi(x) \le c\}$, $c \in \mathbb{R}$, have the same recession cone, namely, the recession cone of ψ. Without loss of generality, we use the notation S_ψ^∞ to denote the recession cone of ψ. Let $E = S_\psi^\infty \cap (-S_\psi^\infty)$. Then E is a subspace and is called the constancy space [16, p. 69] of ψ.

LEMMA 2.1 (see [16, p. 69]). *The constancy space E is the largest subspace contained in S_ψ^∞ which satisfies*

$$E = \{z \in \mathbb{R}^n \ : \ \psi(x + \lambda z) = \psi(x), \ \forall x \in \mathbb{R}^n, \ \forall \lambda \in \mathbb{R}\}.$$

For a convex function ψ, we denote the set of subgradients of ψ at $x \in \mathbb{R}^n$ by $\partial\psi(x)$. It is well known that the directional derivative $\psi'(x;h) = \lim_{t\to 0+} \frac{\psi(x+th)-\psi(x)}{t}$ always exists for $x, \ h \in \mathbb{R}^n$, and one has that

$$\psi'(x;h) = \max\left\{\eta^T h : \eta \in \partial\psi(x)\right\}. \tag{2.1}$$

Let $\mathbb{Z}^n_+ = \{\beta = (\beta_1, \ldots, \beta_n) : \beta_i \in \mathbb{Z}, \beta_i \geq 0,\ \forall i \in [n]\}$. For $\beta \in \mathbb{Z}^n_+$, we use x^β to denote the product $x_1^{\beta_1}, \ldots, x_n^{\beta_n}$ for every $x \in \mathbb{R}^n$. Then, for an mth-order polynomial f on $\mathbb{R}^n$, we can write it as $f(x) = \sum_{|\beta|=0}^m a_\beta x^\beta\ \forall x \in \mathbb{R}^n$, where $\beta \in \mathbb{Z}^n_+$, $a_\beta \in \mathbb{R}$, and $|\beta| = \sum_{i=1}^n \beta_i$. In this paper, we always use f_l, $0 \leq l \leq m$, to denote the lth-order homogeneous polynomial corresponding to f, that is, $f_l(x) = \sum_{|\beta|=l} a_\beta x^\beta$, and so

$$f = \sum_{l=0}^m f_l. \tag{2.2}$$

Note that f_0 is the constant term of f. If $f(x) \geq 0$ ($f(x) > 0$) $\forall x \in \mathbb{R}^n$ ($x \neq 0$), we say that f is a positive semidefinite (definite) polynomial. It is easy to see that if f is a convex polynomial and $m \geq 2$, then m is an even integer.

The Taylor series of an mth-order polynomial f can be written as

$$f(x+y) = \sum_{|\beta|=0}^m \frac{D^\beta f(x)}{\beta!} y^\beta \quad \forall x,\ \forall y \in \mathbb{R}^n, \tag{2.3}$$

where $\beta \in \mathbb{Z}^n_+$, $D^\beta f(x) := \frac{\partial^{\beta_1}}{\partial x_1^{\beta_1}}, \ldots, \frac{\partial^{\beta_n}}{\partial x_n^{\beta_n}} f(x)$ and $\beta! = \beta_1!, \ldots, \beta_n!$. Using this notation, we have $(\nabla f(x))^T y = \sum_{|\beta|=1} (D^\beta f(x))^T y^\beta,\ \forall x \in \mathbb{R}^n,\ \forall y \in \mathbb{R}^n$.

Let h be a kth-order homogeneous polynomial ($k \geq 1$). We define the kernel of h by

$$\mathrm{Ker}(h) := \left\{x \in \mathbb{R}^n\ :\ D^\alpha h(x) = 0\ \forall \alpha \in \mathbb{Z}^n_+ \text{ satisfying } |\alpha| = k-1\right\}.$$

It is easy to see that if $x \in \mathrm{Ker}(h)$, then $D^\beta h(x) = 0$ for every $\beta \in \mathbb{Z}^n_+$ satisfying $0 \leq |\beta| \leq k-1$. In particular, $x \in \mathrm{Ker}(h)$ implies $h(x) = 0$. Moreover, $\mathrm{Ker}(h)$ is a linear space. If h is a homogeneous quadratic function, that is, $h(x) = x^T A x$ for some matrix A, then $\mathrm{Ker}(h) = \mathrm{Ker}(A)$.

The following lemma is a well-known result, and we omit the proof.

LEMMA 2.2. *Let $f(x) = \sum_{|\beta|=0}^m a_\beta x^\beta$ be a polynomial defined on $\mathbb{R}^n$. If $f(x) = 0\ \forall x \in \mathbb{R}^n$, then $a_\beta = 0$ for every $\beta \in \mathbb{Z}^n_+$ satisfying $0 \leq |\beta| \leq m$.*

The next two lemmas, which play an important role in this paper are due to Belousov (see [4, Lemmas 1 and 2]).

LEMMA 2.3. *Let f be a convex polynomial. Let $x, y, d \in \mathbb{R}^n$. If $\mu(t) = f(x+td)$ defined on $\mathbb{R}$ is a convex polynomial of order p, then $\nu(t) = f(y+td)$, $t \in \mathbb{R}$, is also a convex polynomial of order p. Further, if $p \geq 1$, then the coefficient associated with the term t^p in $\mu(t)$ and $\nu(t)$ are identical.*

LEMMA 2.4. *Let f be a convex polynomial satisfying $S_f \neq \emptyset$. For $d \neq 0$, $d \in S_f^\infty$ if and only if $f(td) = f(0) + rt,\ \forall t \in \mathbb{R}$, for some $r \leq 0$.*

COROLLARY 2.1. *Let $h : \mathbb{R}^n \to \mathbb{R}$ be a kth-order homogeneous convex polynomial, where $k \geq 2$. Then h is positive semidefinite.*

Proof. Suppose that $h(y) < 0$ for some y. Since $h(ty) = t^k h(y) < 0$ for each $t > 0$, we have $ty \in S_h\ \forall t \geq 0$. From [16, Theorem 8.3], it follows that $y \in S_h^\infty$. By Lemma 2.4, we have that $h(ty) = \beta t$ for some $\beta \leq 0$, which contradicts $h(ty) = t^k h(y)$. ∎

Note that a positive semidefinite homogeneous polynomial is not necessarily a convex polynomial. Let $g : \mathbb{R}^2 \to \mathbb{R}$ be defined by $g(x_1, x_2) = (x_1^2 - x_2^2)^2$. It is easy to see that g is positive semidefinite but not convex.

LEMMA 2.5. *Let h be a kth-order homogeneous polynomial ($k \geq 1$), and let $K \subset \mathbb{R}^n$ be a subspace. Then $h(x+y) = h(x)$, $\forall x \in \mathbb{R}^n$, $\forall y \in K$, if and only if $K \subset \mathrm{Ker}(h)$.*

Proof. Note that $h(x+y) = \sum_{|\beta|=0}^{k} \frac{D^\beta h(y)}{\beta!} x^\beta$ and $h(x) = \sum_{|\beta|=k} \frac{D^\beta h(0)}{\beta!} x^\beta$ $\forall x, y \in \mathbb{R}^n$. By Lemma 2.2, $h(x+y) = h(x)$, $\forall x \in \mathbb{R}^n$, $\forall y \in K \iff D^\beta h(y) = 0$, $\forall y \in K$, for each β satisfying $|\beta| \leq k-1$. Thus, the claim holds. ∎

LEMMA 2.6. *Let h be a kth-order homogeneous convex polynomial. Then* $\mathrm{Ker}(h) = \{x : h(x) = 0\}$, *and h is positive definite on* $(\mathrm{Ker}(h))^\perp$.

Proof. Let $H = \{x : h(x) = 0\}$. We need only to prove $H \subseteq \mathrm{Ker}(h)$. For $y \in H$, we have $g(ty) = 0 = g(0)$ $\forall t \in \mathbb{R}$. From Lemma 2.3, it follows that $g(x+ty) = g(x)$ $\forall x \in \mathbb{R}^n$ and $t \in \mathbb{R}$. By Lemma 2.5, one has that $y \in \mathrm{Ker}(h)$, which implies $H \subseteq \mathrm{Ker}(h)$. By Corollary 2.1, h is nonnegative on $\mathbb{R}^n$. Suppose $h(x) = 0$ for some $x \in (\mathrm{Ker}(h))^\perp$. Then $H \subseteq \mathrm{Ker}(h)$ implies $x \in \mathrm{Ker}(h)$, and therefore $x = 0$. ∎

The next lemma shows that if a polynomial f is constant along a subspace, then it can be transformed into a polynomial with less variables by changing the basis.

LEMMA 2.7. *Let $f(x)$ be an mth-order polynomial on $\mathbb{R}^n$, and let $L \subset \mathbb{R}^n$ be a subspace of dimension p. If $f(x+y) = f(x)$ $\forall x \in \mathbb{R}^n$ and $y \in L^\perp$, then there exists an orthogonal matrix U such that*

$$f(Ux) = g(x_1, \ldots, x_p) \quad \forall x \in \mathbb{R}^n, \tag{2.4}$$

for some polynomial g in p variables.

Proof. Let $f = \sum_{l=0}^{m} f_l$. By Lemma 2.5, we have $L^\perp \subset \cap_{l=1}^{m} \mathrm{Ker}(f_l)$. Let U be the matrix such that the first p columns of U are an orthonormal basis of L and the rest of the columns are an orthonormal basis of $L^\perp$. By (2.3) and $L^\perp \subset \cap_{l=1}^{m} \mathrm{Ker}(f_l)$, it is easy to obtain (2.4). ∎

The following result can be proved by Lemmas 2.1 and 2.7 in a straightforward way.

COROLLARY 2.2. *Let f be a convex polynomial on $\mathbb{R}^n$, and let $E = S_f^\infty \cap (-S_f^\infty)$. Assume that* $\dim(E) = n - p$. *Then there exists an orthogonal matrix U such that*

$$f(Ux) = g(x_1, \ldots, x_p) \quad \forall x \in \mathbb{R}^n,$$

where g is a convex polynomial satisfying $S_g^\infty \cap (-S_g^\infty) = \{0\}$.

3. Properties of convex polynomials. In this section, we will prove that if S_f^∞ is not a linear space, then we can simplify the form of f to the sum of a convex polynomial in fewer variables and a linear form. In view of Corollary 2.2, we always assume that $S_f^\infty \cap (-S_f^\infty) = \{0\}$ for each convex polynomial f in the remainder of the paper.

THEOREM 3.1. *Let $f = \sum_{l=0}^{m} f_l$ be a convex polynomial on $\mathbb{R}^n$. Then*

$$\cap_{l=2}^{m} \mathrm{Ker}(f_l) = S_f^\infty - S_f^\infty. \tag{3.1}$$

Moreover, $\dim(\cap_{l=2}^{m} \mathrm{Ker}(f_l)) \leq 1$.

Proof. Let $z \in \cap_{l=2}^m \operatorname{Ker}(f_l)$. Then $tz \in \cap_{l=2}^m \operatorname{Ker}(f_l)\ \forall t \in \mathbb{R}$, which together with (2.3) implies that

$$\begin{aligned} f(x+tz) &= \sum_{l=0}^{m} \sum_{|\beta|=0}^{l} \frac{D^\beta f_l(tz)}{\beta!} x^\beta \\ &= \sum_{l=2}^{m} \sum_{|\beta|=l} \frac{D^\beta f_l(tz)}{\beta!} x^\beta + \sum_{|\beta|=1} D^\beta f_1(tz) x^\beta + f_1(tz) + f_0 \\ &= \sum_{l=2}^{m} \sum_{|\beta|=l} \frac{D^\beta f_l(0)}{\beta!} x^\beta + \sum_{|\beta|=1} D^\beta f_1(0) x^\beta + t f_1(z) + f_0 \\ &= f(x) + t f_1(z) \quad \forall x \in \mathbb{R}^n, \forall t \in \mathbb{R}. \end{aligned} \tag{3.2}$$

If $f_1(z) \le 0$, then $f(x+tz) \le f(x)\ \forall t \ge 0$, and so $z \in S_f^\infty$. Since $0 \in S_f^\infty$, we have $z = z - 0 \in S_f^\infty - S_f^\infty$. If $f_1(z) \ge 0$, then $-z \in S_f^\infty$, and so $z = 0 - (-z) \in S_f^\infty - S_f^\infty$. Thus, $\cap_{l=2}^m \operatorname{Ker}(f_l) \subseteq S_f^\infty - S_f^\infty$.

For (3.1), it suffices to prove $S_f^\infty \subset \cap_{l=2}^m \operatorname{Ker}(f_l)$. Fix $d \in S_f^\infty$. By Lemmas 2.3 and 2.4, there exists $\delta \le 0$ such that $f(x+td) = f(x) + \delta t$ for every $t \in \mathbb{R}$ and $x \in \mathbb{R}^n$. Then

$$\begin{aligned} f(x) + \delta t = f(x+td) &= \sum_{l=0}^{m} \sum_{|\beta|=0}^{l} \frac{D^\beta f_l(td)}{\beta!} x^\beta = \sum_{l=0}^{m} \sum_{|\beta|=0}^{l} \frac{D^\beta f_l(d)}{\beta!} x^\beta t^{l-|\beta|} \\ &= \sum_{k=0}^{m} \left(\sum_{|\beta|=0}^{m-k} \frac{D^\beta f_{|\beta|+k}(d)}{\beta!} x^\beta \right) t^k \quad \forall t \in \mathbb{R}, \forall x \in \mathbb{R}^n. \end{aligned}$$

Hence, $\sum_{|\beta|=0}^{m-1} \frac{D^\beta f_{|\beta|+1}(d)}{\beta!} x^\beta = \delta$ for every $x \in \mathbb{R}^n$. By Lemma 2.2, we have $\frac{D^\beta f_{|\beta|+1}(d)}{\beta!} = 0$ for each $\beta \in \mathbb{Z}_+^n$ satisfying $1 \le |\beta| \le m-1$, that is, $d \in \cap_{l=2}^m \operatorname{Ker}(f_l)$.

Let $X_\infty := \cap_{l=2}^m \operatorname{Ker}(f_l)$. If $\dim(X_\infty) > 1$, there exist $y,\ z \in X_\infty$ such that $y \ne \lambda z\ \forall \lambda \in \mathbb{R}$. We can find real numbers μ and ν such that $f_1(\mu y + \nu z) = 0$ and $\mu y + \nu z \ne 0$. Since $y,\ z \in X_\infty = \cap_{l=2}^m \operatorname{Ker}(f_l)$, similar to (3.2), we have

$$f(x + \lambda(\mu y + \nu z)) = f(x) + \lambda f_1(\mu y + \nu z) = f(x) \qquad \forall x \in \mathbb{R}^n,\ \forall \lambda \in \mathbb{R}.$$

By Lemma 2.1, $\mu y + \nu z \in S_f^\infty \cap (-S_f^\infty)$, which contradicts to $S_f^\infty \cap (-S_f^\infty) = \{0\}$. ∎

THEOREM 3.2. *Let f be a convex polynomial as in Theorem* 3.1.

(i) *If $S_f^\infty \ne \{0\}$, then there exists an orthogonal matrix U such that*

$$f(Ux) = g(x_1, \dots, x_{n-1}) + r x_n \quad \forall x \in \mathbb{R}^n,$$

where g is a convex polynomial satisfying $S_g^\infty = \{0\}$ and $r < 0$.

(ii) *If $S_f^\infty = \{0\}$, then f is strictly convex, and so* $\operatorname{argmin}_{x \in \mathbb{R}^n} f(x)$ *is a singleton set.*

Proof. (i). Since $S_f^\infty \ne \{0\}$, by Theorem 3.1, there exists $w \in \mathbb{R}^n$, with $\|w\| = 1$ such that

$$\cap_{l=2}^m \operatorname{Ker}(f_l) = S_f^\infty - S_f^\infty = \{tw \ :\ t \in \mathbb{R}\}.$$

Since S_f^∞ is a cone, we can assume that $S_f^\infty = \{tw \,:\, t \geq 0\}$ without loss of generality. By Lemma 2.4, there exists $r \leq 0$ such that $f(x+tw) = f(x) + rt$. If $r = 0$, it is easy to prove $w \in S_f^\infty \cap (-S_f^\infty)$, which is a contradiction. Thus, $r < 0$. Note that

$$f(0) + rt = f(tw) = \sum_{l=0}^{m} \sum_{|\beta|=0}^{l} \frac{D^\beta f_l(0)}{\beta!} w^\beta t^{|\beta|} \quad \forall t \in \mathbb{R}.$$

By Lemma 2.2, one has that $r = \sum_{|\beta|=1} \frac{D^\beta f_1(0)}{\beta!} w^\beta = (\nabla f_1)^T w$. Let $F(x) = f(x) - (\nabla f_1)^T x \; \forall x \in \mathbb{R}^n$. Then F is a convex polynomial. We also have

$$\begin{aligned} F(x+tw) &= f(x+tw) - (\nabla f_1)^T (x+tw) \\ &= f(x) + rt - (\nabla f_1)^T x - t(\nabla f_1)^T w \\ &= F(x) \quad \forall x \in \mathbb{R}^n, \forall t \in \mathbb{R}, \end{aligned}$$

that is, $w \in S_F^\infty \cap (-S_F^\infty)$. Note that for any vector v satisfying $\pm v \notin S_f^\infty$, F is not linear on the line $\{tv : t \in \mathbb{R}\}$. Hence, $S_F^\infty \cap (-S_F^\infty) = \{tw : t \in \mathbb{R}\}$. Let U be an orthogonal matrix such that the last column of U is w. By Corollary 2.2, $F(Ux) = \phi(x_1, \ldots, x_{n-1})$, where ϕ is a convex polynomial. Note that the last component of vector $U^T(\nabla f_1 - rw)$ is zero and $U^T w = e(n)$, where $e(n) = (0, \ldots, 0, 1)$. Thus,

$$\begin{aligned} f(Ux) &= F(Ux) + (\nabla f_1)^T Ux \\ &= \phi(x_1, \ldots, x_{n-1}) + \left[U^T(\nabla f_1 - rw)\right]^T x + rw^T Ux \\ &= g(x_1, \ldots, x_{n-1}) + rx_n \quad \forall x \in \mathbb{R}^n, \end{aligned}$$

for some convex polynomial g. It is easy to see that $S_g^\infty = \{0\}$.

(ii). If f is not strictly convex, there exists $x, y \in \mathbb{R}^n$ $(x \neq y)$ such that f is linear on the segment $S = \{z : z = tx + (1-t)y, 0 \leq t \leq 1\}$. Then, it is easy to prove that f is linear on the line $\{z : z = tx + (1-t)y, t \in \mathbb{R}\}$, and so $y - x \in S_f^\infty$ (or $x - y \in S_f^\infty$) according to Lemma 2.4, which contradicts $S_f^\infty = \{0\}$. □

COROLLARY 3.1. *Let h be a kth-order homogeneous convex polynomial on $\mathbb{R}^n$. Then ∇h is a one-to-one mapping from $\mathbb{R}^n$ onto itself.*

Proof. Note that $S_h = \{0\}$, and so $S_h^\infty = \{0\}$. By Theorem 3.2 (ii), h is strictly convex. It is easy to see that $\lim_{\lambda \to \infty} \frac{f(\lambda y)}{\lambda} = +\infty$ for every $y \neq 0$. Then the claim follows from [16, Theorem 26.6]. □

For a homogeneous convex polynomial h, by Lemma 2.6, it is easy to see that $\mathrm{Ker}(h) = S_h = S_h^\infty$. In what follows, we study the relationship between $\mathrm{Ker}(h)$ and the range of ∇h. Note that we do not assume $S_h^\infty \cap (-S_h^\infty) = \{0\}$.

COROLLARY 3.2. *Let h be a kth-order homogeneous convex polynomial. Then we have $\mathrm{Ker}(h)^\perp = R(\nabla h)$, where $R(\nabla h) = \{z \in \mathbb{R}^n : z = \nabla h(x)$ for some $x \in \mathbb{R}^n\}$.*

Proof. The case $k = 1$ is trivial. Suppose $k \geq 2$. Let L be a subspace such that $L^\perp = \mathrm{Ker}(h)$. We will prove $R(\nabla h) = L$. Suppose $\dim(L) = p$. Let $U = (\alpha(1), \ldots, \alpha(n))$, where $[\alpha(1), \ldots, \alpha(p)]$ is an orthonormal basis for L and $[\alpha(p+1), \ldots, \alpha(n)]$ is an orthonormal basis for $L^\perp$. We define $\zeta(x) = h(Ux) \; \forall x \in \mathbb{R}^n$. By Lemma 2.6, we have $\mathrm{Ker}(\zeta) = \{x : \zeta(x) = 0\}$ and $\mathrm{Ker}(h) = \{y : h(y) = 0\}$, and so $\mathrm{Ker}(h) = U\,\mathrm{Ker}(\zeta)$. By the construction of U and $L^\perp = \mathrm{Ker}(h)$, we have $\mathrm{Ker}(\zeta) = (0, \mathbb{R}^{n-p})^T$. From $\zeta(x) = h(Ux)$, it follows that $\nabla\zeta(x) = U^T \nabla h(Ux)$, and so $R(\nabla\zeta) = U^T R(\nabla h)$. Hence, to show $R(\nabla h) = L$, it suffices to prove $R(\nabla\zeta) = (\mathbb{R}^p, 0)^T$. Let $\psi : \mathbb{R}^p \to \mathbb{R}$ be defined by $\psi(y) = \zeta(y, 0) \; \forall y \in \mathbb{R}^p$. Then ψ is a convex

polynomial and $\mathrm{Ker}(\psi) = 0$, which implies $S_\psi = \{0\}$. By Corollary 3.1, we have $R(\nabla\psi) = \mathbb{R}^p$, and so $R(\nabla\zeta) = (\mathbb{R}^p, 0)^T$. □

The condition that h is convex cannot be removed as the next example shows.

Example 3.1. Let $h : \mathbb{R}^2 \to \mathbb{R}$ be defined by $h(x_1, x_2) = (x_1^2 - x_2^2)^2$. It is easy to verify that $\mathrm{Ker}(h) = \{0\}$. Note that

$$\frac{1}{4}\nabla h(x) = \left(\left(x_1^2 - x_2^2\right) x_1, -\left(x_1^2 - x_2^2\right) x_2\right)^T \quad \forall x \in \mathbb{R}^2.$$

Next we show that $R(\nabla h) \neq \mathbb{R}^2$. If $(x_1^2 - x_2^2)x_1 = -(x_1^2 - x_2^2)x_2$, then $(x_1 - x_2)(x_1 + x_2)^2 = 0$. We obtain that $x_1 = x_2$ or $x_1 = -x_2$, and so $\nabla h(x) = (0, 0)^T$. Hence, $(y_1, y_2)^T \notin R(\nabla h)$ if $y_1 = y_2 \neq 0$.

4. Error bounds for convex polynomials. In this section, for each convex polynomial f, we always assume

$$S_f \neq \emptyset \quad \text{and} \quad S_f^\infty \cap (-S_f^\infty) = \{0\}. \tag{4.1}$$

Then, by Theorem 3.1, S_f^∞ is a ray or equal to $\{0\}$. The next theorem shows that if $S_f^\infty \neq \{0\}$, then f has a linear error bound.

THEOREM 4.1. *Let f be a convex polynomial satisfying $S_f^\infty \neq \{0\}$. Then f is unbounded below, and there exists $\tau > 0$ such that*

$$d(x, S_f) \leq \tau[f(x)]_+ \quad \forall x \in \mathbb{R}^n. \tag{4.2}$$

Proof. By Theorem 3.2(i), there exists an orthogonal matrix U such that $f(Ux) = g(x_1, \ldots, x_{n-1}) + rx_n$ for some $r < 0$. It is obvious that f is unbounded below. Let $\bar{f}(x) = f(Ux)\ \forall x \in \mathbb{R}^n$. Then $US_{\bar{f}} = S_f$. Since Euclidean distance is invariant under orthogonal transformations, we have $d(Ux, S_f) = d(x, S_{\bar{f}})$. Thus, it is equivalent to prove that (4.2) holds for $\bar{f}$. Let $x' = x + te(n)$, where $e(n) = (0, \ldots, 0, 1)$. Since $\frac{\partial \bar{f}}{\partial x_n} = r$, we have that $t \leq -\frac{r}{2}(\bar{f}(x) - \bar{f}(x'))$ if $t > 0$ is sufficiently small. Then (4.2) follows from [13, Lemma 2.3]. □

To deal with the case $S_f^\infty = \{0\}$, we need to study convex polynomials which have no linear term.

THEOREM 4.2. *Let $g = \sum_{l=k}^m g_l$ be a convex polynomial, where $k \geq 2$. Then g_k is a convex polynomial.*

Proof. If g_k is not convex, by [3, Theorem 3.3.7], $\nabla^2 g_k(\bar{x})$ is not positive semidefinite for some $\bar{x} \in \mathbb{R}^n$ (obviously $\bar{x} \neq 0$). Then there exists $v \in \mathbb{R}^n$, which satisfies $v^T \nabla^2 g_k(\bar{x}) v < 0$. Hence, we have

$$v^T \nabla^2 g(t\bar{x}) v = \sum_{l=k}^m v^T \nabla^2 g_l(t\bar{x}) v = \sum_{l=k}^m t^{l-2} v^T \nabla^2 g_l(\bar{x}) v \quad \forall t \in \mathbb{R}.$$

If $t > 0$ is small enough, from $v^T \nabla^2 g_k(\bar{x}) v < 0$, it follows that $v^T \nabla^2 g(t\bar{x}) v < 0$, which shows that $\nabla^2 g(t\bar{x})$ is not positive semidefinite. By [3, Theorem 3.3.7], g is not convex, which is a contradiction. Thus, g_k must be convex. □

Stimulated by the idea of [1, p. 46], in the following we will define subspaces $X_k,\ k = 0, \ldots, \frac{m}{2}$, which satisfy

$$X_{k+1} \subseteq X_k \quad \text{and} \quad X_{\frac{m}{2}} = \{0\}. \tag{4.3}$$

For a convex polynomial $f = \sum_{l=2}^m f_l$, let

$$X_0 = \mathbb{R}^n \quad \text{and} \quad X_k = \cap_{i=1}^k \mathrm{Ker}(f_{2i}) \quad \text{for} \quad k = 1, \ldots, \frac{m}{2}. \tag{4.4}$$

LEMMA 4.1. *For each* $1 \le k \le \frac{m}{2}$*, the following assertions hold:*

(i) *If* $2 \le i \le 2k+1$*, then* $X_k \subseteq \mathrm{Ker}(f_i)$.

(ii) $f_{2k}(x) > 0$ *for every* $x \in X_{k-1} \backslash X_k$.

Proof. We prove (i) by induction. For $k = 1$, we need only to prove $X_1 \subseteq \mathrm{Ker}(f_3)$. Let $\bar{f} : X_1 \to \mathbb{R}$ be the restriction of f on X_1. Then $\bar{f}$ is convex, and $\bar{f}(x) = \sum_{i=3}^{m} f_i(x)\ \forall x \in X_1$. By Theorem 4.2, f_3 is a convex function on X_1, which implies that $f_3(x) = 0\ \forall x \in X_1$. From Lemma 2.6, it follows that $X_1 \subseteq \mathrm{Ker}(f_3)$.

Assume that (i) holds for $k = l-1$, where $2 \le l \le \frac{m}{2}$. Since $X_l \subseteq X_{l-1}$, we have $X_l \subseteq \mathrm{Ker}(f_i)$ for each $2 \le i \le 2l-1$, and $X_l \subseteq \mathrm{Ker}(f_{2l})$ by the definition of X_l. Let $\bar{f} : X_l \to \mathbb{R}$ be the restriction of f on X_l. Then $\bar{f}$ is convex, and $\bar{f}(x) = \sum_{i=2l+1}^{m} f_i(x)\ \forall x \in X_l$. Similar to the proof of $k = 1$, we can prove $X_l \subseteq \mathrm{Ker}(f_{2l+1})$. Then (i) holds for $k = l$.

(ii). Let $g : X_{k-1} \to \mathbb{R}$ be the restriction of f on X_{k-1}. Then g is convex and by (i), $g(x) = \sum_{i=2k}^{m} f_i(x)\ \forall x \in X_{k-1}$. From Theorem 4.2, it follows that f_{2k} is convex on X_{k-1}. If $f_{2k}(x) = 0$ for some $x \in X_{k-1}$, then $x \in \mathrm{Ker}(f_{2k})$ by Lemma 2.6, and so $x \in X_k$. Thus, f_{2k} is positive on $X_{k-1} \backslash X_k$. □

If $d \in X_{\frac{m}{2}}$, then by Lemma 4.1(i), $f(td) = \sum_{l=2}^{m} f_l(td) = 0\ \forall t \in \mathbb{R}$. From Lemma 2.3, it follows that $f(y + td) = f(y),\ \forall y \in \mathbb{R}^n,\ \forall t \in \mathbb{R}$, and so $d \in S_f^\infty \cap (-S_f^\infty)$. By (4.1), we obtain $d = 0$, that is, $X_{\frac{m}{2}} = \{0\}$.

LEMMA 4.2. *Let* $f = \sum_{l=2}^{m} f_l$ *be a convex polynomial. Then* $\mathrm{argmin}_{x \in \mathbb{R}^n} f(x) = \{0\}$*, and so* $S_f^\infty = \{0\}$.

Proof. First, we prove f is nonnegative. Suppose there exists $\bar{x} \in \mathbb{R}^n$ such that $f(\bar{x}) < 0$. If we show that $\bar{x} \in X_{\frac{m}{2}}$, then $\bar{x} = 0$ by (4.3), which is a contradiction. Note that

$$f(t\bar{x}) \le (1-t)f(0) + tf(\bar{x}) < 0 \tag{4.5}$$

for every $0 < t < 1$. If $f_2(\bar{x}) > 0$, then $f(t\bar{x}) = \sum_{l=3}^{m} f_l(\bar{x})t^l + f_2(\bar{x})t^2$ is positive for $t > 0$ sufficiently small, which contradicts (4.5). Thus, $f_2(\bar{x}) \le 0$. By Theorem 4.2, f_2 is a convex polynomial, and so $f_2(\bar{x}) \ge 0$. Thus $f_2(\bar{x}) = 0$, which implies $\bar{x} \in X_1$. Assume that $\bar{x} \in X_{k-1}$, where $2 \le k \le \frac{m}{2}$. Next we prove $\bar{x} \in X_k$. By Lemma 4.1 (i), we have $f(t\bar{x}) = \sum_{l=2k+1}^{m} f_l(\bar{x})t^l + f_{2k}(\bar{x})t^{2k}$. If $f_{2k}(\bar{x}) > 0$, then $f(t\bar{x})$ is positive for $t > 0$ sufficiently small, which contradicts (4.5). Thus, $f_{2k}(\bar{x}) \le 0$. Note that $f(z) = \sum_{l=2k}^{m} f_l(z)\ \forall z \in X_{k-1}$. By Theorem 4.2, f_{2k} is a convex polynomial on X_{k-1}, and so f_{2k} is positive semidefinite on X_{k-1}. Hence, $f_{2k}(\bar{x}) = 0$, which implies $\bar{x} \in X_{k-1} \cap \mathrm{Ker}(f_{2k}) = X_k$. By the proof above, we obtain that $\bar{x} \in X_{\frac{m}{2}}$ and arrive at a contradiction. Thus, f is nonnegative.

If $f(\bar{y}) = 0$ for some $\bar{y} \ne 0$, then $f(t\bar{y}) = 0,\ \forall t \in [0,1]$, from the convexity of f. Since f is a polynomial, it is easy to see that $f(t\bar{y}) = 0\ \forall t \in \mathbb{R}$. From Lemma 2.3, it follows that $\bar{y} \in S_f^\infty \cap (-S_f^\infty)$, which contradicts (4.1).

The claim $S_f^\infty = \{0\}$ follows from the boundness of the set $S_f = \{0\}$. □

The following result shows that a convex polynomial $f = \sum_{l=2}^{m} f_l$ has an error bound of order $\frac{1}{m}$ locally.

THEOREM 4.3. *For a convex polynomial* $f = \sum_{l=2}^{m} f_l$*, there exists* $\kappa > 0$ *such that*

$$\|x\| \le \kappa f(x)^{\frac{1}{m}} \quad \text{if} \quad \|x\| < 1. \tag{4.6}$$

Proof. We prove the claim via a contrapositive argument. If (4.6) does not hold, there exists a sequence $\{x(k)\} \subset \mathbb{R}^n$ satisfying $\|x(k)\| < 1$ and

$$\|x(k)\| > kf(x(k))^{\frac{1}{m}} \qquad \forall\ k \ge 1. \tag{4.7}$$

Since $\{x(k)\}$ is bounded, there exists a subsequence $x(n_k)$ such that $x(n_k) \to y$. By (4.7), we have $\|y\| > kf(y)^{\frac{1}{m}}\ \forall k \geq 1$. Then $f(y) = 0$, and so $y = 0$ according to Lemma 4.2. Let $d(k) = \frac{x(n_k)}{\|x(n_k)\|}$ for each $k \geq 1$. Then $\{d(k)\}$ has a cluster d, with $\|d\| = 1$. Without loss of generality, assume that $d = \lim_{k\to\infty} d(k)$.

Now we prove that $d \in X_k$ for each $k \geq 1$. For each $k \geq 1$, let

$$\sigma_k = k^m \frac{f(\|x(n_k)\|d(k))}{\|x(n_k)\|^m} = k^m \sum_{l=2}^{m} f_l(d(k))\|x(n_k)\|^{i-m}.$$

Then by (4.7), $\sigma_k < 1\ \forall k \geq 1$. If $d \notin X_1$, by Lemma 4.1(ii), we have $f_2(d) > 0$. Hence, $f_2(d(k)) > 0$ for all large enough k. Since $\|d(k)\| = 1$, there exists $M > 0$ such that $|f_i(d(k))| < M\ \forall k \geq 1$ and $2 \leq i \leq m$. Then

$$\sigma_k \geq \frac{k^m}{\|x(n_k)\|^{m-3}} \left(\frac{f_2(d(k))}{\|x(n_k)\|} - M \sum_{i=3}^{m} \|x(n_k)\|^{i-3} \right). \tag{4.8}$$

Since $x(n_k) \to 0$, the right-hand side of (4.8) is unbounded as $k \to \infty$, which is a contradiction. Thus, $d \in X_1$. Assume that $d \in X_{k-1}$, where $2 \leq k \leq \frac{m}{2}$. By Lemma 4.1(i), we have

$$\sigma_k = k^m \sum_{l=2k}^{m} f_l(d(k))\|x(n_k)\|^{l-m}.$$

If $f_{2k}(d) > 0$, similar to the proof above, we can derive a contradiction. Then $d \in X_k$. By induction, we have $d \in X_{\frac{m}{2}}$, and so $d = 0$ by (4.3), which is contradiction. The proof is complete. □

THEOREM 4.4. *Let $f = \sum_{l=0}^{m} f_l$ be a convex polynomial satisfying $S_f^\infty = \{0\}$.*

(i) *If there exists $\bar{x} \in \mathbb{R}^n$ such that $f(\bar{x}) < 0$, then f has a linear error bound.*

(ii) *If f is nonnegative on $\mathbb{R}^n$, then there exists a unique $z \in \mathbb{R}^n$ such that $f(z) = 0$. In particular, there exists $\tau > 0$ such that*

$$\|x - z\| \leq \tau \left(f(x) + f(x)^{\frac{1}{m}} \right) \quad \forall x \in \mathbb{R}^n. \tag{4.9}$$

Proof. (i). Note that $\mathrm{bd}(S_f) = \{x \in \mathbb{R}^n : f(x) = 0\}$. Then $\mathrm{bd}(S_f)$ contains none of the optimal solution set. By the fact that f is a convex function, we have $\|\nabla f(x)\| > 0\ \forall x \in bd(S_f)$. Since $S_f^\infty = \{0\}$, S_f is bounded by [16, Theorem 8.4]. Then there exists $\sigma > 0$ such that $\|\nabla f(x)\| \geq \sigma$ for each $x \in \mathrm{bd}(S_f)$. By [13, Theorem 3.1(iii)], f has a linear error bound.

(ii). Note that $\min_{x\in\mathbb{R}^n} f(x) = 0$. By Theorem 3.2(ii), there exists a unique $z \in \mathbb{R}^n$ such that $f(z) = 0$. Let $g(x) = f(x + z)$. Then $S_g^\infty = \{0\}$. Assume that $g = \sum_{l=0}^{m} g_l$. Since $g(0) = 0$, we have $g_0 = 0$. From $g(0) = \min_x g(x)$, it follows that $\nabla g(0) = g_1 = 0$. Hence, $g = \sum_{l=2}^{m} g_l$, and so (4.6) holds for g.

To prove (4.9), it is equivalent to show that there exists $\tau > 0$ such that

$$\|x\| \leq \tau \left(g(x) + g(x)^{\frac{1}{m}} \right) \quad \forall x \in \mathbb{R}^n. \tag{4.10}$$

For $\delta > 0$, let $C_\delta = \{x : g(x) \leq \delta\}$. Select $\delta > 0$ such that $C_\delta \subset \{x : \|x\| \leq \frac{1}{3}\}$. Let $\bar{g}(x) = g(x) - \delta\ \forall x \in \mathbb{R}^n$. Then $S_{\bar{g}} = C_\delta$. By Theorem 4.1, there exists $\tau_1 > 0$ such that

$$d(x, C_\delta) \leq \tau_1 [\bar{g}(x)]_+ \leq \tau_1 g(x) \quad \forall x \in \mathbb{R}^n. \tag{4.11}$$

If $\|x\| > \frac{2}{3}$, $\|x\| \le d(x, C_\delta) + \frac{1}{3} \le 2d(x, C_\delta)$. Thus, from (4.11), it follows that

$$\|x\| \le 2d(x, C_\delta) \le 2\tau_1 g(x) \quad \forall x \text{ satisfying } \|x\| > \frac{2}{3}. \tag{4.12}$$

Combining (4.6) and (4.12), it is easy to see that (4.10) holds for $\tau = \max\{\kappa, 2\tau_1\}$. □

Remark 4.1. By Theorems 4.1 and 4.4(i), for any convex polynomial f, if f satisfies the Slater condition (i.e., there exists $\bar{x}$ such that $f(\bar{x}) < 0$), then f has a linear error bound.

For a convex polynomial $f = \sum_{l=2}^{m} f_l$ $(m \ge 2)$, let

$$Z_0 = \mathbb{R}^n \quad \text{and} \quad Z_i = \cap_{l=1}^{i} \operatorname{Ker}(f_{m-2l+2}) \quad \text{for} \ \ i = 1, \ldots, \frac{m}{2}.$$

LEMMA 4.3. *For each $0 \le i \le \frac{m}{2} - 1$, let $s_i(x) = \sum_{l=m-2i}^{m} f_l(x)$ $\forall x \in \mathbb{R}^n$. Then for each $1 \le i \le \frac{m}{2} - 1$, the following assertions hold:*

(i) *For $m - 2i + 1 \le l \le m$, we have $Z_i \subset \operatorname{Ker}(f_l)$.*

(ii) *There exist $\tau > 0$ and $\delta > 0$ such that*

$$s_i(x) \ge \tau \|x\|^{m-2i} \quad \forall x \in (Z_{i+1})^\perp \cap V_\delta, \tag{4.13}$$

where $V_\delta := \{x \in \mathbb{R}^n : \|x\| \ge \delta\}$.

Proof. We prove the assertions by induction. Let $i = 1$. For (i), by the definition of Z_1, it suffices to prove $Z_1 \subset \operatorname{Ker}(f_{m-1})$. Let $\bar{f} : Z_1 \to \mathbb{R}$ be the restriction of f on Z_1. Then $\bar{f}(x) = \sum_{l=2}^{m-1} f_l(x)$ $\forall x \in Z_1$, and $\bar{f}$ is a convex polynomial on Z_1. By [1, p. 40 Lemma 1], f_{m-1} is a convex polynomial on Z_1. Since $m - 1$ is odd, f_{m-1} must vanish on Z_1. From Lemma 2.6, it follows that $Z_1 \subset \operatorname{Ker}(f_{m-1})$.

(ii). We use a contrapositive argument. If (4.13) does not hold, then there exists a sequence $\{x(k)\} \subset (Z_2)^\perp$ satisfying $\|x(k)\| \ge k$ such that

$$s_1(x(k)) = f_m(x(k)) + f_{m-1}(x(k)) + f_{m-2}(x(k)) < \frac{1}{k}\|x(k)\|^{m-2}. \tag{4.14}$$

Let $d(k) = \frac{x(k)}{\|x(k)\|} \in (Z_2)^\perp$ for each $k \ge 1$. Then $\{d(k)\}$ has a cluster $d \in (Z_2)^\perp$ satisfying $\|d\| = 1$. Without loss of generality, assume that $d = \lim_{k\to\infty} d(k)$. Dividing (4.14) by $\|x(k)\|^m$ and letting $k \to \infty$, we obtain $f_m(d) \le 0$. By [1, p. 40 Lemma 1], f_m is a convex polynomial, which together with Corollary 2.1 implies that $f_m(d) = 0$, that is, $d \in Z_1$. Let $\bar{f} : Z_1 \to \mathbb{R}$ be the restriction of f on Z_1. Then $\bar{f}$ is a convex polynomial on Z_1, and $\bar{f}(x) = \sum_{l=2}^{m-2} f_l(x)$ $\forall x \in Z_1$. By [1, p. 40 Lemma 1], f_{m-2} is a convex polynomial on Z_1. According to Corollary 2.1, f_{m-2} is nonnegative on Z_1. If we prove $f_{m-2}(d) \le 0$, then $f_{m-2}(d) = 0$, and so $d \in \operatorname{Ker}(f_{m-2})$ by Lemma 2.6. Hence, $d \in Z_1 \cap \operatorname{Ker}(f_{m-2}) = Z_2$. We obtain a contradiction, and so (ii) holds for $i = 1$.

Now we prove $f_{m-2}(d) \le 0$. For $x \in \mathbb{R}^n$, we use x_u (x_v) to denote the projection of x on Z_1 $((Z_1)^\perp)$. By $Z_1 = \operatorname{Ker}(f_m) \subset \operatorname{Ker}(f_{m-1})$ and Lemma 2.5,

$$f_m(x(k)) + f_{m-1}(x(k)) = f_m(x_v(k)) + f_{m-1}(x_v(k)). \tag{4.15}$$

Since f_m is a convex polynomial, by Lemma 2.6, f_m is positive on $(Z_1)^\perp$. Then there exists $\rho > 0$ such that $f_m(x) + f_{m-1}(x) > 0$ $\forall x \in (Z_1)^\perp \cap V_\rho$. If $x_v(k) \in V_\rho$ for infinitely k, then the left-hand side of (4.15) is nonnegative for such k, which

together with (4.14) implies that $f_{m-2}(x(k)) < \frac{1}{k}\|x(k)\|^{m-2}$, that is, $f_{m-2}(d(k)) < \frac{1}{k}$. Letting $k \to \infty$, we obtain $f_{m-2}(d) \le 0$. If $\|x_v(k)\| \le \rho$ for k sufficiently large, then $f_m(x(k)) + f_{m-1}(x(k))$ is bounded. Dividing (4.14) by $\|x(k)\|^{m-2}$ and letting $k \to \infty$, we obtain $f_{m-2}(d) \le 0$ also.

Assume the assertions hold for $i = j-1$, where $2 \le j \le \frac{m}{2}$. Now we prove (i) for $i = j$. We need only to prove $Z_j \subset \mathrm{Ker}(f_{m-2j+1})$. Let $\bar{f} : Z_j \to \mathbb{R}$ be the restriction of f on Z_j. Then $\bar{f}$ is a convex polynomial on Z_j. By the inductive hypothesis and $Z_j \subset \mathrm{Ker}(f_{m-2j+2})$, we have $\bar{f}(x) = \sum_{l=2}^{m-2j+1} f_l(x)\ \forall x \in Z_j$. By [1, p. 40 Lemma 1], f_{m-2j+1} is a convex polynomial on Z_j. Since $m-2j+1$ is odd, f_{m-2j+1} must vanish on Z_j. From Lemma 2.6, it follows that $Z_j \subset \mathrm{Ker}(f_{m-2j+1})$.

For (ii), if (4.13) does not hold, then there exists a sequence $\{x(k)\} \subset (Z_{j+1})^\perp$ satisfying $\|x(k)\| \ge k$ such that

$$s_j(x(k)) = s_{j-1}(x(k)) + f_{m-2j+1}(x(k)) + f_{m-2j}(x(k)) < \frac{1}{k}\|x(k)\|^{m-2j}. \tag{4.16}$$

Let $d(k) = \frac{x(k)}{\|x(k)\|} \in (Z_{j+1})^\perp$ for each $k \ge 1$. As above, we assume that $d = \lim_{k\to\infty} d(k)$. Then $d \in (Z_{j+1})^\perp$, and $\|d\| = 1$. For $x \in \mathbb{R}^n$, let x_u (x_v) be the projection of x on Z_j ($(Z_j)^\perp$). By $Z_j \subset \mathrm{Ker}(f_{m-2j+1})$ and the definition of Z_j, we have

$$s_{j-1}(x(k)) + f_{m-2j+1}(x(k)) = s_{j-1}(x_v(k)) + f_{m-2j+1}(x_v(k)). \tag{4.17}$$

If $\{x_v(k)\}$ is bounded, then $d = \lim_{k\to\infty} x_u(k)/\|x(k)\| \in Z_j$. If $\{x_v(k)\}$ is unbounded, by the inductive hypothesis, (4.13) holds for $i = j-1$, and so there exists $\tau > 0$ such that

$$s_{j-1}(x_v(k)) \ge \tau\|x_v(k)\|^{m-2j+2} \tag{4.18}$$

for infinitely k. By (4.16), (4.17), and (4.18), we deduce that $\lim_{k\to\infty} \frac{\|x_v(k)\|}{\|x(k)\|} = 0$, which implies $d \in Z_j$ also. Similar to the proof of $i = 1$, if we show that $f_{m-2j}(d) \le 0$, then we obtain $d \in Z_j \cap \mathrm{Ker}(f_{m-2j}) = Z_{j+1}$, which is a contradiction.

Now we prove $f_{m-2j}(d) \le 0$. Since (4.13) holds for $i = j-1$, there exists $\rho > 0$ such that $s_{j-1}(x) + f_{m-2j+1}(x) > 0$ for any $x \in (Z_j)^\perp \cap V_\rho$. If $x_v(k) \in V_\rho$ for infinitely k, by (4.16) and (4.17), we have $f_{m-2j}(x(k)) < \frac{1}{k}\|x(k)\|^{m-2j}$ for such k, that is, $f_{m-2j}(d(k)) < \frac{1}{k}$. Letting $k \to \infty$, we obtain $f_{m-2j}(d) \le 0$. If $\|x_v(k)\| \le \rho$ for k sufficiently large, then $s_{j-1}(x(k)) + f_{m-2j+1}(x(k))$ is bounded. Dividing (4.16) by $\|x(k)\|^{m-2j}$ and letting $k \to \infty$, we obtain $f_{m-2j}(d) \le 0$ also. The proof is complete. □

If $d \in Z_{\frac{m}{2}}$, by Lemma 4.3(i) and the definition of $Z_{\frac{m}{2}}$, $f(td) = \sum_{l=2}^{m} f_l(td) = 0\ \forall t \in \mathbb{R}$. From Lemma 2.3, it follows that $f(y + td) = f(y),\ \forall y \in \mathbb{R}^n,\ t \in \mathbb{R}$, and so $d \in S_f^\infty \cap (-S_f^\infty)$. By (4.1), we obtain $d = 0$, that is, $Z_{\frac{m}{2}} = \{0\}$. By Lemma 4.3(ii) and the fact $Z_{\frac{m}{2}} = \{0\}$, there exist $\tau > 0$ and $\delta > 0$ such that

$$f(x) \ge \tau\|x\|^2 \quad \forall x \in V_\delta. \tag{4.19}$$

Combining (4.19) and Lemma 4.2, it is easy to obtain the following stronger result.

COROLLARY 4.1. *Let $f = \sum_{l=2}^{m} f_l$ be a convex polynomial. For any $\delta > 0$, there exists $\tau > 0$ such that*

$$f(x) \ge \tau\|x\|^2 \quad \forall x \in V_\delta.$$

PROPOSITION 4.1. *Let $f = \sum_{l=2}^{m} f_l$ be a convex polynomial. For any $\delta > 0$, there exists $\tau > 0$ such that $\|\nabla f(x)\| \geq \tau \|x\|$ $\forall x \in V_\delta$, where $V_\delta := \{x \in \mathbb{R}^n : \|x\| \geq \delta\}$.*

Proof. Note that $f(0) = f(x) - x^T \nabla f(x) + \frac{1}{2} x^T \nabla^2 f(\xi) x$ for some $\xi \in \mathbb{R}^n$. Since $\nabla^2 f(\xi)$ is positive semidefinite, we have

$$\|x\| \cdot \|\nabla f(x)\| \geq x^T \nabla f(x) \geq f(x) - f(0) \geq \tau \|x\|^2 \quad \forall x \in V_\delta,$$

where the last inequality follows from Corollary 4.1. Thus, the claim is true. ∎

Using the fact $Z_{\frac{m}{2}} = \{0\}$ and a contrapositive argument, we can establish the following result, which generalizes Corollary 3.1 to the nonhomogeneous convex polynomials.

PROPOSITION 4.2. *Let $f = \sum_{l=2}^{m} f_l$ be a convex polynomial. Then ∇f is a one-to-one mapping from $\mathbb{R}^n$ onto itself.*

Proof. By Lemma 4.2, we have $S_f^\infty = \{0\}$, which together with Theorem 3.2(ii) implies that f is strictly convex. If we prove $\lim_{\lambda \to \infty} \frac{f(\lambda y)}{\lambda} = +\infty$ for every $y \neq 0$, then the claim follows from [16, Theorem 26.6]. Suppose the contrary. There exists $\bar{y} \neq 0$ such that

$$\lim_{\lambda \to \infty} \frac{f(\lambda \bar{y})}{\lambda} = \delta \tag{4.20}$$

for some $\delta \geq 0$. We must have $\bar{y} \in Z_1$. Otherwise, $f_m(\bar{y}) > 0$, and so $\lim_{\lambda \to \infty} \frac{f(\lambda \bar{y})}{\lambda} = +\infty$. Suppose $\bar{y} \in Z_{k-1}$, where $2 \leq k \leq \frac{m}{2}$. By Lemma 4.3(i), $f(\lambda \bar{y}) = \sum_{l=2}^{m-2k+2} f_l(\lambda \bar{y})$. We must have $\bar{y} \in Z_k$. Otherwise, we will obtain a contradiction to (4.20). Thus, we arrive at $\bar{y} \in Z_{\frac{m}{2}}$, and so $\bar{y} = 0$, which is a contradiction. ∎

Remark 4.2. For a convex polynomial $f = \sum_{l=2}^{m} f_l$, if $S_f^\infty \cap (-S_f^\infty) = \{0\}$ does not hold, according to Lemma 4.2, S_f^∞ must be a linear space. Thus, by Theorem 3.1, one has that $S_f^\infty = \cap_{l=2}^{m} \mathrm{Ker}(f_l)$. Moreover, using a similar argument of Corollary 3.2, it is easy to prove $(S_f^\infty)^\perp = R(\nabla f)$.

5. Error bounds for polyhedral-constrained convex polynomials. In this section, we consider the error bounds of the following problem:

$$\begin{aligned} &\min f(x) \\ &\text{subject to } x \in P, \end{aligned} \tag{5.1}$$

where f is an mth-order convex polynomial and P is a polyhedral set. Let $P = \{x \in \mathbb{R}^n : (a^i)^T x \leq b^i,\ 1 \leq i \leq k\}$, where $a^i \in \mathbb{R}^n$, $b^i \in \mathbb{R}$ for each $i \in [k]$. Let $S = S_f \cap P \neq \emptyset$. By Lemma 2.4, if $d \in S^\infty \cap (-S^\infty)$, then f is constant along the direction $\pm d$ and $P + td = P$ $\forall t \in \mathbb{R}$. Hence, without loss of generality, we assume that

$$S^\infty \cap (-S^\infty) = \{0\}.$$

First, we will derive a linear error bound under the Slater condition.

$$\text{There exists } x^* \in S \text{ such that } f(x^*) < 0. \tag{5.2}$$

For each $x \in \mathrm{bd}(S)$, we define the index set of x by $I(x) := \{i \in [k] : (a^i)^T x = b^i\}$.

THEOREM 5.1. *If there exists $x^* \in S$ such that $f(x^*) < 0$, then there exists $\tau > 0$ such that*

$$d(x, S) \leq \tau [f(x)]_+ \quad \forall x \in P. \tag{5.3}$$

Proof. Let $F(x) = f(x) + \iota_P(x)$, $\forall x \in \mathbb{R}^n$, where $\iota_P(x) = 0$ $\forall x \in P$ and $\iota_P(x) = +\infty$ $\forall x \notin P$. Then $\partial F(x) = \nabla f(x) + N_P(x)$ for each $x \in P$. Fix $y \in \mathrm{bd}(S)$, which satisfies $f(y) = 0$. By (5.2) and [2, Corollary 3],

$$N_S(y) = N_{S_f}(y) + N_P(y) = \left\{ \lambda \nabla f(y) + \sum_{i \in I(y)} \lambda_i a^i : \lambda \geq 0, \lambda_i \geq 0 \ \forall i \in I(y) \right\}.$$

Hence, $N_S(y) = \{\theta \partial F(y) : \theta \geq 0\}$. If $h \in N_S^1(y)$, then $h = \theta(\nabla f(y) + \sum_{i \in I(y)} \mu_i a^i)$ for some $\theta > 0$ and $\mu_i \geq 0$ $\forall i \in I(y)$. By (2.1), we have

$$\begin{aligned} F'(y;h) &= \max\left\{h^T \xi : \xi \in \partial F(y)\right\} \\ &\geq \theta \left\| \nabla f(y) + \sum_{i \in I(y)} \mu_i a^i \right\|^2 = \left\| \nabla f(y) + \sum_{i \in I(y)} \mu_i a^i \right\|. \end{aligned} \tag{5.4}$$

Now we prove that for each $x \in \mathrm{bd}(S)$ satisfying $f(x) = 0$, the following holds:

$$\left\| \nabla f(x) + \sum_{i \in I(x)} \lambda_i a^i \right\| \geq \frac{1}{\tau} \quad \forall \lambda_i \geq 0, \ i \in I(x). \tag{5.5}$$

Then (5.3) follows from [13, Theorem 3.1(iii)], (5.4), and (5.5). Suppose the contrary. There exist sequences of $\{x^r\}$ and $\{\lambda^r\}$ such that $x^r \in \mathrm{bd}(S)$, $f(x^r) = 0$, $\lambda^r \geq 0$, and

$$\nabla f(x^r) + \sum_{i \in I(x^r)} \lambda_i^r a^i \to 0. \tag{5.6}$$

Since $I(x^r) \subset [k]$ and $[k]$ has finite number of subsets, there is a subsequence $\{y^r\}$ of $\{x^r\}$ such that $I(y^r) = I$ for some index set $I \subset [k]$. Without loss of generality, we may assume that $I(x^r) = I$. Let $L = \{x \in \mathbb{R}^n : (a^i)^T x = b^i \ \forall i \in I\}$. Then $\{x^r\} \subset L$.

Now we show that $\{x^r\}$ is unbounded. Otherwise, $\{x^r\}$ has a cluster, say x^∞. Then $f(x^\infty) = 0$ and $(a^i)^T x^\infty = b^i$ $\forall i \in I$. From (5.6), it follows that $\| \sum_{i \in I(x^r)} \lambda_i^r a^i \|$ is bounded. By Hoffman's error bound we may assume, without loss of generality, that $\{\lambda_i^r\}$, $i \in I$ is bounded. Therefore, by passing to a subsequence if necessary, we can assume that the sequence $\{\lambda_i^r\}$ converges to $\lambda_i^\infty \geq 0$ for each $i \in I$. Then (cf. (5.6))

$$\nabla f(x^\infty) + \sum_{i \in I} \lambda_i^\infty a^i = 0. \tag{5.7}$$

Let $\varphi(x) = f(x) + \sum_{i \in I} \lambda_i^\infty ((a^i)^T x - b^i)$ $\forall x \in \mathbb{R}^n$. Then $\varphi(x^\infty) = 0$ and $\nabla \varphi(x^\infty) = 0$. Since φ is convex, φ attains its global minimum, which is 0, at x^∞. However, by (5.2) and the fact $x^* \in P$, we have $\varphi(x^*) < 0$, a contradiction. Hence, $\{x^r\}$ is unbounded.

If S_f^∞ is a linear space, without loss of generality, we assume that $S_f^\infty = (0, \mathbb{R}^{n-p})^T$ for some $p < n$. For each $x \in \mathbb{R}^n$, let x_u denote the vector of the first p components of x, that is, $x_u = (x_1, \ldots, x_p)$. Then $f(x) = g(x_u)$, where g is a convex polynomial on $\mathbb{R}^p$ satisfying $S_g^\infty = \{0\}$. Thus, S_g is bounded, which implies $\{x_u^r\}$ is bounded. Therefore, $\{x_u^r\}$ has a cluster $z \in \mathbb{R}^p$. Let $x^\infty = (z^T, 0)^T \in \mathbb{R}^n$. Then $f(x^\infty) = 0$. Similar to the proof above, we can assume that the sequence $\{\lambda_i^r\}$ converges to $\lambda_i^\infty \geq 0$ for each $i \in I$. Then (5.7) holds for x^∞ and λ^∞. By the same argument as above, we can derive a contradiction.

Thus, S_f^∞ is not a space. By Theorem 3.1 and Theorem 3.2(i), without loss of generality, we assume that $S_f^\infty = \{x \in \mathbb{R}^n : x_{p+1} \geq 0, x_i = 0 \text{ for } i > p+1\}$ for some $p \leq n-1$ and

$$f(x) = g(x_1, \ldots, x_p) + cx_{p+1},$$

where g is a convex polynomial on $\mathbb{R}^p$ satisfying $S_g^\infty = \{0\}$ and $c < 0$. By Theorem 3.2(ii), g is strictly convex. It is not hard to see that there exist r_1 and r_2 such that $f(\frac{x^{r_1}+x^{r_2}}{2}) < 0$. Let $\bar{f}$ be the restriction of f on L. Then $\bar{f}$ satisfies the Slater condition. By coordinate transformation, if necessary, and using Remark 4.1, $\bar{f}$ has a linear error bound.

Fix $q \in \{x \in L : \bar{f}(x) = 0\}$. Let w be the projection of $\nabla f(q)$ onto L^∞. Then $w \neq 0$. Otherwise, $\bar{f}$ will attain its minimum at q. Note that $\|w\| = \min\{\|\nabla f(q)+\xi\| : \xi \in (L^\infty)^\perp\}$. We also have $N^1_{S_{\bar{f}}}(q) \cap L^\infty = \{w/\|w\|\}$. Since $\bar{f}$ has a linear error bound, by [13, Theorem 3.1(iii)], there exists $\kappa > 0$ such that

$$\kappa < \bar{f}\left(q; \frac{w}{\|w\|}\right) = \frac{w^T \nabla f(q)}{\|w\|} = \|w\| = \min\left\{\|\nabla f(q) + \xi\| : \xi \in (L^\infty)^\perp\right\},$$

which contradicts (5.6) since $(L^\infty)^\perp = \{\sum_{i\in I} \lambda_i a^i : \lambda_i \in \mathbb{R}\}$. The proof is complete. □

THEOREM 5.2. *If f is nonnegative on P, then there exists $\tau > 0$ such that*

$$d(x, S) \leq \tau\left(f(x) + f(x)^{\frac{1}{m}}\right) \quad \forall x \in P. \tag{5.8}$$

Proof. It is easy to see that S is a singleton set. Assume that $S = \{x^*\}$. By KKT condition of problem (5.1), there exist $\lambda_i \geq 0,\ \forall i \in I(x^*)$, such that

$$\nabla f(x^*) + \sum_{i\in I(x^*)} \lambda_i a^i = 0.$$

Let $\varphi(x) = f(x) + \sum_{i\in I(x^*)} \lambda_i((a^i)^T x - b^i)\ \forall x \in \mathbb{R}^n$. Then φ is a convex polynomial and $\nabla\varphi(x^*) = 0$, and so φ attains its minimum at x^*. By Theorem 4.4(ii), (5.8) holds thanks to $\varphi(x) \leq f(x)\ \forall x \in P$. □

Acknowledgments. The author is grateful to the referees for their valuable comments and suggestions in improving the quality of the manuscript.

REFERENCES

[1] B. BANK AND R. MANDEL, *Parametric Integer Optimization, Mathematical Research,* Vol. 39, Akademie-Verlag, Berlin, 1988.

[2] H. H. BAUSCHKE, J. M. BORWEIN, AND W. LI, *Strong conical hull intersection property, bounded linear regularity, Jameson's property (G), and error bounds in convex optimization,* Math. Program. Ser. A, 86 (1999), pp. 135–160.

[3] M. S. BAZARAA, H. D. SHERALI, AND C. M. SHETTY, *Nonlinear programming,* Theory and Algorithms, 2nd ed., John Wiley & Sons, New York, 1993.

[4] E. G. BELOUSOV AND D. KLATTE, *A Frank-Wolfe type theorem for convex polynomial programs,* Comput. Optim. Appl., 22 (2002), pp. 37–48.

[5] E. G. BELOUSOV, *Introduction to Convex Analysis and Integer Programming,* Moscow University, Moscow, 1977 (in Russian).

[6] M. KOJIMA, S. KIM, AND H. WAKI, *A general framework for convex relaxation of polynomial optimization problems over cones,* J. Oper. Res. Soc. Japan, 46 (2003), pp. 125–144.

[7] J. B. Lasserre, *Global optimization with polynomials and the problem of moments,* SIAM J. Optim., 11 (2001), pp. 796–817.
[8] X. D. Luo and Z. Q. Luo, *Extension of Hoffman's error bound to polynomial systems,* SIAM J. Optim., 4 (1994), pp. 383–392.
[9] Z. Q. Luo and J. S. Pang, *Error bounds for analytic systems and their applications,* Math. Program., 67 (1994), pp. 1–28.
[10] Z. Q. Luo and J. F. Sturm, *Error bounds for quadratic systems,* in High Perpormance Optimization, H. Frenk, K. Roos, T. Terlaky, and S. Zhang, eds., Kluwer, Dordrecht, The Netherlands, 2000, pp. 383–404.
[11] Y. Nesterov, *Squared functional systems and optimization problems,* in High Performance Optimization, H. Frenk, K. Roos, T. Terlaky, and S. Zhang, eds., Kluwer, Dordrecht, The Netherlands, 2000, 1992, pp. 405–440.
[12] K. F. Ng and X. Y. Zheng, *Global error bounds with fractional exponents,* Math. Program. Ser. B, 88 (2000), pp. 357–370.
[13] K. F. Ng and X. Y. Zheng, *Error bounds for lower semicontinuous functions in normed spaces,* SIAM J. Optim., 12 (2001), pp. 1–17.
[14] P. A. Parrilo, *Semidefinite programming relaxation for semialgebraic problems,* Math. Program., 96 (2003), pp. 293–320.
[15] L. Qi, W. Sun, and Y. Wang, *Numerical multilinear algebra and its applications,* Front. Math. China, 2 (2007), pp. 501–526.
[16] R. T. Rockafellar, *Convex Analysis,* Princeton University Press, Princeton, NJ, 1970.
[17] N. Z. Shor, *Class of global minimum bounds of polynomial functions,* Cybernetics, 23 (1987), pp. 731–734.
[18] H. Ngai and M. Thera, *Error bounds for convex differentiable inequality systems in Banach spaces,* Math. Program., 104 (2005), pp. 465–482.
[19] T. Wang and J. S. Pang, *Global error bound for convex quadratic inequality systems,* Optimization, 31 (1994), pp. 1–12.
[20] Z. Wu and J. Ye, *On error bounds for lower semicontinuous functions,* Math. Program., 92 (2002), pp. 301–314.

SIAM J. OPTIM.
Vol. 19, No. 4, pp. 1648–1673

CALMNESS FOR L-SUBSMOOTH MULTIFUNCTIONS IN BANACH SPACES*

XI YIN ZHENG† AND KUNG FU NG‡

Abstract. Using variational analysis techniques, we study subsmooth multifunctions in Banach spaces. In terms of the normal cones and coderivatives, we provide some characterizations for such multifunctions to be calm. Sharper results are obtained for Asplund spaces. We also present some exact formulas of the modulus of the calmness. As applications, we provide some error bound results on nonconvex inequalities, which improve and generalize the existing error bound results.

Key words. subsmoothness, calmness, metric subregularity, error bound, multifunction

AMS subject classifications. 90C31, 90C25, 49J52, 46B20

DOI. 10.1137/080714129

1. Introduction. As an extension of convexity, prox-regularity of a set expresses a variational behavior of "order two" and plays an important role in variational analysis (see [5, 32, 34] and the references therein). Recently, Aussel, Daniilidis, and Thibault [1] considered a variational behavior of "order one" of a set and introduced subsmoothness, extending the notions of the smoothness and the prox-regularity. Motivated by their work, we consider (in section 3) a further weakened notion (called L-subsmooth).

The calmness property plays an important role in many issues in mathematical programming like exact penalty functions, optimality conditions, local error bounds, weak sharp minima, and so on. Recently, many authors studied calmness (cf. [8, 11, 9, 10, 17, 34, 43, 45] and the references therein). Let Y, X be Banach spaces and $M : Y \rightrightarrows X$ a multifunction. For $\bar{y} \in Y$ and $\bar{x} \in M(\bar{y})$, recall that M is calm at $(\bar{y}, \bar{x})$ if there exist $\eta, \delta \in (0, +\infty)$ such that

$$d(x, M(\bar{y})) \le \eta \|y - \bar{y}\| \quad \forall y \in B(\bar{y}, \delta) \text{ and } x \in M(y) \cap B(\bar{x}, \delta), \tag{1.1}$$

where $B(\bar{x}, \delta)$ denotes the open ball with center $\bar{x}$ and radius r. Let $F(x) := \{y \in Y : x \in M(y)\}$ for all $x \in X$. As observed by Henrion and Outrata [10], the calmness of M at $(\bar{y}, \bar{x})$ is equivalent to the condition that there exist $\eta, \delta \in (0, +\infty)$ such that

$$d(x, F^{-1}(\bar{y}) \le \eta d(\bar{y}, F(x)) \quad \forall x \in B(\bar{x}, \delta). \tag{1.2}$$

Following Dontchev and Rockafellar [6], (1.2) means that the generalized equation $\bar{y} \in F(x)$ is metrically subregular at $\bar{x}$. This property provides an estimate on how far a candidate x can be from the solution set of the generalized equation. A stronger property is the following: a multifunction F is said to be metrically regular at $\bar{x}$ for $\bar{y}$

*Received by the editors January 24, 2008; accepted for publication (in revised form) September 26, 2008; published electronically January 21, 2009. This research was supported by an earmarked grant from the Research Grant Council of Hong Kong and by the National Natural Science Foundation of People's Republic of China (grant 10761012).
http://www.siam.org/journals/siopt/19-4/71412.html
†Department of Mathematics, Yunnan University, Kunming 650091, People's Republic of China (xyzheng@ynu.edu.cn).
‡Department of Mathematics, The Chinese University of Hong Kong, Shatin, New Territory, Hong Kong (kfng@math.cuhk.edu.hk).

if there exist $\tau, \delta \in (0, +\infty)$ such that

$$d(x, F^{-1}(y)) \le \tau d(y, F(x)) \quad \forall (x, y) \in B(\bar{x}, \delta) \times B(\bar{y}, \delta)). \tag{1.3}$$

Both notions (of the metric regularity and the metric subregularity) have been studied by many authors (see [7, 6, 14, 20, 23, 24, 25, 28, 40, 43] and the references therein). In particular, it is well known (cf. [23, 24, 25, 34]) that if X and Y are finite-dimensional, then F is metrically regular at $\bar{x}$ for $\bar{y}$ if and only if $D^*F(\bar{x}, \bar{y})^{-1}(0) = \{0\}$; moreover

$$\inf\{\tau > 0 : \text{ (1.3) holds}\} = \|D^*F(\bar{x}, \bar{y})^{-1}\|^- = \limsup_{(x,y) \overset{\mathrm{Gr}(F)}{\longrightarrow} (\bar{x}, \bar{y})} \|D^*F(x, y)^{-1}\|^-, \tag{1.4}$$

where $D^*F(x, y)$ is the coderivative of F at (x, y) and $\|D^*F(x, y)^{-1}\|^-$ denotes the inner norm of $D^*F(x, y)^{-1}$ (see section 2 for undefined terms and further notation). The modulus of the calmness of M at $(\bar{y}, \bar{x})$ is denoted by $\eta(M; \bar{y}, \bar{x})$ and defined by

$$\eta(M; \bar{y}, \bar{x}) := \inf\{\eta \in (0, \infty) : \text{ (1.1) holds for some } \delta \in (0, +\infty)\}. \tag{1.5}$$

The case $\eta(M; \bar{y}, \bar{x}) = \infty$ indicates that M is not calm at $(\bar{y}, \bar{x})$ (here and throughout we adopt the convention that the infimum over the empty set is ∞). In terms of the normal cone of $M(\bar{y})$, the derivative, or subdifferential, Henrion and Outrata [9], Henrion, Jourani, and Outrata [11], and Henrion and Jourani [8] gave sufficient conditions for $\eta(M; \bar{y}, \bar{x}) < +\infty$ in some special cases. Recently, in terms of the normal cone and coderivative, the authors [43] considered the case when M is a general closed convex multifunction between Banach spaces and provided some characterizations for $\eta(M; \bar{y}, \bar{x}) < +\infty$. This and (1.4) motivate us to seek some formulas for $\eta(M; \bar{y}, \bar{x})$ in terms of coderivative in the case when M is not necessarily convex. In section 4, for L-subsmooth multifunctions, we establish some such formulas and provide several sufficient and/or necessary conditions for the calmness. In section 5, as an application, we consider error bounds for inequalities. In particular, we extend some existing error bound results from the convex case to the nonconvex case.

2. Preliminaries. Let X be a Banach space. Let X^* and B_X denote the dual space and the closed unit ball of X, respectively.

For a closed subset A of X and $a \in A$, let $T_c(A, a)$ and $T(A, a)$ denote, respectively, the Clarke tangent cone and the contingent (Bouligand) cone of A at a; they are defined by

$$T_c(A, a) := \liminf_{x \overset{A}{\to} a, t \to 0^+} \frac{A - x}{t} \quad \text{and} \quad T(A, a) := \limsup_{t \to 0^+} \frac{A - a}{t},$$

where $x \overset{A}{\to} a$ means that $x \to a$ with $x \in A$. Thus, $v \in T_c(A, a)$ if and only if, for each sequence $\{a_n\}$ in A converging to a and each sequence $\{t_n\}$ in $(0, \infty)$ decreasing to 0, there exists a sequence $\{v_n\}$ in X converging to v such that $a_n + t_n v_n \in A$ for all n, while $v \in T(A, a)$ if and only if there exist a sequence $\{v_n\}$ converging to v and a sequence $\{t_n\}$ in $(0, \infty)$ decreasing to 0 such that $a + t_n v_n \in A$ for all n. We denote by $N_c(A, a)$ the Clarke normal cone of A at a, that is,

$$N_c(A, a) := \{x^* \in X^* : \langle x^*, h\rangle \le 0 \ \ \forall h \in T_c(A, a)\}.$$

For $\varepsilon \ge 0$ and $a \in A$, the nonempty set

$$\hat{N}_\varepsilon(A, a) := \left\{ x^* \in X^* : \limsup_{x \overset{A}{\to} a} \frac{\langle x^*, x - a\rangle}{\|x - a\|} \le \varepsilon \right\}$$

is called the set of Fréchet ε-normals of A at a. When $\varepsilon = 0$, $\hat{N}_\varepsilon(A,a)$ is a convex cone which is called the Fréchet normal cone of A at a and is denoted by $\hat{N}(A,a)$.

Let $N(A,a)$ denote the Mordukhovich normal cone (also known as the limiting or basic normal cone) of A at a, that is,

$$N(A,a) = \limsup_{x \xrightarrow{A} a, \varepsilon \to 0^+} \hat{N}_\varepsilon(A,x).$$

Thus, $x^* \in N(A,a)$ if and only if there exists a sequence $\{(x_n, \varepsilon_n, x_n^*)\}$ in $A \times R_+ \times X^*$ such that $(x_n, \varepsilon_n) \to (a,0)$, $x_n^* \xrightarrow{w^*} x^*$ and $x_n^* \in \hat{N}_{\varepsilon_n}(A, x_n)$ for each $n \in \mathbb{N}$, where $\mathbb{N}$ denotes the set of all natural numbers. It is known that

$$\hat{N}(A,a) \subset N(A,a) \subset N_c(A,a) \tag{2.1}$$

(cf. [24, 25, 26]). It is known that if A is convex, then $T_c(A,a) = T(A,a)$ and

$$N_c(A,a) = \hat{N}(A,a) = \{x^* \in X^* : \ \langle x^*, x\rangle \le \langle x^*, a\rangle \ \ \forall x \in A\}.$$

Recall that a Banach space X is called an Asplund space if every continuous convex function on X is Fréchet differentiable at each point of a dense subset of X (for other definitions and their equivalents, see [31, Definition 1.22 and Corollary 2.35]). It is well known (cf. [31]) that X is an Asplund space if and only if every separable subspace of X has a separable dual space. In the case when X is an Asplund space, Mordukhovich and Shao [26] proved that

$$N_c(A,a) = \mathrm{cl}^*(\mathrm{co}(N(A,a))) \ \text{ and } \ N(A,a) = \limsup_{x \xrightarrow{A} a} \hat{N}(A,x).$$

The following approximate projection result (recently established in [44]) will play an important role in the proofs of our main results.

LEMMA 2.1. *Let be A a nonempty closed subset of a Banach space X and let $\gamma \in (0,\ 1)$. Then for any $x \notin A$ there exist $a \in \mathrm{bd}(A)$ and $a^* \in N_c(A,a)$ with $\|a^*\| = 1$ such that*

$$\gamma\|x - a\| < \min\{d(x,A),\ \langle a^*, x - a\rangle\}.$$

If X is assumed to be an Asplund space, then above a^* can be chosen from $\hat{N}(A,a)$.

For a multifunction F between Banach spaces X and Y, the graph of F is defined by

$$\mathrm{Gr}(F) := \{(x,y) \in X \times Y : \ y \in F(x)\}.$$

As usual, F is said to be closed (resp., convex) if $\mathrm{Gr}(F)$ is a closed (resp., convex) subset of $X \times Y$. Let $(x,y) \in \mathrm{Gr}(F)$. The Clarke tangent and contingent derivatives $D_cF(x,y)$, $DF(x,y)$ of F at (x,y) are defined by

$$\mathrm{Gr}(D_cF(x,y)) = T_c(\mathrm{Gr}(F),(x,y)) \ \text{ and } \ \mathrm{Gr}(DF(x,y)) = T(\mathrm{Gr}(F),(x,y)),$$

respectively. Let $\hat{D}^*F(x,y)$, $D^*F(x,y)$, and $D_c^*F(x,y)$ denote the coderivatives of F at (x,y) associated, respectively, with the Fréchet, Mordukhovich, and Clarke normal structures; they are defined by

$$\hat{D}^*F(x,y)(y^*) := \{x^* \in X^* : \ (x^*, -y^*) \in \hat{N}(\mathrm{Gr}(F),(x,y))\} \quad \forall y^* \in Y^*,$$
$$D^*F(x,y)(y^*) := \{x^* \in X^* : \ (x^*, -y^*) \in N(\mathrm{Gr}(F),(x,y))\} \quad \forall y^* \in Y^*,$$

and

$$D_c^*F(x,y)(y^*) := \{x^* \in X^* :\ (x^*, -y^*) \in N_c(\mathrm{Gr}(F), (x,y))\} \quad \forall y^* \in Y^*.$$

The history of the coderivatives can be found in Mordukhovich's book [24, 25].

Let $G : X \rightrightarrows Y$ be a positively homogeneous multifunction (i.e., $\mathrm{Gr}(G)$ is a cone in $X \times Y$). Following Dontchev, Lewis, and Rockefeller [7], the inner norm of G is defined by $\|G\|^- := \sup_{x\in B_X} \inf_{y\in Gx} \|y\|$. For a cone K in X, let $\|G|_K\|^-$ be defined by $\|G|_K\|^- := \sup_{x\in B_X\cap K} \inf_{y\in Gx} \|y\|$. It is not difficult to verify that

$$\|G^{-1}|_C\|^- = \inf\{\tau > 0 :\ C \cap B_Y \subset \tau G(B_X)\}. \tag{2.2}$$

3. Subsmoothness of multifunctions. Throughout the remainder of this paper, X, Y, and Z denote Banach spaces. If additional conditions are imposed, they will be explicitly specified.

Let A be a subset of X and $a \in A$. Recall (see [5, 32, 34]) that A is prox-regular at a if there exist $\sigma, \delta \in (0, +\infty)$ such that

$$\langle x^* - u^*, x - u\rangle \geq -\sigma\|x - u\|^2$$

whenever $x, u \in B(a,\delta) \cap A$, $x^* \in N_c(A, x) \cap B_{X^*}$, and $u^* \in N_c(A, u) \cap B_{X^*}$. As an interesting extension of the prox-regularity, Aussel, Daniilidis, and Thibault [1] introduced and studied the following subsmoothness and semisubsmoothness: A is said to be

(a) subsmooth at $a \in A$ if for any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\langle x^* - u^*, x - u\rangle \geq -\varepsilon\|x - u\|$$

whenever $x, u \in B(a,\delta) \cap A$, $x^* \in N_c(A, x) \cap B_{X^*}$, and $u^* \in N_c(A, u) \cap B_{X^*}$;

(b) semisubsmooth at $a \in A$ if

$$\langle x^* - a^*, x - a\rangle \geq -\varepsilon\|x - a\|$$

whenever $x \in B(a,\delta) \cap A$, $x^* \in N_c(A, x) \cap B_{X^*}$, and $a^* \in N_c(A, a) \cap B_{X^*}$.

It is easy to verify that A is subsmooth at $a \in A$ if and only if for any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\langle u^*, x - u\rangle \leq \varepsilon\|x - u\|$$

whenever $x, u \in B(a,\delta) \cap A$ and $u^* \in N_c(A, u) \cap B_{X^*}$. In the above (b), setting $x^* = 0$, one can define a weaker notion: A satisfies condition (S) at a if for any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\langle a^*, x - a\rangle \leq \varepsilon\|x - a\| \quad \forall x \in B(a,\delta) \cap A \ \text{ and } \ \forall a^* \in N_c(A, a) \cap B_{X^*}.$$

Clearly, if A satisfies condition (S), then $N_c(A, a) \subset \hat{N}(A, a)$ and so, by (2.1),

$$N_c(A, a) = N(A, a) = \hat{N}(A, a).$$

It is known (and easily verified) that

convexity⇒prox-regularity⇒subsmoothness⇒semisubsmoothness⇒condition (S).

In what follows, let $F : X \rightrightarrows Y$ be a closed multifunction, and let $a \in X$ and $b \in F(a)$.

DEFINITION 3.1. *We say that F is subsmooth (resp., satisfies condition* (S)*) at (a, b) if* $\mathrm{Gr}(F)$ *is subsmooth (resp., satisfies condition* (S)*) at (a, b).*

Now we introduce a few new notions which are weaker than the subsmoothness but stronger than condition (S). They will play an important role in our analysis.

DEFINITION 3.2. *We say that*

(i) *F is L-subsmooth at (a, b) if for any $\varepsilon > 0$ there exists $\delta > 0$ such that*

$$\langle u^*, x - a\rangle + \langle v^*, y - v\rangle \le \varepsilon(\|x - a\| + \|y - v\|) \tag{3.1}$$

whenever $v \in F(a) \cap B(b, \delta)$, $(u^*, v^*) \in N_c(\mathrm{Gr}(F), (a, v)) \cap (B_{X^*} \times B_{Y^*})$, and $(x, y) \in \mathrm{Gr}(F)$ with $\|x - a\| + \|y - b\| < \delta$;

(i′) *F is $\mathcal{L}$-subsmooth at (a, b) if F^{-1} is L-subsmooth at (b, a);*

(ii) *F is weakly L-subsmooth if same as in* (i) *but the Clarke normal cone $N_c(\mathrm{Gr}(F), \cdot)$ is replaced with the Mordukhovich normal cone $N(\mathrm{Gr}(F), \cdot)$;*

(ii′) *F is weakly $\mathcal{L}$-subsmooth at (a, b) if F^{-1} is weakly L-subsmooth at (b, a).*

It is clear that the subsmoothness of F at (a, b) implies both the L-subsmoothness and the $\mathcal{L}$-subsmoothness of F at (a, b) and that the L-subsmoothness implies the weak L-subsmoothness.

Below we provide some sufficient conditions for subsmoothness of multifunctions.

PROPOSITION 3.3. *Suppose that F is defined by $F(x) = g(x) + \Omega$ for all $x \in X$, where $g : X \to Y$ is a smooth function and Ω is a closed subset of Y. Let $(a, b) \in \mathrm{Gr}(F)$. Then the following assertions hold.*

(i) $T_c(\mathrm{Gr}(F), (a, b)) = \{(u, v) \in X \times Y : \ v \in g'(a)(u) + T_c(\Omega, b - g(a))\}$.

(ii) $N_c(\mathrm{Gr}(F), (a, b)) = \{(-(g'(a))^*(y^*), y^*) \in X^* \times Y^* : \ y^* \in N_c(\Omega, b - g(a))\}$.

(iii) *If, in addition, Ω is subsmooth at $b - g(a)$, F is subsmooth at (a, b).*

Proof. (i) Let $(u, v) \in T_c(\mathrm{Gr}(F), (a, b))$ and take sequences $\omega_n \xrightarrow{\Omega} b - g(a)$ and $t_n \downarrow 0$. Then $(a, g(a) + \omega_n) \xrightarrow{\mathrm{Gr}(F)} (a, b)$ and hence there exists a sequence $\{(u_n, v_n)\}$ converging to (u, v) such that for all $n \in \mathbb{N}$,

$$(a, g(a) + \omega_n) + t_n(u_n, v_n) \in \mathrm{Gr}(F),$$

that is, $g(a) + \omega_n + t_n v_n \in g(a + t_n u_n) + \Omega$. This means that

$$\omega_n + t_n \left(v_n - \frac{g(a + t_n u_n) - g(a)}{t_n} \right) \in \Omega \quad \forall n \in \mathbb{N}.$$

Since $v_n - \frac{g(a+t_n u_n)-g(a)}{t_n} \to v - g'(a)(u)$, we obtain that $v - g'(a)(u) \in T_c(\Omega, b - g(a))$. This shows that the set on the left-hand side of (i) is contained in the set on the right-hand side. To prove the converse inclusion, let $u \in X$ and $v \in g'(a)(u) + T_c(\Omega, b - g(a))$; take arbitrary sequences $(x_n, y_n) \xrightarrow{\mathrm{Gr}(F)} (a, b)$ and $t_n \downarrow 0$. Then there exists a sequence $\{\omega_n\}$ in Ω such that $\omega_n = y_n - g(x_n) \to b - g(a)$, and so there exists a sequence $\{\tilde{\omega}_n\}$ in Ω such that $\frac{\tilde{\omega}_n - \omega_n}{t_n} \to v - g'(a)(u)$. By the smoothness of g, it follows that

$$v_n := \frac{g(x_n + t_n u) - g(x_n)}{t_n} + \frac{\tilde{\omega}_n - \omega_n}{t_n} \to v.$$

Note that, for each $n \in \mathbb{N}$,

$$y_n + t_n v_n = g(x_n) + \omega_n + t_n v_n = g(x_n + t_n u) + \tilde{\omega}_n \in F(x_n + t_n u),$$

that is, $(x_n, y_n) + t_n(u, v_n) \in \mathrm{Gr}(F)$. Therefore $(u, v) \in T_c(\mathrm{Gr}(F), (a, b))$. This shows that the converse inclusion holds.

(ii) This follows easily from (i).

(iii) Let $\varepsilon > 0$. Since g is smooth, there exist $M, r \in (0, +\infty)$ such that

$$\|g(x) - g(u)\| \le M\|x - u\| \quad \forall x, u \in B(a, r) \tag{3.2}$$

and

$$\|g(x) - g(u) - g'(u)(x - u)\| \le \frac{\varepsilon}{2}\|x - u\| \quad \forall x, u \in B(a, r). \tag{3.3}$$

By the subsmoothness of Ω at $b - g(a)$, there exists $\delta_1 > 0$ such that

$$\langle z^*, y - z\rangle \le \frac{\varepsilon}{2(1 + M)}\|y - z\| \tag{3.4}$$

whenever $y, z \in \Omega \cap B(b - g(a), \delta_1)$ and $z^* \in N_c(\Omega, z) \cap B_{Y^*}$. By the continuity of g and the definition of F, there exists $\delta \in (0,\ r)$ such that

$$y - g(x) \in \Omega \cap B(b - g(a), \delta_1) \quad \forall (x, y) \in \mathrm{Gr}(F) \cap B((a, b), \delta). \tag{3.5}$$

Let $(x, y), (u, v) \in \mathrm{Gr}(F) \cap B((a, b), \delta)$ and $(u^*, v^*) \in N_c(\mathrm{Gr}(F), (u, v)) \cap (B_{X^*} \times B_{Y^*})$. Then, by (ii), $u^* = -(g'(u))^*(v^*)$ and $v^* \in N_c(\Omega, v - g(u)) \cap B_{Y^*}$. Thanks to (3.5), one can apply (3.4) to $y - g(x), v - g(u)$ in place of y, z and conclude that

$$\langle v^*, y - g(x) - (v - g(u))\rangle \le \frac{\varepsilon}{2(1 + M)}\|y - g(x) - (v - g(u))\|;$$

it follows from (3.3) and (3.2) that

$$\begin{aligned}
\langle (u^*, v^*), (x, y) - (u, v)\rangle &= \langle v^*, -g'(u)(x - u) + y - v\rangle \\
&\le \langle v^*, -(g(x) - g(u)) + y - v\rangle + \frac{\varepsilon\|x - u\|}{2} \\
&\le \frac{\varepsilon}{2(1 + M)}\|y - v - g(x) + g(u)\| + \frac{\varepsilon}{2}\|x - u\| \\
&\le \frac{\varepsilon}{2(1 + M)}\|y - v\| + \left(\frac{\varepsilon M}{2(M + 1)} + \frac{\varepsilon}{2}\right)\|x - u\| \\
&\le \varepsilon(\|x - u\| + \|y - v\|).
\end{aligned}$$

This shows that F is subsmooth at (a, b). The proof is complete.

In the case when $g'(a)$ is surjective, Proposition 3.3 can be strengthened as follows.

PROPOSITION 3.4. *Suppose that $F : X \rightrightarrows Y$ is defined by $F(x) = G(g(x))$ for all $x \in X$, where $g : X \to Z$ is a smooth function and $G : Z \rightrightarrows Y$ is a closed multifunction. Let $(a, b) \in \mathrm{Gr}(F)$. Suppose that $g'(a)$ is surjective and that G is $\mathcal{L}$-subsmooth (resp., subsmooth) at $(g(a), b)$. Then F is $\mathcal{L}$-subsmooth (resp., subsmooth) at (a, b).*

To prove Proposition 3.4, we need the following lemma, which is of some independent interest.

LEMMA 3.5. *Let Θ be a closed subset of Y. Let $g : X \to Y$ be strictly differentiable at $\bar{x} \in g^{-1}(\Theta)$ and suppose that $g'(\bar{x})$ is surjective. Then*

$$T_c(g^{-1}(\Theta), \bar{x}) = (g'(\bar{x}))^{-1}(T_c(\Theta, g(\bar{x}))) \tag{3.6}$$

and

$$N_c(g^{-1}(\Theta),\bar{x}) = g'(\bar{x})^*(N_c(\Theta, g(\bar{x}))). \tag{3.7}$$

Proof. Let $h \in T_c(g^{-1}(\Theta),\bar{x})$ and take any sequences $y_n \overset{\Theta}{\to} g(\bar{x})$ and $t_n \downarrow 0$. By our assumptions on $\bar{x}$, the Lyusternik–Graves theorem (cf. [24, Theorem 1.57]) can be applied and so there exists $\mu \in (0, +\infty)$ such that for all large enough n,

$$d(\bar{x}, g^{-1}(y_n)) \le \mu\|g(\bar{x}) - y_n\|.$$

It follows that there exists $x_n \in g^{-1}(y_n) \subset g^{-1}(\Theta)$ such that $x_n \to \bar{x}$. Since $h \in T_c(g^{-1}(\Theta),\bar{x})$, there exists a sequence $h_n \to h$ such that $x_n + t_n h_n \in g^{-1}(\Theta)$ for all n. On the other hand, the strict differentiability assumption implies that

$$g(x_n + t_n h_n) = y_n + g'(\bar{x})(t_n h_n) + t_n\|h_n\|\alpha_n, \tag{3.8}$$

where $\{\alpha_n\}$ is a sequence in Y converging to 0. Since $g'(\bar{x})$ is surjective, the open mapping theorem implies that there exists a sequence $\{u_n\}$ in X converging to 0 such that $g'(\bar{x})(u_n) = \alpha_n$. Hence $g'(\bar{x})(h_n + \|h_n\|u_n) \to g'(\bar{x})(h)$. Noting (by (3.8)) that

$$y_n + t_n g'(\bar{x})(h_n + \|h_n\|u_n) = g(x_n + t_n h_n) \in \Theta,$$

it follows that $g'(\bar{x})(h) \in T_c(\Theta, g(\bar{x}))$. Therefore,

$$T_c(g^{-1}(\Theta),\bar{x}) \subset (g'(\bar{x}))^{-1}(T_c(\Theta, g(\bar{x}))).$$

Conversely, let $u \in (g'(\bar{x}))^{-1}(T_c(\Theta, g(\bar{x})))$. Then $g'(\bar{x})(u) \in T_c(\Theta, g(\bar{x}))$. To prove (3.6), we have to show that $u \in T_c(g^{-1}(\Theta),\bar{x})$. To do this, let $x_n \overset{g^{-1}(\Theta)}{\longrightarrow} \bar{x}$ and $t_n \searrow 0$. Then $g(x_n) \overset{\Theta}{\to} g(\bar{x})$. Hence, there exists a sequence $v_n \to g'(\bar{x})(u)$ such that $g(x_n) + t_n v_n \in \Theta$ for all n. By the Lyusternik–Graves theorem, we assume without loss of generality that

$$d(x_n + t_n u, g^{-1}(g(x_n) + t_n v_n)) \le \mu\|g(x_n + t_n u) - g(x_n) - t_n v_n\| \tag{3.9}$$

for some $\mu \in (0, +\infty)$ and all $n \in \mathbb{N}$. By the strict differentiability of g at $\bar{x}$,

$$g(x_n + t_n u) - g(x_n) = g'(\bar{x})(t_n u) + o(t_n).$$

This and (3.9) imply that there exists $\tilde{x}_n$ with

$$\tilde{x}_n \in g^{-1}(g(x_n) + t_n v_n) \subset g^{-1}(\Theta)$$

such that

$$\|x_n + t_n u - \tilde{x}_n\| \le 2\mu(t_n\|g'(\bar{x})(u) - v_n\| + \|o(t_n)\|).$$

Then $u_n := \frac{\tilde{x}_n - x_n}{t_n} \to u$ and $x_n + t_n u_n = \tilde{x}_n \in g^{-1}(\Theta)$. This shows that $u \in T_c(g^{-1}(\Theta),\bar{x})$ as required to show. Since $g'(\bar{x})(X) = Y$, (3.7) is immediate from (3.6) and [39, Corollary 2.8.4 (ii)] (applied to $g'(\bar{x}), X, T_c(\Theta, g(\bar{x}))$ in place of A, L, M). The proof is complete.

Remark 3.1. Our formula (3.7) was inspired by Mordukhovich [24, Corollary 1.15] where the same relation was established but for Fréchet normal cones in place of Clarke normal cones. In the literature, study on the calculus of the Clarke tangent cone and

normal cone seems to be quite scarce. Nevertheless, Clarke [4, p. 108, Corollary 1] did prove the same formula but required that $g'(\bar{x})(X) \cap \operatorname{int}(T_c(\Theta, g(\bar{x})) \neq \emptyset$ and that Θ admit a hypertangent vector at $g(\bar{x})$, namely, there exist $v \in Y$ and $r > 0$ such that

$$B(g(\bar{x}), r) \cap \Theta + tB(v, r) \subset \Theta \ \ \forall t \in (0,\ r).$$

For Proposition 3.4, we shall also need the following lemma.

LEMMA 3.6. *Let $g : X \to Z$ be smooth and $a \in X$, and suppose that $g'(a)$ is surjective. Then there exist $l, r \in (0,\ +\infty)$ such that*

$$lB_Z \subset g'(u)(B_X) \ \text{ and } \ l\|z^*\| \leq \|(g'(u))^*(z^*)\| \quad \forall u \in B(a, r) \text{ and } \forall z^* \in Z^*. \tag{3.10}$$

Proof. We need only show that the inclusion in (3.10) holds for some $l, r \in (0,\ +\infty)$ (the inequality then follows easily). By the surjectivity assumption and the open mapping theorem, there exists $l \in (0,\ +\infty)$ such that $2lB_Z \subset g'(a)(B_X)$; by the smoothness of g, there exists $r > 0$ such that $\|g'(u) - g'(a)\| < \frac{l}{2}$ for all $u \in B(a, r)$. Hence,

$$2lB_Z \subset (g'(u) + (g'(a) - g'(u)))(B_X) \subset g'(u)(B_X) + \frac{l}{2}B_Z \quad \forall u \in B(a, r).$$

By the Radstrom cancellation lemma (cf. [42, Lemma 2.3]), this implies that

$$\frac{3l}{2}B_Z \subset \operatorname{cl}(g'(u)(B_X)) \quad \forall u \in B(a, r). \tag{3.11}$$

Since X, Z are Banach spaces and $g'(u)$ is a bounded linear operator from X to Z, $g'(u)(B_X)$ and $\operatorname{cl}(g'(u)(B_X))$ have the same interior (by [15, p. 183, Theorem A.1]). It follows from (3.11) that the inclusion in (3.10) holds. This completes the proof.

Proof of Proposition 3.4. We shall prove only the assertion regarding the $\mathcal{L}$-subsmoothness (the corresponding assertion regarding the subsmoothness can be proved similarly). By the smoothness and surjectivity assumption and Lemma 3.6, there exist $M, l, r \in (0,\ +\infty)$ such that (3.2) and (3.10) hold. Suppose that G is $\mathcal{L}$-subsmooth at $(g(a), b)$. Let $\varepsilon > 0$ and $\sigma := \frac{l\varepsilon}{(l+1)(M+1)}$. Then there exists $\eta > 0$ such that

$$\langle w^*, z - w\rangle + \langle v^*, y - b\rangle \leq \sigma(\|z - w\| + \|y - b\|)$$

for any $w \in G^{-1}(b) \cap B(g(a), \eta)$, $(w^*, v^*) \in N_c(\operatorname{Gr}(G), (w, b)) \cap (B_{Z^*} \times B_{Y^*})$, and $(z, y) \in \operatorname{Gr}(G)$ with $\|z - g(a)\| + \|y - b\| \leq \eta$. On the other hand, the smoothness of g implies that there exists $\delta \in (0,\ r)$ such that

$$\|g(x) - g(u) - g'(u)(x - u)\| \leq \sigma\|x - u\| \quad \forall x, u \in B(a, \delta), \tag{3.12}$$
$$g(B(a, \delta)) \subset B(g(a), \eta),$$

and

$$\|x - a\| + \|y - b\| < \delta \Longrightarrow \|g(x) - g(a)\| + \|y - b\| < \eta.$$

Let $u \in F^{-1}(b) \cap B(a, \delta)$. Then $g(u) \in G^{-1}(b) \cap B(g(a), \eta)$ and hence

$$\langle w^*, g(x) - g(u)\rangle + \langle v^*, y - b\rangle \leq \sigma(\|g(x) - g(u)\| + \|y - b\|) \tag{3.13}$$

for any $(w^*, v^*) \in N_c(\mathrm{Gr}(G), (g(u), b)) \cap (B_{Z^*} \times B_{Y^*})$ and $(x, y) \in \mathrm{Gr}(F)$ with $\|x - a\| + \|y - b\| < \delta$. Let $\tilde{g} : X \times Y \to Z \times Y$ be defined by $\tilde{g}(x, y) = (g(x), y)$ for all $(x, y) \in X \times Y$. Then $\tilde{g}$ is smooth and $\tilde{g}'(u, b)(B_X \times B_Y) = g'(u)(B_X) \times B_Y$; hence $\tilde{g}'(u, b)$ is surjective (by the first equality of (3.10)). Noting that $\mathrm{Gr}(F) = \tilde{g}^{-1}(\mathrm{Gr}(G))$, it follows from Lemma 3.5 that $N_c(\mathrm{Gr}(F), (u, b)) = (\tilde{g}'(u, b))^*(N_c(\mathrm{Gr}(G), (g(u), b)))$. This and the definition of $\tilde{g}$ imply that

$$N_c(\mathrm{Gr}(F), (u, b)) = \{((g'(u))^*(z^*), y^*) : \ (z^*, y^*) \in N_c(\mathrm{Gr}(G), (g(u), b))\}.$$

Now let $(u^*, y^*) \in N_c(\mathrm{Gr}(F), (u, b)) \cap (B_{X^*} \times B_{Y^*})$. Then there exists $z^* \in Z^*$ such that $u^* = (g'(u))^*(z^*)$ and $(z^*, y^*) \in N_c(\mathrm{Gr}(G), (g(u), b))$. It follows from (3.10) that $\|z^*\| \le \frac{1}{l}$. Thus, applying (3.13) (with $\frac{l}{1+l}(z^*, y^*)$ in place of (w^*, v^*)) and making use of (3.2), one has

$$\begin{aligned} \langle z^*, g(x) - g(u)\rangle + \langle y^*, y - b\rangle &\le \frac{\sigma(l+1)}{l}(\|g(x) - g(u)\| + \|y - b\|) \\ &\le \frac{\sigma(l+1)}{l}(M\|x - u\| + \|y - b\|) \end{aligned} \tag{3.14}$$

for any $(x, y) \in \mathrm{Gr}(F)$ with $\|x - a\| + \|y - b\| < \delta$. Moreover, (3.12) entails that for any $x \in B(a, \delta)$,

$$\begin{aligned} -\frac{\sigma}{l}\|x - u\| &\le \langle z^*, g(x) - g(u) - g'(u)(x - u)\rangle \\ &= \langle z^*, g(x) - g(u)\rangle - \langle u^*, x - u\rangle. \end{aligned}$$

This and (3.14) imply that

$$\begin{aligned} \langle u^*, x - u\rangle + \langle y^*, y - b\rangle &\le \frac{\sigma(l+1)(M+1)}{l}(\|x - u\| + \|y - b\|) \\ &= \varepsilon(\|x - u\| + \|y - b\|) \end{aligned}$$

for any $(x, y) \in \mathrm{Gr}(F)$ with $\|x - a\| + \|y - b\| < \delta$. This shows that F is $\mathcal{L}$-subsmooth at (a, b). The proof is complete.

Note that every closed convex multifunction is subsmooth at each point of its graph. The following corollary is immediate from Proposition 3.4.

COROLLARY 3.7. *Suppose that F is defined by $F = G \circ g$, namely, $F(x) = G(g(x))$ for all $x \in X$, where $g : X \to Z$ is a smooth function and $G : Z \rightrightarrows Y$ is a closed convex multifunction. Let $(a, b) \in \mathrm{Gr}(F)$ and suppose that $g'(a)$ is surjective. Then F is subsmooth at (a, b).*

4. Calmness for multifunctions. Throughout this section, let $M : Y \rightrightarrows X$ be a closed multifunction. We also fix (arbitrary) $\bar{y} \in Y$ and $\bar{x} \in M(\bar{y})$.

It is easy to verify that M is calm at $(\bar{y}, \bar{x})$ if and only if there exist $\tau, \delta \in (0, +\infty)$ such that

$$M(y) \cap B(\bar{x}, \delta) \subset M(\bar{y}) + \tau\|y - \bar{y}\|B_X \ \ \forall\ y \text{ close to } \bar{y}. \tag{4.1}$$

Motivated by the notion of linear cover property (cf. [7, 23, 28]), let us say that a multifunction $\Phi : X \rightrightarrows Y$ has the linear cover-like property at $(\bar{x}, \bar{y})$ if there exists $\tau \in (0, +\infty)$ such that for all x close to $\bar{x}$ and $r > 0$

$$\bar{y} \in \Phi(x) + \mathrm{int}(rB_Y) \implies \bar{y} \in \Phi(x + \mathrm{int}(\tau r B_X)). \tag{4.2}$$

In terms of (4.1) and (4.2), the following proposition provides formulas for the calmness modulus $\eta(M;\bar{y},\bar{x})$ (which is defined by (1.5)); we omit its proof as it is immediate from the related definitions.

PROPOSITION 4.1.

$$\begin{aligned}\eta(M;\bar{y},\bar{x}) &= \inf\{\tau>0:\ (4.1)\text{ holds for some }\delta>0\}\\ &= \inf\{\tau>0:\ (4.2)\text{ holds with }\Phi=M^{-1}\ \forall\ r>0\text{ and }\forall\ x\text{ close to }\bar{x}\}.\end{aligned}$$

The remainder of this section is devoted to a study on the duality aspect of the calmness. We divide our discussion into two subsections addressing the necessary conditions and the sufficient conditions for calmness.

4.1. Necessary conditions for calmness. There are two results in this subsection: one is on the Banach space setting and the other on the Asplund spaces.

THEOREM 4.2. *Suppose that there exist* $\eta,\delta\in(0,\ +\infty)$ *such that* (1.1) *holds. Then*

$$\hat{N}(M(\bar{y}),u)\cap B_{X^*}\subset \eta D^*_cM^{-1}(u,\bar{y})(B_{Y^*})\quad\forall u\in M(\bar{y})\cap B(\bar{x},\delta). \tag{4.3}$$

Proof. Let $\mathcal{I}_{\mathrm{Gr}(M^{-1})}$ denote the indicator function of $\mathrm{Gr}(M^{-1})$. Then (1.1) can be rewritten as

$$d(x,M(\bar{y}))\le \mathcal{I}_{\mathrm{Gr}(M^{-1})}(x,y)+\eta\|y-\bar{y}\|\quad\forall(x,y)\in B(\bar{x},\delta)\times B(\bar{y},\delta). \tag{4.4}$$

Let $u\in M(\bar{y})\cap B(\bar{x},\delta)$ and $u^*\in\hat{N}(M(\bar{y}),u)\cap B_{X^*}$. Noting (cf. [24, Corollary 1.96]) that $\hat{N}(M(\bar{y}),u)\cap B_{X^*}=\hat{\partial}d(\cdot,M(\bar{y})))(u)$, it follows that for any $\sigma>0$ there exists $r\in(0,\ \delta)$ such that $B(u,r)\subset B(\bar{x},\delta)$ and

$$\langle u^*,x-u\rangle\le d(x,M(\bar{y}))+\sigma\|x-u\|\quad\forall x\in B(u,r). \tag{4.5}$$

Hence, by (4.4),

$$\langle u^*,x-u\rangle\le \mathcal{I}_{\mathrm{Gr}(M^{-1})}(x,y)+\eta\|y-\bar{y}\|+\sigma\|x-u\|\quad\forall(x,y)\in B(u,r)\times B(\bar{y},\delta),$$

that is, $(u,\bar{y})$ is a local minimizer of ϕ defined by

$$\phi(x,y):=-\langle u^*,x-u\rangle+\mathcal{I}_{\mathrm{Gr}(M^{-1})}(x,y)+\eta\|y-\bar{y}\|+\sigma\|x-u\|\quad\forall(x,y)\in X\times Y.$$

Hence, $(0,0)\in\partial_c\phi(u,\bar{y})$. It follows from [4, Theorem 2.9.8] that

$$(0,0)\in(-u^*,0)+N_c(\mathrm{Gr}(M^{-1}),(u,\bar{y}))+\{0\}\times\eta B_{Y^*}+(\sigma B_{X^*})\times\{0\},$$

that is,

$$(u^*+\sigma x^*_\sigma,-\eta y^*_\sigma)\in N_c(\mathrm{Gr}(M^{-1}),(u,\bar{y}))$$

for some $x^*_\sigma\in B_{X^*}$ and $y^*_\sigma\in B_{Y^*}$. Since B_{Y^*} is weak* compact, without loss of generality we can assume $(u^*+\sigma x^*_\sigma,-\eta y^*_\sigma)\overset{w^*}{\to}(u^*,-\eta v^*)$ for some v^* in B_{Y^*} as $\sigma\to 0^+$. Hence $(u^*,-\eta v^*)\in N_c(\mathrm{Gr}(M^{-1}),(u,\bar{y}))$ (because $N_c(\mathrm{Gr}(M^{-1}),(u,\bar{y}))$ is weak*-closed). This implies that

$$u^*\in D^*_cM^{-1}(u,\bar{y})(\eta v^*)\subset\eta D^*_cM^{-1}(u,\bar{y})(B_{Y^*}).$$

This shows that (4.3) holds. The proof is complete.

When Y, X are Asplund spaces, the conclusion in Theorem 4.2 can be strengthened with $\hat{N}(M(\bar{y}), u)$ and $D_c^* M^{-1}(u, \bar{y})$ replaced, respectively, by $N(M(\bar{y}), u)$ and $D^* M^{-1}(u, \bar{y})$.

THEOREM 4.3. *Suppose that Y, X are Asplund spaces and that there exist $\eta, \delta \in (0, +\infty)$ such that (1.1) holds. Then*

$$N(M(\bar{y}), u) \cap B_{X^*} \subset \eta D^* M^{-1}(u, \bar{y})(B_{Y^*}) \quad \forall u \in M(\bar{y}) \cap B(\bar{x}, \delta). \tag{4.6}$$

Proof. Let $u \in M(\bar{y}) \cap B(\bar{x}, \delta)$ and $u^* \in B_{X^*} \cap N(M(\bar{y}), u)$. Then there exist sequences $\{u_n\}$ in $M(\bar{y}) \cap B(\bar{x}, \delta)$ and $\{u_n^*\}$ in X^* such that

$$u_n \to u, \quad u_n^* \xrightarrow{w^*} u^*, \ \text{ and } \ u_n^* \in \hat{N}(M(\bar{y}), u_n) \quad \forall n \in \mathbb{N}.$$

Similar to the proof of (4.5), there exists $r \in (0, \delta)$ such that $B(u_n, r) \subset B(\bar{x}, \delta)$ and

$$\langle u_n^*, x - u_n \rangle \le d(x, M(\bar{y})) + \frac{1}{n}\|x - u_n\| \quad \forall x \in B(u_n, r).$$

Letting

$$\phi(x, y) := -\langle u_n^*, x - u_n \rangle + \mathcal{I}_{\mathrm{Gr}(M^{-1})}(x, y) + \eta\|y - \bar{y}\| + \frac{1}{n}\|x - u_n\| \quad \forall (x, y) \in X \times Y,$$

from the corresponding part of the proof of Theorem 4.2, it follows that $(u_n, \bar{y})$ is a local minimizer of ϕ. This and [24, Theorem 2.33] imply that there exists $(w_n, y_n) \in \mathrm{Gr}(M^{-1})$ such that $\|w_n - u_n\| + \|y_n - \bar{y}\| < \frac{1}{n}$ and

$$(0, 0) \in (-u_n^*, 0) + \hat{N}(\mathrm{Gr}(M^{-1}), (w_n, y_n)) + \{0\} \times \eta B_{Y^*} + \frac{2}{n}(B_{X^*} \times B_{Y^*}).$$

Therefore, $(w_n, y_n) \to (u, \bar{y})$ and there exist $x_n^* \in B_{X^*}$ and $y_n^*, v_n^* \in B_{Y^*}$ such that

$$\left(u_n^* + \frac{2}{n}x_n^*, -\eta y_n^* - \frac{2}{n}v_n^*\right) \in \hat{N}(\mathrm{Gr}(M^{-1}), (w_n, y_n)).$$

Since B_{Y^*} is sequentially weak*-compact (as Y is an Asplund space), we can assume that $y_n^* \xrightarrow{w^*} y^* \in B_{Y^*}$ as $n \to \infty$. It follows that $(u^*, -\eta y^*) \in N(\mathrm{Gr}(M^{-1}), (u, \bar{y}))$ and so $u^* \in D^* M^{-1}(u, \bar{y})(\eta y^*)$. Therefore, (4.6) holds. The proof is complete.

Remark 4.1. In Asplund spaces, the limiting subdifferential enjoys, like the Clarke subdifferential, the full sum rule, but, on the other hand, the Mordukhovich normal cone is not necessarily weak*-closed. This is why the last part of the proof of Theorem 4.3 differs from that of Theorem 4.2.

4.2. Sufficient conditions for calmness of L-subsmooth multifunctions. Under a suitable L-subsmoothness assumption, we show in the next result that a slightly stronger condition than (4.3) turns out to be sufficient for calmness.

THEOREM 4.4. *Suppose that M is L-subsmooth (resp., weakly L-subsmooth) at $(\bar{y}, \bar{x})$ and that there exist $\eta, \delta \in (0, +\infty)$ such that*

$$N_c(M(\bar{y}), u) \cap B_{X^*} \subset \eta D_c^* M^{-1}(u, \bar{y})(B_{Y^*}) \quad \forall u \in \mathrm{bd}(M(\bar{y})) \cap B(\bar{x}, \delta) \tag{4.7}$$

$$(\textit{resp.}, \ N_c(M(\bar{y}), u) \cap B_{X^*} \subset \eta D^* M^{-1}(u, \bar{y})(B_{Y^*}) \quad \forall u \in \mathrm{bd}(M(\bar{y})) \cap B(\bar{x}, \delta)).$$

Then M is calm at $(\bar{y},\bar{x})$ and, more precisely, for any $\varepsilon \in (0,\ \frac{1}{1+\eta})$ there exists $\delta_\varepsilon > 0$ such that

$$d(x,M(\bar{y})) \le \frac{\eta+(1+\eta)\varepsilon}{1-(1+\eta)\varepsilon}\|y-\bar{y}\| \quad \forall y \in B(\bar{y},\delta_\varepsilon) \text{ and } \forall x \in M(y)\cap B(\bar{x},\delta_\varepsilon). \tag{4.8}$$

Proof. We provide only the proof for the assertion under the L-subsmoothness assumption (the proof for the other part is similar). Let $\varepsilon \in (0,\ \frac{1}{1+\eta})$. Then, by the L-subsmoothness assumption, there exists $\delta_\varepsilon \in (0,\ \frac{\delta}{2})$ such that

$$-\langle v^*, y-\bar{y}\rangle + \langle u^*, x-u\rangle \le \varepsilon(\|y-\bar{y}\|+\|x-u\|) \tag{4.9}$$

whenever $y \in B(\bar{y},2\delta_\varepsilon)$, $x \in M(y)\cap B(\bar{x},2\delta_\varepsilon)$, $u \in M(\bar{y})\cap B(\bar{x},2\delta_\varepsilon)$, $v^* \in B_{Y^*}$, and $u^* \in D_c^*M^{-1}(u,\bar{y})(v^*)\cap B_{X^*}$. To verify (4.8), let $y \in B(\bar{y},\delta_\varepsilon)$ and $x \in M(y)\cap (B(\bar{x},\delta_\varepsilon)\setminus M(\bar{y}))$. Then $d(x,M(\bar{y})) \le \|x-\bar{x}\| < \delta_\varepsilon$. Let

$$\gamma \in \left(\max\left\{\frac{d(x,M(\bar{y}))}{\delta_\varepsilon}, (1+\eta)\varepsilon, \frac{1}{2}\right\},\ 1\right).$$

By Lemma 2.1 there exist $u \in \mathrm{bd}(M(\bar{y}))$ and $u^* \in N_c(M(\bar{y}),u)$ with $\|u^*\| = 1$ such that

$$\gamma\|x-u\| \le \min\{\langle u^*, x-u\rangle, d(x,M(\bar{y}))\}. \tag{4.10}$$

Thus, $\|x-u\| \le \frac{d(x,M(\bar{y}))}{\gamma} < \delta_\varepsilon$. Hence

$$\|u-\bar{x}\| \le \|u-x\| + \|x-\bar{x}\| < 2\delta_\varepsilon < \delta.$$

By (4.7), there exists $v^* \in \eta B_{Y^*}$ such that $u^* \in D_c^*M^{-1}(u,\bar{y})(v^*)$. Applying (4.9) with $(\frac{u^*}{1+\eta}, \frac{v^*}{1+\eta})$ in place of (u^*,v^*), it follows that

$$-\langle v^*, y-\bar{y}\rangle + \langle u^*, x-u\rangle \le (1+\eta)\varepsilon(\|y-\bar{y}\|+\|x-u\|)$$

and so

$$\begin{aligned}\langle u^*, x-u\rangle - (1+\eta)\varepsilon\|x-u\| &\le \langle v^*, y-\bar{y}\rangle + (1+\eta)\varepsilon\|y-\bar{y}\| \\ &\le (\eta+(1+\eta)\varepsilon)\|y-\bar{y}\|.\end{aligned}$$

This and (4.10) imply that

$$(\gamma-(1+\eta)\varepsilon)\|x-u\| \le (\eta+(1+\eta)\varepsilon)\|y-\bar{y}\|$$

and hence

$$d(x,M(\bar{y})) \le \frac{\eta+(1+\eta)\varepsilon}{\gamma-(1+\eta)\varepsilon}\|y-\bar{y}\|$$

(because $u \in M(\bar{y})$). Letting $\gamma \to 1$, it follows that (4.8) holds. The proof is complete.

The following example shows that the L-subsmoothness assumption cannot be dropped in Theorem 4.4.

Example 4.5. Let $X = Y = R$ and let

$$\Omega_1 = \{(s,t) \in R^2 :\ s^2 + (t-1)^2 \le 1 \text{ and } (s-1)^2 + t^2 \le 1\},$$
$$\Omega_2 = \{(s,t) \in R^2 :\ s^2 + (t+1)^2 \le 1 \text{ and } (s-1)^2 + t^2 \le 1\},$$
$$\Omega_3 = \{(s,t) \in R^2 :\ (s+1)^2 + t^2 \le 1 \text{ and } s^2 + (t+1)^2 \le 1\},$$
$$\Omega_4 = \{(s,t) \in R^2 :\ (s+1)^2 + t^2 \le 1 \text{ and } s^2 + (t-1)^2 \le 1\}.$$

Define the multifunction $M : Y \rightrightarrows X$ such that $\mathrm{Gr}(M) = \bigcup_{i=1}^4 \Omega_i$. Then $M(0) = \{0\}$ and so $N_c(M(0), 0) = X^*$. It is easy to verify that $N_c(\mathrm{Gr}(M), (0,0)) = X^* \times Y^*$. Hence

$$N_c(M(0),0) \cap B_{X^*} = B_{X^*} \subset \tau D_c^* M^{-1}(0,0)(B_{Y^*}) = X^* \quad \forall \tau \in (0,\ +\infty).$$

On the other hand, note that

$$\frac{1}{n} \in M\left(\frac{1}{n^2\left(1+\sqrt{1-\frac{1}{n^2}}\right)}\right) \quad \text{and} \quad \frac{d(\frac{1}{n}, M(0))}{\left\|\frac{1}{n^2\left(1+\sqrt{1-\frac{1}{n^2}}\right)} - 0\right\|} \to +\infty.$$

Hence M is not calm at $(0,0)$.

Recall [43] that M is strongly calm at $(\bar{y}, \bar{x})$ if there exist $\eta, \delta \in (0,\ +\infty)$ such that

$$\|x - \bar{x}\| \le \eta \|y - \bar{y}\| \quad \forall y \in B(\bar{y}, \delta) \text{ and } x \in M(y) \cap B(\bar{x}, \delta).$$

It is clear that M is strongly calm at $(\bar{y}, \bar{x})$ if and only if $\bar{x}$ is an isolated point of $M(\bar{y})$ (i.e., $M(\bar{y}) \cap B(\bar{x}, r) = \{\bar{x}\}$ for some $r > 0$) and M is calm at $(\bar{y}, \bar{x})$.

COROLLARY 4.6. *Suppose that M satisfies condition* (S) *at $(\bar{y}, \bar{x})$. Then M is strongly calm at $(\bar{y}, \bar{x})$ if and only if*

$$D_c^* M^{-1}(\bar{x}, \bar{y})(Y^*) = X^*. \tag{4.11}$$

Proof. Suppose that (4.11) holds. Since $D_c^* M^{-1}(\bar{x}, \bar{y})$ is a closed convex multifunction from Y^* to X^*, (4.11) and the Robinson–Ursescu theorem (cf. [33, 36]) imply that there exists $\eta > 1$ such that

$$\frac{1}{\eta} B_{X^*} \subset D_c^* M^{-1}(\bar{x}, \bar{y})(B_{Y^*}) \cap B_{X^*}. \tag{4.12}$$

Hence, by the Hahn–Banach theorem,

$$\frac{1}{\eta}\|u\| \le \max\{\langle u^*, u\rangle :\ u^* \in D_c^* M^{-1}(\bar{x}, \bar{y})(B_{Y^*}) \cap B_{X^*}\} \quad \forall u \in X. \tag{4.13}$$

Consider $\varepsilon \in (0,\ \frac{1}{\eta})$. The condition (S) assumption implies that there exists $\delta > 0$ such that

$$\langle u^*, x - \bar{x}\rangle \le \varepsilon \|x - \bar{x}\| \quad \forall x \in M(\bar{y}) \cap B(\bar{x}, \delta) \text{ and } \forall u^* \in D_c^* M^{-1}(\bar{x}, \bar{y})(B_{Y^*}) \cap B_{X^*}.$$

It follows from (4.13) that $M(\bar{y}) \cap B(\bar{x}, \delta) = \{\bar{x}\}$. This entails that M is L-subsmooth at $(\bar{y}, \bar{x})$ and (4.7) holds (due to the condition (S) assumption and (4.12), respectively). Therefore, Theorem 4.4 can be applied to conclude that M is calm at $(\bar{y}, \bar{x})$.

Conversely, suppose that M is strongly calm at $(\bar{y},\bar{x})$. Then $M(\bar{y})\cap B(\bar{x},r)=\{\bar{x}\}$ for some $r>0$ (so $\hat{N}(M(\bar{y}),\bar{x})=X^*$), and Theorem 4.2 implies that there exist $\eta,\delta\in(0,\,+\infty)$ such that (4.3) holds and so does (4.11). The proof is complete.

COROLLARY 4.7. *Suppose that Y,X are finite-dimensional and that $\mathrm{Gr}(M)$ is Clarke regular at $(\bar{y},\bar{x})$ (i.e., $T_c(\mathrm{Gr}(M),(\bar{y},\bar{x}))=T(\mathrm{Gr}(M),(\bar{y},\bar{x}))$). Then M is strongly calm at $(\bar{y},\bar{x})$ if and only if $D_c^*M^{-1}(\bar{x},\bar{y})(Y^*)=X^*$.*

Proof. By Corollary 4.6, we need only show that M satisfies condition (S) at $(\bar{y},\bar{x})$. To do this, suppose to the contrary that there exist $\varepsilon_0>0$, a sequence $\{(y_n,x_n)\}$ in $\mathrm{Gr}(M)\setminus\{(\bar{y},\bar{x})\}$, and a sequence $\{(v_n^*,u_n^*)\}$ in $N_c(\mathrm{Gr}(M),(\bar{y},\bar{x}))\cap(B_{Y^*}\times B_{X^*})$ such that $(y_n,x_n)\to(\bar{y},\bar{x})$ and

$$\langle v_n^*,y_n-\bar{y}\rangle+\langle u_n^*,x_n-\bar{x}\rangle>\varepsilon_0(\|y_n-\bar{y}\|+\|x_n-\bar{x}\|)\quad\forall n.$$

Since Y,X are finite-dimensional, we can assume that

$$\frac{(y_n-\bar{y},x_n-\bar{x})}{\|y_n-\bar{y}\|+\|x_n-\bar{x}\|}\to(v,u)\ \text{ and }\ (v_n^*,u_n^*)\to(v^*,u^*)$$

for some $(v,u)\in Y\times X$ and $(v^*,u^*)\in Y^*\times X^*$. Then $\langle v^*,v\rangle+\langle u^*,v\rangle\ge\varepsilon_0$, $(v,u)\in T(\mathrm{Gr}(M),(\bar{y},\bar{x}))$, and $(v^*,u^*)\in N_c(\mathrm{Gr}(M),(\bar{y},\bar{x}))$. This contradicts the Clarke regularity assumption. The proof is complete.

When X is an Asplund space, the assumption in Theorem 4.4 can be weakened with $N_c(M(\bar{y}),u)$ replaced by $\hat{N}(M(\bar{y}),u)$.

THEOREM 4.8. *Suppose that X is an Asplund space and that M is L-subsmooth (resp., weakly L-subsmooth) at $(\bar{y},\bar{x})$ and that there exist $\eta,\delta\in(0,\,+\infty)$ such that*

$$\hat{N}(M(\bar{y}),u)\cap B_{X^*}\subset\eta D_c^*M^{-1}(u,\bar{y})(B_{Y^*})\quad\forall u\in M(\bar{y})\cap B(\bar{x},\delta)$$

$$(\textit{resp.},\ \hat{N}(M(\bar{y}),u)\cap B_{X^*}\subset\eta D^*M^{-1}(u,\bar{y})(B_{Y^*})\quad\forall u\in M(\bar{y})\cap B(\bar{x},\delta)).$$

Then for any $\varepsilon>0$ there exists $\delta_\varepsilon>0$ such that (4.8) *holds.*

The proof of Theorem 4.8 is the same as that of Theorem 4.4, but with the Asplund space version of Lemma 2.1 applied in place of the Banach space version.

Making use of (2.2) and the equivalence

$$x^*\in(D_c^*M(\bar{y},u))^{-1}(y^*)\Leftrightarrow -x^*\in D_c^*M^{-1}(u,\bar{y})(-y^*),$$

it is easy from Theorem 4.2 to verify part (i) of the following corollary. Similarly, part (ii) follows from Theorem 4.4.

COROLLARY 4.9. *The following assertions hold.*

(i) $\eta(M;\bar{y},\bar{x})\ge\limsup_{u\in M(\bar{y}),u\to\bar{x}}\|D_c^*M(\bar{y},u)|_{-\hat{N}(M(\bar{y}),u)}\|^-$.

(ii) *If M is L-subsmooth at $(\bar{y},\bar{x})$, then*

$$\eta(M;\bar{y},\bar{x})\le\limsup_{u\in M(\bar{y}),u\to\bar{x}}\|D_c^*M(\bar{y},u)|_{-N_c(M(\bar{y}),u)}\|^-.$$

Similarly, one can use Theorems 4.3 and 4.8 to show the following corollary.

COROLLARY 4.10. *Suppose that Y,X are Asplund spaces. Then the following assertions hold.*

(i) $\eta(M;\bar{y},\bar{x})\ge\limsup_{u\in M(\bar{y}),u\to\bar{x}}\|D^*M(\bar{y},u)|_{-\hat{N}(M(\bar{y}),u)}\|^-$.

(ii) *If M is weakly L-subsmooth at $(\bar{y},\bar{x})$, then the equality in* (i) *holds.*

Note that, in the Asplund space setting, the Fréchet normal cone $\hat{N}(M(\bar{y}),u)$ is used in Theorem 4.8. However, for the general Banach spaces, one needs to use the

Clarke normal cone $N_c(M(\bar{y}),u)$ in Theorem 4.4; we have only the inequality version in Corollary 4.9(ii) while we have the equality version in Corollary 4.10(ii).

The following result concerns "convex-composite" multifunctions. Optimization problems involving convex-composite functions have been extensively studied (for details, see [16, 30, 35, 37] and the references therein).

THEOREM 4.11. *Suppose that M is defined by $M = g^{-1} \circ G$, namely, $M(y) = g^{-1}(G(y))$ for all $y \in Y$, where $G : Y \rightrightarrows Z$ is a closed convex multifunction and $g : X \to Z$ is a smooth function. Let $(\bar{y},\bar{x}) \in \mathrm{Gr}(M)$ and suppose that $g'(\bar{x})$ is surjective. Then the following assertions hold.*

(i) $\eta(M;\bar{y},\bar{x}) = \limsup_{u\in M(\bar{y}),u\to\bar{x}} \|D^*M(\bar{y},u)|_{-N(M(\bar{y}),u)}\|^-$.

(ii) *If there exist $\eta,\delta \in (0,\ +\infty)$ such that*

$$N(M(\bar{y}),u)\cap B_{X^*} \subset \eta D^*M^{-1}(u,\bar{y})(B_{Y^*}) \quad \forall u \in M(\bar{y})\cap B(\bar{x},\delta), \tag{4.14}$$

then for any $\varepsilon > 0$ there exists $\delta_\varepsilon \in (0,\ \delta)$ such that

$$d(x,M(\bar{y})) \le \frac{\eta}{1-\varepsilon}\|y-\bar{y}\| \tag{4.15}$$

whenever $y \in B(\bar{y},\delta_\varepsilon)$ and $x \in M(y)\cap B(\bar{x},\delta_\varepsilon)$.

Proof. By Corollary 3.7, M^{-1} $(= G^{-1}\circ g)$ is subsmooth at $(\bar{x},\bar{y})$ and so is M at $(\bar{y},\bar{x})$. Let $\tilde{g}(y,x) := (y,g(x))$ for all $(y,x) \in Y \times X$. Then it follows from the surjectivity of $g'(\bar{x})$ that $\tilde{g}'(\bar{y},\bar{x})$ is surjective. By Lemma 3.6, take $r > 0$ such that $g'(u)$ and $\tilde{g}'(y,u)$ are surjective for all $(y,u) \in B(\bar{y},r)\times B(\bar{x},r)$. Noting that $M(\bar{y}) = g^{-1}(G(\bar{y}))$ and $\mathrm{Gr}(M) = \tilde{g}^{-1}(\mathrm{Gr}(G))$, it follows from Lemma 3.5 and [24, Theorem 1.17] that

$$\hat{N}(M(\bar{y}),u) = N(M(\bar{y}),u) = N_c(M(\bar{y}),u) \quad \forall u \in M(\bar{y})\cap B(\bar{x},r) \tag{4.16}$$

and

$$\hat{N}(\mathrm{Gr}(M),(y,u)) = N(\mathrm{Gr}(M),(y,u)) = N_c(\mathrm{Gr}(M),(y,u))$$

for all $(y,u) \in \mathrm{Gr}(M)\cap(B(\bar{y},r)\times B(\bar{x},r))$; hence

$$D^*M(\bar{y},u) = D^*_cM(\bar{y},u) \quad \forall u \in M(\bar{y})\cap B(\bar{x},r).$$

Thus (i) follows from Corollary 4.9.

To prove (ii), let $\eta,\delta \in (0,\ +\infty)$ satisfy (4.14). Let $\varepsilon \in (0,\ 1)$. Then, by the subsmoothness of M at $(\bar{y},\bar{x})$, there exists $\delta_\varepsilon \in (0,\ r)$ such that

$$\langle u^*, x-u\rangle - \langle v^*, y-\bar{y}\rangle \le \varepsilon\|x-u\| \tag{4.17}$$

whenever $y \in B(\bar{y},2\delta_\varepsilon)$, $x \in M(y)\cap B(\bar{x},2\delta_\varepsilon)$, $u \in M(\bar{y})\cap B(\bar{x},2\delta_\varepsilon)$, $v^* \in Y^*$, and $u^* \in D^*M^{-1}(u,\bar{y})(v^*)\cap B_{X^*}$. Let $y \in B(\bar{y},\delta_\varepsilon)$ and $x \in B(\bar{x},\delta_\varepsilon)\setminus M(\bar{y})$. We have to show that (4.15) holds. To do this, let $\gamma \in \big(\max\big\{\frac{d(x,M(\bar{y}))}{\delta_\varepsilon},\ \frac{1}{2}\big\},\ 1\big)$ sufficiently close to 1. By Lemma 2.1, as in the corresponding part of the proof of Theorem 4.4, there exist $u \in \mathrm{bd}(M(\bar{y}))\cap B(\bar{x},2\delta_\varepsilon)$ and $u^* \in N_c(M(\bar{y}),u)$ with $\|u^*\| = 1$ such that (4.10) holds. It follows from (4.14) and (4.16) that there exists $v^* \in \eta B_{Y^*}$ such that $u^* \in D^*M^{-1}(u,\bar{y})(v^*)$. By (4.10) and (4.17), one has

$$(\gamma-\varepsilon)\|x-u\| \le \langle u^*, x-u\rangle - \varepsilon\|x-u\| \le \langle v^*, y-\bar{y}\rangle \le \eta\|y-\bar{y}\|$$

and so $(\gamma-\varepsilon)d(x,M(\bar{y}))\le\eta\|y-\bar{y}\|$. Letting $\gamma\to 1$, it follows that (4.15) holds. The proof is complete.

Remark 4.2. Motivated by (1.4), a natural problem is whether the upper limit

$$\limsup_{u\in M(\bar{y}),u\to\bar{x}}\|D_c^*M(\bar{y},u)|_{-N_c(M(\bar{y}),u)}\|^-$$

in Corollaries 4.9 and 4.10 and Theorem 4.11 can be replaced with

$$\|D_c^*M(\bar{y},\bar{x})|_{-N_c(M(\bar{y}),\bar{x})}\|^-.$$

The answer is negative even when M is convex and X,Y are finite-dimensional. Below we give an example of a closed convex multifunction M between two finite-dimensional spaces such that

$$\|D_c^*M(\bar{x},\bar{y})|_{-N_c(M(\bar{y}),\bar{x})}\|^-=0\quad\text{but}\quad\eta(M;\bar{y},\bar{x})=+\infty.$$

Let $S=\{(u,v)\in R^2:\ u^2+v^2\le 1\}$ and $C=\{(u,v)\in R^2:\ v\le 1\}$. Let $M:R\rightrightarrows R^2$ be defined by

$$M(y):=\{x\in C:\ d^2(x,S)\le y\}\quad\forall y\in R.$$

Then M is a closed convex multifunction (because C and S are closed convex sets). Take $\bar{y}=0$ and $\bar{x}=(0,1)$. Then $S=M(\bar{y})$ and $\bar{x}\in M(\bar{y})$. It is clear that

$$d(x,M(\bar{y})=d(x,S)\ \text{ and }\ x\in M(d^2(x,S))\quad\forall x\in C.$$

This shows that M is not calm at $(\bar{y},\bar{x})$, that is, $\eta(M;\bar{y},\bar{x})=+\infty$. Next we show that $\|D_c^*M(\bar{y},\bar{x})|_{-N_c(M(\bar{y}),\bar{x}))}\|^-=0$. It is easy from the convexity of $M(\bar{y})$ to verify that

$$N_c(M(\bar{y}),\bar{x})=N(S,\bar{x})=N(C,\bar{x})=\{0\}\times R_+.$$

Let $t,r\in(0,\ +\infty)$. Then, for any $x\in C$ and $y\in M^{-1}(x)=[d^2(x,S),\ +\infty)$, one has

$$\langle(0,t),x\rangle-r(y-\bar{y})\le\langle(0,t),\bar{x}\rangle.$$

This and the convexity of M imply that $((0,t),-r)\in N_c(\mathrm{Gr}(M^{-1}),(\bar{x},\bar{y}))$ and so $(0,t)\in D_c^*M^{-1}(\bar{x},\bar{y})(r)$. Since t and r are arbitrary in $(0,\ +\infty)$, $N_c(M(\bar{y}),\bar{x})\subset D_c^*M^{-1}(\bar{x},\bar{y})(r)$. This shows that

$$-N_c(M(\bar{y}),\bar{x})\subset(D_c^*M(\bar{y},\bar{x}))^{-1}(-r)\subset(D_c^*M(\bar{y},\bar{x}))^{-1}(rB_{Y^*})\ \ \forall r>0.$$

It follows from (2.2) that $\|D_c^*M(\bar{y},\bar{x})|_{-N_c(M(\bar{y}),\bar{x})}\|^-=0$.

Nevertheless, in the convex-composite case, $\|D_c^*M(\bar{y},\bar{x})|_{-N_c(M(\bar{y}),\bar{x})}\|^-<+\infty$ does imply the calmness of the sublinear multifunction $D_cM(\bar{y},\bar{x})$. First, we provide a result in a general case.

PROPOSITION 4.12. *Suppose that*

$$T_c(M(\bar{y}),\bar{x})\subset D_cM(\bar{y},\bar{x})(0).\tag{4.18}$$

Then

$$\|D_c^*M(\bar{y},\bar{x})|_{-N_c(M(\bar{y}),\bar{x})}\|^-=\inf\{\eta>0:\ d(x,T_c(M(\bar{y}),\bar{x}))\le\eta\|y\|\ \ \forall y\in Y\ \textit{and}\ x\in D_cM(\bar{y},\bar{x})(y))\}.$$

If, in addition, $\|D_c^*M(\bar{y},\bar{x})_{-N_c(M(\bar{y}),\bar{x})}\|^- < +\infty$, *then*

$$T_c(M(\bar{y}),\bar{x}) = D_cM(\bar{y},\bar{x})(0)$$

and so $\eta(D_cM(\bar{y},\bar{x});0,0) = \|D_c^*M(\bar{y},\bar{x})|_{-N_c(M(\bar{y},\bar{x})}\|^-$.

Proof. Let

$$\eta_* := \inf\{\eta > 0 :\ d(x, T_c(M(\bar{y}),\bar{x})) \le \eta\|y\| \ \ \forall y \in Y \text{ and } x \in D_cM(\bar{y},\bar{x})(y))\}.$$

First, we assume that $\eta_* < \infty$. Consider any $\eta \in (\eta_*,\ \infty)$. Then

$$d(x, T_c(M(\bar{y},\bar{x})) \le \eta\|y\| \ \ \forall y \in Y \text{ and } x \in D_cM(\bar{y},\bar{x})(y).$$

It follows that $D_cM(\bar{y},\bar{x})(0) \subset T_c(M(\bar{y}),\bar{x})$. This and (4.18) imply that

$$T_c(M(\bar{y}),\bar{x}) = D_cM(\bar{y},\bar{x})(0).$$

Hence $\eta(D_c(M(\bar{y};\bar{x}),0,0) = \eta_*$. Noting that $D_cM(\bar{y},\bar{x})$ is a closed convex multifunction, it follows from [43, Theorem 4.3] that

$$\|D_c^*M(\bar{y},\bar{x})|_{-N_c(M(\bar{y}),\bar{x})}\|^- = \|D_c^*(D_cM(\bar{y},\bar{x}))(0,0)|_{-N_c(D_cM(\bar{y},\bar{x})(0),0)}\|^- = \eta_*.$$

It remains to show that $\|D_c^*M(\bar{y},\bar{x})|_{-N_c(M(\bar{y}),\bar{x})}\|^- = +\infty$ if $\eta_* = +\infty$. Suppose that $\|D_c^*M(\bar{y},\bar{x})|_{-N_c(M(\bar{y}),\bar{x})}\|^- < +\infty$. We need only show that $\eta_* < +\infty$. Let $x \in X \setminus T_c(M(\bar{y}),\bar{x})$ and $\gamma \in (0,\ 1)$. By Lemma 2.1 there exist

$$z \in T_c(M(\bar{y}),\bar{x}) \ \text{ and } \ z^* \in N_c(T_c(M(\bar{y}),\bar{x}),z)$$

such that

$$\|z^*\| = 1 \ \text{ and } \ \langle z^*, x-z\rangle \ge \gamma\|x-z\|. \tag{4.19}$$

Noting that $T_c(M(\bar{y}),\bar{x})$ is a convex cone, it is easy to verify that

$$N_c(T_c(M(\bar{y}),\bar{x}),z) \subset N_c(T_c(M(\bar{y}),\bar{x}),0) = N_c(M(\bar{y}),\bar{x}).$$

Therefore, $z^* \in N_c(M(\bar{y}),\bar{x})$. Let $\eta \in (\|D_c^*M(\bar{y},\bar{x})|_{-N_c(M(\bar{y}),\bar{x})}\|^-,\ \infty)$. Then there exists $y^* \in D_c^*M(\bar{y},\bar{x})(-z^*)$ such that $\|y^*\| < \eta$. It follows that

$$(y^*,z^*) \in N_c(\mathrm{Gr}(M),(\bar{y},\bar{x})) = N_c(\mathrm{Gr}(D_cM(\bar{y},\bar{x})),(0,0)).$$

Since $\mathrm{Gr}(D_cM(\bar{y},\bar{x}))$ is convex,

$$\langle y^*, y-0\rangle + \langle z^*, x-0\rangle \le 0 \quad \forall (y,x) \in \mathrm{Gr}(D_cM(\bar{y},\bar{x})).$$

Noting that $\langle z^*, z\rangle = 0$ (because $z^* \in N_c(T_c(M(\bar{y}),\bar{x}),z)$ and $T_c(M(\bar{y}),\bar{x})$ is a convex cone), it follows from (4.19) that

$$\gamma d(x, T_c(M(\bar{y}),\bar{x})) \le \gamma\|x-z\| \le -\langle y^*, y\rangle \le \eta\|y\| \quad \forall (y,x) \in \mathrm{Gr}(D_cM(\bar{y},\bar{x})).$$

Letting $\gamma \to 1$, one has

$$d(x, T_c(M(\bar{y}),\bar{x})) \le \eta\|y\| \quad \forall (y,x) \in \mathrm{Gr}(D_cM(\bar{y},\bar{x})).$$

Hence, $\eta_* \le \eta < +\infty$. The proof is complete.

Remark 4.3. If one drops the condition $\|D_c^*M(\bar{y},\bar{x})|_{-N_c(M(\bar{y}),\bar{x})}\|^- < \infty$, it is possible that $T_c(M(\bar{y}),\bar{x}) \neq D_cM(\bar{y},\bar{x})(0)$. For example, let $M : R \rightrightarrows R$ be defined by

$$\mathrm{Gr}(M) := \{(y,x) \in R \times R :\ y^2 + x^2 \leq 1\}.$$

Thus, $M(1) = \{0\}$ and so $T_c(M(1),0) = \{0\}$; but $T_c(\mathrm{Gr}(M),(1,0)) = R_+ \times R$. Hence $D_cM(1,0)(0) = R$. This shows that $T_c(M(1),0) \neq D_cF(1,0)(0)$.

In Proposition 4.12, the assumption (4.18) is a mild one. Indeed, the corresponding assertion for contingent derivative always holds: $T(M(\bar{y}),\bar{x}) \subset DM(\bar{y},\bar{x})(0)$ (which is easy to verify). Thus, (4.18) is satisfied if $\mathrm{Gr}(M)$ is regular at $(\bar{y},\bar{x})$ in the Clarke sense. Hence, (4.18) is satisfied if M (resp., M^{-1}) is subsmooth at $(\bar{y},\bar{x})$ (resp., $(\bar{x},\bar{y})$); in particular, (4.18) is satisfied under the assumption of Theorem 4.11. Hence, the following corollary is immediate from Corollary 3.7 and Proposition 4.12.

COROLLARY 4.13. *Let G, g, M, and $(\bar{y},\bar{x})$ be as in Theorem* 4.11. *Then*

$$\|D^*M(\bar{y},\bar{x})|_{-N(M(\bar{y}),\bar{x})}\|^- = \inf\{\eta > 0 :\ d(x, T(M(\bar{y}),\bar{x})) \leq \eta\|y\| \ \ \forall y \in Y \text{ and } x \in DM(\bar{y},\bar{x})(y))\}.$$

If, in addition, $\|D^*M(\bar{y},\bar{x})_{-N(M(\bar{y}),\bar{x})}\|^- < +\infty$, then

$$\eta(DM(\bar{y},\bar{x});0,0) = \|D^*M(\bar{y},\bar{x})|_{-N(M(\bar{y},\bar{x})}\|^-.$$

In what follows, we consider the multifunction $M : Y \rightrightarrows X$ defined by

$$M(y) := \{x \in X :\ g(x) + y \in \Lambda\} \quad \forall y \in Y, \tag{4.20}$$

where $g : X \to Y$ is a function and Λ is a closed subset of Y.

THEOREM 4.14. *Let M be given by* (4.20) *and $(\bar{y},\bar{x}) \in \mathrm{Gr}(M)$. Suppose that g is smooth and that Λ is subsmooth at $g(\bar{x}) + \bar{y}$. Further suppose that there exist $\eta, \delta \in (0,\ +\infty)$ such that*

$$N_c(M(\bar{y}),u) \cap B_{X^*} \subset \eta(g'(u))^*(N_c(\Lambda, g(u) + \bar{y}) \cap B_{Y^*}) \quad \forall u \in M(\bar{y}) \cap B(\bar{x},\delta).$$

Then M is calm at $(\bar{y},\bar{x})$, and, more precisely, for any $\varepsilon > 0$ there exists $\delta_\varepsilon > 0$ such that

$$d(x, M(\bar{y})) \leq \frac{\eta + (1+\eta)\varepsilon}{1 - (1+\eta)\varepsilon}\|y - \bar{y}\| \quad \forall y \in B(\bar{y},\delta_\varepsilon) \text{ and } \forall x \in M(y) \cap B(\bar{x},\delta_\varepsilon).$$

Proof. Note that $M^{-1}(x) = -g(x) + \Lambda$ for each $x \in X$. It follows from Proposition 3.3(ii) (applied to $M^{-1}, -g$ in place of F, g) that

$$D_c^*M^{-1}(u,\bar{y})(B_{Y^*}) = (g'(u))^*(N_c(\Lambda, g(u) + \bar{y}) \cap B_{Y^*}) \quad \forall u \in M(\bar{y}).$$

Similarly, since Λ is subsmooth at $g(\bar{x}) + \bar{y}$, Proposition 3.3(iii) implies that M^{-1} is subsmooth at $(\bar{x},\bar{y})$ and so is M at $(\bar{y},\bar{x})$. Thus, the assertion of Theorem 4.14 follows from Theorem 4.4. The proof is complete.

Let T be a compact topological space and let $\mathcal{C}(T)$ denote the Banach space of all continuous functions on T equipped with the sup-norm. Let $\psi : X \times T \to R$ be a function and consider the multifunction $M : \mathcal{C}(T) \rightrightarrows X$ defined by

$$M(y) := \{x \in X :\ \psi(x,t) \leq -y(t) \ \ \forall t \in T\} \quad \forall y \in \mathcal{C}(T); \tag{4.21}$$

equivalently one can write (4.21) as

$$M(y) = \{x \in X : \ g(x) + y \in \Lambda\} \quad \forall y \in \mathcal{C}(T),$$

where $g(x) = \psi(x, \cdot)$ and Λ is the convex cone of all nonpositive continuous functions on T. In the special case when $X = R^n$, $T \subset R^m$, and ψ is a continuously differentiable function on $R^n \times R^m$ such that $\psi_1'(x,t)$ is locally Lipschitzian on $R^n \times R^m$, where $\psi_1'(x,t)$ denotes the derivative of $\psi(x,t)$ with respect the first variable x, Henrion and Outrata [10] recently considered the calmness of M defined by (4.21) at $(0, \bar{x}) \in \mathrm{Gr}(M)$. For $x \in M(0)$, let $T(x) := \{t \in T : \ \psi(x,s) \le \psi(x,t) \ \text{ for all } s \in T\}$ and let

$$\mathcal{J} := \{S \in \mathcal{K}(T) : \ \exists x_i \overset{\mathrm{bd}M(0)\setminus\{\bar{x}\}}{\longrightarrow} \bar{x} \ \text{ s.t. } d_H(S, T(x_i)) \to 0\},$$

where $\mathcal{K}(T)$ denotes the family of all compact subsets of T and d_H denotes the Hausdorff distance between compact sets. Henrion and Outrata established the following sufficient condition for calmness (see [10, Theorem 4]).

THEOREM A. *Consider* (4.21) *with* $X = R^n$, $T \subset R^m$, *and* ψ *being a smooth function on* $R^n \times R^m$ *such that* $\psi_1'(x,t)$ *is locally Lipschitzian on* $R^n \times R^m$. *Let* $\bar{x} \in M(0)$ *with* $\psi(\bar{x}, \bar{t}) = 0$ *for some* $\bar{t} \in T$. *Suppose that the following two conditions are satisfied.*

(1) $T(M(0), \bar{x}) = \{h \in R^n : \ \langle \psi_1'(\bar{x}, t), h\rangle \le 0$ *for all* $t \in T(\bar{x})\}$.

(2) *There exists* $\rho > 0$ *such that* $d(0, \mathrm{co}\{\psi_1'(\bar{x}, t) : \ t \in S\}) \ge \rho$ *for all* $S \in \mathcal{J}$.

Then M *is calm at* $(0, \bar{x})$.

Recently, Zheng and Yang [45] proved that the conditions (1) and (2) in Theorem A can be replaced by the following weaker condition: there exist $\eta, \delta \in (0, \ +\infty)$ such that

$$\text{(WC)} \quad N_c(M(0), u) \cap B_{X^*} \subset [0, \ \eta]\mathrm{co}\{\psi_1'(u,t) : \ t \in T(u)\} \quad \forall u \in M(0) \cap B(\bar{x}, \delta).$$

As an application of Theorem 4.14, we can improve and generalize Theorem A to the general Banach space case. To do this, it would be convenient to recall some standard notation. Let $\mathcal{B}(T)$ denote the family of all Borel sets in T and let $\mathrm{rca}(T)$ denote the space of all regular finite real-valued Borel measures on T equipped with the total variation norm $\|\mu\| = |\mu|(T)$ for any $\mu \in \mathrm{rca}(T)$. Recall that a Borel measure μ on T is said to be supported on $A \in \mathcal{B}(T)$ if $\mu(B) = 0$ for all $B \in \mathcal{B}(T)$ with $B \cap A = \emptyset$. Let

$$\mathrm{rca}^+(T) := \{\mu \in \mathrm{rca}(T) : \ \mu(B) \ge 0 \ \ \forall B \in \mathcal{B}(T)\}$$

and

$$\mathrm{rac}_A^+(T) := \{\mu \in \mathrm{rac}^+(T) : \ \mu \text{ is supported on } A\},$$

where $\mathrm{rac}_A^+(T)$ is interpreted as $\{0\}$ if $A = \emptyset$. It is well known, as the Riesz representation theorem, that $C(T)^* = \mathrm{rca}(T)$ and that

$$\mu \in \mathrm{rca}(T) \text{ and } \int_T y(t)d\mu \ge 0 \ \ \forall y \in C^+(T) \Longrightarrow \mu \in \mathrm{rca}^+(T),$$

where $C^+(T)$ denotes the set of all nonnegative continuous functions on T. For $y \in \mathcal{C}(T)$, let $I(y) = \{t \in T : \ y(t) = 0\}$.

PROPOSITION 4.15. *Let X be a general Banach space, T a compact topological space, $\psi(x,t)$ a continuous function on $X \times T$ such that $\psi_1'(x,t)$ is continuous on $X \times T$, and $g : X \to \mathcal{C}(T)$ defined by $g(x) := \psi(x,\cdot)$ for all $x \in X$. Let $M : \mathcal{C}(T) \rightrightarrows X$ be defined by* (4.21) *and let $\bar{x} \in M(0)$. Suppose that there exist $\eta, \delta \in (0, +\infty)$ such that for all $u \in M(0) \cap B(\bar{x}, \delta)$,*

(4.22)

$$N_c(M(0),u) \cap B_{X^*} \subset [0,\ \eta] \left\{ \int_T \psi_1'(u,t)d\mu :\ \mu \in \mathrm{rac}^+_{I(g(u))}(T) \text{ and } \mu(T) \le 1 \right\}.$$

Then M is calm at $(0, \bar{x})$.

Proof. By the assumption on ψ, it is easy to verify that $g'(x) = \psi_1'(x,\cdot)$ for all $x \in X$, and so

$$(g'(x))^*(\mu) = \int_T \psi_1'(x,t)d\mu \quad \forall \mu \in \mathrm{rac}(T) = \mathcal{C}(T)^*.$$

By Theorem 4.14 (applied to $Y = \mathcal{C}(T)$, $\Lambda = -\mathcal{C}^+(T)$, $g(x) = \psi(x,\cdot)$ for all $x \in X$ and $\bar{y} = 0$), we need only show that

$$N(-\mathcal{C}^+(T), -y) = \mathrm{rac}^+_{I(y)}(T) \quad \forall y \in \mathcal{C}^+(T). \tag{4.23}$$

Let $y \in \mathcal{C}^+(T)$ and $\mu \in N(-\mathcal{C}^+(T), -y)$. Since $-\mathcal{C}^+(T)$ is a closed convex cone in $\mathcal{C}(T)$, the Riesz representation theorem implies that

$$\int_T -z(t)d\mu \le \int_T -y(t)d\mu = 0 \quad \forall z \in \mathcal{C}^+(T).$$

It follows that $\mu \in \mathrm{rac}^+(T)$ and $\int_T y(t)d\mu = 0$. This shows that $\mu \in \mathrm{rac}^+_{I(y)}(T)$. Hence, $N(-\mathcal{C}(T), -y) \subset \mathrm{rac}^+_{I(y)}(T)$. Since the reverse inclusion is clear, (4.23) holds. The proof is complete.

Remark 4.4. Note that in the special case when $X = R^n$ and under the assumption of Proposition 4.15,

$$\mathrm{co}\{\psi_1'(u,t) :\ t \in I(g(u))\} = \mathrm{cl}(\mathrm{co}\{\psi_1'(u,t) :\ t \in I(g(u))\})$$

and so

$$\mathrm{co}\{\psi_1'(u,t) :\ t \in I(g(u))\} = \left\{ \int_T \psi_1'(u,t)d\mu :\ \mu \in \mathrm{rac}^+_{I(g(u))}(T) \text{ and } \mu(T) \le 1 \right\}.$$

Thus, (4.22) and (WC) are the same. Hence Proposition 4.15 improves and generalizes Theorem A by Henrion and Outrata. Moreover, Proposition 4.15 does not require that $\psi_1'(x,t)$ is locally Lipschitzian.

5. Application to error bounds for inequality systems. Let $f : X \to R \cup \{+\infty\}$ be a proper lower semicontinuous function and consider the following inequality system:

$$f(x) \le 0. \tag{5.1}$$

Let $f_1, \ldots, f_n : X \to R \cup \{+\infty\}$ be proper lower semicontinuous functions and $f(x) = \max\{f_i(x) :\ i = 1, \ldots, n\}$. Then (5.1) is the following system of finitely many inequalities:

$$f_i(x) \le 0,\ \ i = 1, \ldots, n. \tag{5.2}$$

Recall that inequality (5.1) has a local error bound (or metric regularity) at $\bar{x}$ if there exists $\tau > 0$ such that

$$d(x, S) \le \tau [f(x)]_+ \quad \forall x \text{ close to } \bar{x}, \tag{5.3}$$

where S is the solution set of (5.1) and $[f(x)]_+ = \max\{0, f(x)\}$.

In the case when f (resp., f_i) is convex, many authors studied the error bound issues for (5.1) (resp., (5.2)) (see [12, 13, 18, 19, 21, 38, 39, 41] and the references therein). In particular, in the case when f is convex, it is known (cf. [12, 13, 41]) that (5.1) has a local error bound at a point a of the solution set S if and only if there exist $\tau, \delta \in (0, +\infty)$ such that

$$N(S, z) \cap B_{X^*} \subset [0, \tau] \partial f(z) \quad \forall\, z \in \mathrm{bd}(S) \cap B(a, \delta).$$

Under the condition that X is finite-dimensional and each f_i is convex and smooth, Li [19] proved that inequality system (5.2) has a local error bound at $a \in S$ if and only if

$$N(S, z) = R_+ \mathrm{co}\{f_i'(z) : \ i \in I(z)\} \ \ \forall z \in \mathrm{bd}(S) \text{ close to } a,$$

where $I(z) := \{1 \le i \le n : \ f_i(z) = 0\}$.

As applications of the main results obtained in section 4, we consider local error bounds for (5.1) and (5.2) when f and f_i are not necessarily convex. For the sake of simplicity in presentation, let us assume, in the remainder of this section, that $f : X \to R$ is a local Lipschitz (not necessarily convex) function.

As an extension of the convexity, Ngai, Luc, and Thera [29] introduced the approximate convexity. Recall that a function $f : X \to R$ is said to be approximately convex at $a \in X$ if for any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$f(tx_1 + (1-t)x_2) \le tf(x_1) + (1-t)f(x_2) + \varepsilon t(1-t)\|x_1 - x_2\|$$

for all $x_1, x_2 \in B(a, \delta)$ and $t \in (0, 1)$. Recently, Aussel, Daniilidis, and Thibault [1] proved that a local Lipschitz function $f : X \to R$ is approximately convex at a if and only if for any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$(*) \qquad f(x) - f(u) - \langle u^*, x - u\rangle \ge -\varepsilon\|x - u\| \quad \forall x, u \in B(a, \delta) \text{ and } \forall u^* \in \partial_c f(u).$$

Slightly weakened conditions can be introduced as follows: f is said to be L-subsmooth (resp., weak L-subsmooth) at $a \in X$ if for any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$f(x) - f(u) - \langle u^*, x - u\rangle \ge -\varepsilon\|x - u\|$$

whenever $x \in B(u, \delta)$ and $u \in B(a, \delta)$ with $f(u) = f(a)$, $u^* \in \partial_c f(u)$ (resp., $u^* \in \partial f(u)$).

Let $M : R \rightrightarrows X$ be defined by

$$M(y) := \{x \in X : \ f(x) \le y\} \quad \forall y \in R.$$

Then $\mathrm{Gr}(M^{-1}) = \mathrm{epi}(f)$. Hence,

$$\mathrm{dom}(D_c^* M^{-1}(x, f(x))) = [0, +\infty) \ \text{ and } \ D_c^* M^{-1}(x, f(x))(r) = r\partial_c f(x) \tag{5.4}$$

for all $x \in X$ and $r \in [0, +\infty)$. Note that $N_c(\mathrm{Gr}(M), (t, x)) = \{(0, 0)\}$ for any $x \in X$ and $t > f(x)$ (because $(x, t) \in \mathrm{int}(\mathrm{Gr}(M^{-1}))$). By the local Lipschitz property of

f, it is easy to verify that f is (weak) L-subsmooth at a if and only if M is (weak) L-subsmooth at $(f(a), a)$. Note that $M(0) = S$, and (5.1) has a local error bound at $a \in S$ if and only if M is calm at $(0, a)$. Thus, the following result is immediate from Theorems 4.2 and 4.4.

THEOREM 5.1. *The following assertions hold.*

(i) *If inequality* (5.1) *has a local error bound at* $a \in S$*, then there exist* $\tau, \delta \in (0, +\infty)$ *such that*

$$\hat{N}(S, x) \cap B_{X^*} \subset [0, \tau]\partial_c f(x) \quad \forall x \in S \cap B(a, \delta).$$

(ii) *If* f *is L-subsmooth at* $a \in S$ *and there exist* $\tau, \delta \in (0, +\infty)$ *such that*

$$N_c(S, x) \cap B_{X^*} \subset [0, \tau]\partial_c f(x) \quad \forall x \in \mathrm{bd}(S) \cap B(a, \delta), \tag{5.5}$$

then (5.1) *has a local error bound at* a.

By the same argument but using Theorems 4.3 and 4.8 (in place of Theorems 4.2 and 4.4), we have the following characterization of a local error bound for inequality (5.1) when X is an Asplund space.

THEOREM 5.2. *Suppose that* X *is an Asplund space and that* f *is weakly L-subsmooth at* $a \in S$*. Then inequality* (5.1) *has a local error bound at* $a \in S$ *if and only if there exist* $\tau, \delta \in (0, +\infty)$ *such that*

$$\hat{N}(S, x) \cap B_{X^*} \subset [0, \tau]\partial f(x) \quad \forall x \in \mathrm{bd}(S) \cap B(a, \delta).$$

The next two theorems (Theorems 5.3 and 5.6) concern convex-composite functions.

THEOREM 5.3. *Let* $\phi : Z \to R$ *be a continuous convex function and* $g : X \to Z$ *be a smooth function. Let* $f(x) = \phi(g(x))$ *for all* $x \in X$*. Let* $a \in S$ *and suppose that* $g'(a)$ *is surjective. Then* (5.1) *has a local error bound at* a *if and only if there exist* $\tau, \delta \in (0, +\infty)$ *such that* (5.5) *holds.*

Proof. Let $G : R \rightrightarrows Z$ and $M : R \rightrightarrows X$ be defined by

$$M(y) := \{x \in X : \ f(x) \le y\} \ \text{ and } \ G(y) := \{z \in Z : \ \phi(z) \le y\} \quad \forall y \in R.$$

Then $M(y) = g^{-1}(G(y))$ for all $y \in R$. It follows from Theorem 4.11 and (5.4) that M is calm at $(0, a)$ if and only if that there exist $\tau, \delta \in (0, +\infty)$ such that (5.5) holds. Since M is calm at $(0, a)$ if and only if (5.1) has a local error bound, the proof is complete.

Theorems 5.1–5.3 can be regarded as generalizations of the main result in [41] from the convex case to the nonconvex case. Next we consider local error bounds for inequality system (5.2).

PROPOSITION 5.4. *Let* $f_1, \ldots, f_n : X \to R$ *be smooth (not necessarily convex) functions. Let* $a \in S := \{x \in X : \ f_i(x) \le 0, \ i = 1, \ldots, n\}$ *and suppose that there exists* $\tau \in (0, +\infty)$ *such that*

$$N_c(S, z) \cap B_{X^*} \subset [0, \tau]\mathrm{co}(\{f_i'(z) : \ i \in I(z)\}) \ \forall \ z \in \mathrm{bd}(S) \text{ close to } a.$$

Then (5.2) *has a local error bound at* a.

Proof. Let $f(x) = \max\{f_i(x) : \ i = 1, \ldots, n\}$ for all $x \in X$. Then, by [4, Proposition 2.3.12], one has

$$\partial_c f(u) = \mathrm{co}(\{f_i'(u) : \ i \in I(u)\}) \quad \forall u \in X.$$

By (ii) of Theorem 5.1, we need only show that f is L-subsmooth at a. Since each f_i is smooth on X, for any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$f_i(x_1) - f_i(x_2) - \langle f_i'(x_2), x_1 - x_2\rangle \geq -\varepsilon\|x_1 - x_2\| \quad \forall x_1, x_2 \in B(a,\delta). \tag{5.6}$$

Let $x, u \in B(a,\delta)$ and $u^* \in \partial_c f(u)$. Then there exist $t_i \geq 0$ $(i \in I(u))$ such that $\sum_{i\in I(u)} t_i = 1$ and $u^* = \sum_{i\in I(u)} t_i f_i'(u)$. Hence, it follows from (5.6) that

$$\begin{aligned} f(x) - f(u) - \langle u^*, x - u\rangle &= \sum_{i\in I(u)} t_i(f(x) - f_i(u) - \langle f_i'(u), x - u\rangle) \\ &\geq -\varepsilon\|x - u\|. \end{aligned}$$

This shows that f is approximately convex (and thus L-subsmooth) at a. The proof is complete.

Now we extend Li's result on local error bounds (i.e., the metric regularity) for a system of smooth and convex inequalities to the nonconvex case. First we prove a lemma.

LEMMA 5.5. *Let $f_1, \ldots, f_n : X \to R$ be smooth functions. Let $a \in \mathrm{bd}(S)$ be such that for any $J \subset I(a)$,*

$$0 \in \mathrm{co}\{f_i'(a) :\ i \in J\} \Rightarrow a \text{ is a local minimizer of } \max\{f_i(x) :\ i \in J\}. \tag{5.7}$$

Then there exists $\tau \in (0,\ +\infty)$ such that

$$N_c(S,z) \cap B_{X^*} \subset [0,\ \tau]\mathrm{co}\{f_i'(z) :\ i \in I(z)\}\ \ \forall z \in \mathrm{bd}(S) \text{ close to } a \tag{5.8}$$

if and only if

$$N_c(S,z) = R_+\mathrm{co}(\{f_i'(z) :\ i \in I(z)\})\ \ \forall z \in \mathrm{bd}(S) \text{ close to } a. \tag{5.9}$$

The corresponding result also holds if $N_c(S,z)$ is replaced with $\hat{N}(S,z)$ in (5.8) *and* (5.9).

Proof. We prove only the first assertion (the proof for the last assertion is similar). Since each f_i is smooth, it is easy to verify that

$$R_+\mathrm{co}(\{f_i'(z) :\ i \in I(z)\}) \subset N_c(S,z) \quad \forall z \in \mathrm{bd}(S).$$

We need only show that (5.9) implies that there exists $\tau \in (0,\ +\infty)$ such that (5.8) holds. To do this, suppose to the contrary that there exist a sequence $\{z_k\}$ in $\mathrm{bd}(S)$ and a sequence $\{z_k^*\}$ in X^* such that

$$z_k \to a \text{ and } z_k^* \in N_c(S,z_k) \cap B_{X^*} \setminus [0,\ k]\mathrm{co}\{f_i'(z_k) :\ i \in I(z_k)\}. \tag{5.10}$$

By the continuity of f_i and by considering k large if necessary, we assume without loss of generality that $I(z_k) \subset I(a)$ for each k. Similarly, by (5.9), we may assume that for each k there exist $t_k(i) > 0$ $(i \in I(z_k))$ such that $z_k^* = \sum_{i\in I(z_k)} t_k(i) f_i'(z_k)$. It follows from the Carathéodory theorem (cf. [3, p. 25]) that there exist $J_k \subset I(z_k)$ and $r_k(i) > 0$ such that $\{f_i'(z_k) :\ i \in J_k\}$ is linearly independent and $z_k^* = \sum_{i\in J_k} r_k(i) f_i'(z_k)$. This and (5.10) imply that $\sum_{i\in J_k} r_k(i) > k$. Noting that $J_k \subset \{1, \ldots, n\}$, without loss of generality we can assume that $J_k = J$ for each k and

$$\frac{r_k(i)}{\sum_{j\in J} r_k(j)} \to r_i, \text{ as } k \to \infty \text{ and } i \in J$$

(passing to a subsequence if necessary). Then $J \subset I(z_k) \subset I(a)$ and $\sum_{i\in J} r_i = 1$. Since $z_k^* \in B_{X^*}$, it follows that

$$0 = \lim_{k\to\infty} \frac{z_k^*}{\sum_{j\in J} r_k(j)} = \lim_{k\to\infty} \frac{\sum_{i\in J} r_k(i) f_i'(z_k)}{\sum_{j\in J} r_k(j)} = \sum_{i\in J} r_i f_i'(a) \in \mathrm{co}\{f_i'(a) :\ i \in J\}.$$

This and (5.7) imply that a is a local minimizer of f_J defined by $f_J(x) := \max\{f_i(x) : i \in J\}$. Thus there exists an open neighborhood U of a such that $f_J(x) \geq f_J(a)$ for all $x \in U$. Noting that $f_J(a) = f_J(z_k)$ (as $I(z_k) \subset I(a)$), it follows from (5.10) that z_k is also a local minimizer of f_J for each k large enough. Thus, for each k large enough, $0 \in \mathrm{co}\{f_i'(z_k) :\ i \in J\}$, contradicting the fact that $\{f_i'(z_k) :\ i \in J\}$ is linearly independent.

The following theorem clearly improves and extends Li's result (from the convex and finite-dimensional case to the nonconvex and infinite-dimensional case).

THEOREM 5.6. *Let $f_i(x) = \phi_i(g(x))$ for all $x \in X$ $(i = 1, \ldots, n)$, where $g : X \to Z$ is a smooth mapping, $\phi_i : Z \to R$ is a smooth convex function, and Z is another Banach space. Let $a \in \mathrm{bd}(S)$. Suppose that $g'(a)$ is surjective. Then* (5.2) *has a local error bound at a if and only if*

$$N_c(S, z) = R_+\mathrm{co}(\{f_i'(z) :\ i \in I(z)\}) \ \text{ for } z \in \mathrm{bd}(S) \text{ close to } a.$$

Proof. Note that (5.2) has a local error bound at a if and only if (5.1) also does with $f(x) = \max\{f_i(x) :\ i = 1, \ldots, n\}$ for all $x \in X$, and also note that

$$\partial_c f(x) = \mathrm{co}(\{f_i'(x) :\ i \in I(x)\}) \quad \forall x \in X.$$

By Theorem 5.3 and Lemma 5.5, we need only show that (5.7) holds. To do this, let $J \subset I(a)$ and $0 \in \mathrm{co}\{f_i'(a) :\ i \in J\}$. Then there exist $\lambda_i \geq 0$ with $\sum_{i\in J} \lambda_i = 1$ such that

$$0 = \sum_{i\in J} \lambda_i f_i'(a) = \sum_{i\in J} \lambda_i [g'(a)]^*(\phi_i'(g(a))) = [g'(a)]^* \left(\sum_{i\in J} \lambda_i \phi_i'(g(a)) \right).$$

Noting that $[g'(a)]^*$ is injective (because $g'(a)$ is surjective), it follows that

$$0 = \sum_{i\in J} \lambda_i \phi_i'(g(a)). \tag{5.11}$$

Let $\phi(u) := \max\{\phi_i(u) :\ i \in J\}$ for all $u \in Z$. Then ϕ is a continuous convex function and, by $J \subset I(a)$, $\phi(g(a)) = \phi_i(g(a))$ for all $i \in J$. This and (5.11) imply that $g(a)$ is a global minimizer of ϕ. It follows that a is a global minimizer of $\max\{f_i(x) :\ i \in J\}$. This shows that (5.7) holds.

Acknowledgment. The authors wish to thank the referees for careful reading of the paper and for many valuable comments, which helped to improve our presentation.

REFERENCES

[1] D. AUSSEL, A. DANIILIDIS, AND L. THIBAULT, *Subsmooth sets: Functional characterizations and related concepts*, Trans. Amer. Math. Soc., 357 (2005), pp. 1275–1301.

[2] M. BERGER, *Nonlinearity and Functional Analysis*, Academic Press, New York, 1977.

[3] J. M. BORWEIN AND A. S. LEWIS, *Convex Analysis and Nonlinear Optimization*, Springer-Verlag, New York, 2000.

[4] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.
[5] F. H. Clarke, R. Stern, and P. Wolenski, *Proximal smoothness and the lower-C^2 property*, J. Convex Anal., 2 (1995), pp. 117–144.
[6] A. L. Dontchev and R. T. Rockafellar, *Regularity and conditioning of solution mappings in variational analysis*, Set-Valued Anal., 12 (2004), pp. 79–109.
[7] A. L. Dontchev, A. S. Lewis, and R. T. Rockafellar, *The radius of metric regularity*, Trans. Amer. Math. Soc., 355 (2003), pp. 493–517.
[8] R. Henrion and A. Jourani, *Subdifferential conditions for calmness of convex constraints*, SIAM J. Optim., 13 (2002), pp. 520–534.
[9] R. Henrion and J. Outrata, *A subdifferential condition for calmness of multifunctions*, J. Math. Anal. Appl., 258 (2001), pp. 110–130.
[10] R. Henrion and J. Outrata, *Calmness of constraint systems with applications*, Math. Program., 104 (2005), pp. 437–464.
[11] R. Henrion, A. Jourani, and J. Outrata, *On the calmness of a class of multifunctions*, SIAM J. Optim., 13 (2002), pp. 603–618.
[12] H. Hu, *Characterizations of the strong basic constraint qualification*, Math. Oper. Res., 30 (2005), pp. 956–965.
[13] H. Hu, *Characterizations of local and global error bounds for convex inequalities in Banach spaces*, SIAM J. Optim., 18 (2007), pp. 309–321.
[14] A. D. Ioffe, *Metric regularity and subdifferential calculus*, Russian Math. Surveys, 55 (2000), pp. 501–558.
[15] G. Jameson, *Ordered Linear Spaces*, Springer-Verlag, Berlin, 1970.
[16] V. Jeyakumar, D. T. Luc, and P. N. Tinh, *Convex composite non-Lipschitz programming*, Math. Program., 92 (2002), pp. 177–195.
[17] D. Klatte and B. Kummer, *Nonsmooth Equations in Optimization: Regularity Calculus, Methods and Applications*, Kluwer Academic, Dordrecht, The Netherlands, 2002.
[18] A. Lewis and J. S. Pang, *Error bounds for convex inequality systems*, in Generalized Convexity, Generalized Monotonicity: Recent Results, Proceedings of the Fifth Symposium on Generalized Convexity (Luminy, 1996), J.-P. Crouzeix, J.-E. Martinez-Legaz, and M. Volle, eds., Kluwer Academic, Dordrecht, The Netherlands, 1997, pp. 75–100.
[19] W. Li, *Abadie's constraint qualification, metric regularity, and error bounds for differentiable convex inequalities*, SIAM J. Optim., 7 (1997), pp. 966–978.
[20] W. Li and I. Singer, *Global error bounds for convex multifunctions and applications*, Math. Oper. Res., 23 (1998), pp. 443–462.
[21] W. Li, C. Nahak, and I. Singer, *Constraint qualifications for semi-infinite systems of convex inequalities*, SIAM J. Optim., 11 (2000), pp. 31–52.
[22] R. E. Megginson, *An Introduction to Banach Space Theory*, Springer-Verlag, New York, 1998.
[23] B. S. Mordukhovich, *Complete characterization of openness, metric regularity, and Lipschitzian properties of multifunctions*, Trans. Amer. Math. Soc., 340 (1994), pp. 1–35.
[24] B. S. Mordukhovich, *Variational Analysis and Generalized Differentiation.* I. *Basic Theory*, Springer-Verlag, Berlin, Heidelberg, 2006.
[25] B. S. Mordukhovich, *Variational Analysis and Generalized Differentiation.* II. *Applications*, Springer-Verlag, Berlin, Heidelberg, 2006.
[26] B. S. Mordukhovich and Y. Shao, *Nonsmooth sequential analysis in Asplund spaces*, Trans. Amer. Math. Soc., 348 (1996), pp. 1235–1280.
[27] K. F. Ng, *An open mapping theorem*, Proc. Cambridge Philos. Soc., 74 (1973), pp. 61–66.
[28] K. F. Ng and X. Y. Zheng, *Characterizations of error bounds for convex multifunctions on Banach spaces*, Math. Oper. Res., 29 (2004), pp. 45–63.
[29] H. V. Ngai, D. T. Luc, and M. Thera, *Approximate convex functions*, J. Nonlinear Convex Anal., 1 (2000), pp. 155–176.
[30] J. P. Penot, *Optimality conditions in mathematical programming and composite optimization*, Math. Program., 67 (1994), pp. 225–245.
[31] R. R. Phelps, *Convex Functions, Monotone Operators and Differentiability*, Lecture Notes in Math. 1364, Springer-Verlag, New York, 1989.
[32] R. Poliquin, R. T. Rockafellar, and L. Thibault, *Local differentiability of distance functions*, Trans. Amer. Math. Soc., 352 (2000), pp. 5231–5249.
[33] S. M. Robinson, *Regularity and stability for convex multivalued functions*, Math. Oper. Res., 1 (1976), pp. 130–143.
[34] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998.
[35] A. Shapiro, *On a class of nonsmooth composite functions*, Math. Oper. Res., 28 (2003), pp. 677–692.
[36] C. Ursescu, *Multifunctions with closed graph*, Czech. Math. J., 25 (1975), pp. 438–441.

[37] R. S. Womersley, *Local properties of algorithms for minimizing nonsmooth composite functions*, Math. Program., 32 (1985), pp. 69–89.
[38] C. Zalinescu, *Weak sharp minima, well-behaving functions and global error bounds for convex inequalities in Banach spaces*, in Proceedings of the 12th Baikal International Conference on Optimization Methods and Their Applications, Irkutsk, Russia, 2001, pp. 272–284.
[39] C. Zalinescu, *Convex Analysis in General Vector Spaces*, World Scientific, Singapore, 2002.
[40] C. Zalinescu, *A nonlinear extension of Hoffman's error bounds for linear inequalities*, Math. Oper. Res., 28 (2003), pp. 524–532.
[41] X. Y. Zheng and K. F. Ng, *Metric regularity and constraint qualifications for convex inequalities on Banach spaces*, SIAM J. Optim., 14 (2004), pp. 757–772.
[42] X. Y. Zheng and K. F. Ng, *Perturbation analysis of error bounds for systems of conic linear inequalities in Banach spaces*, SIAM J. Optim., 15 (2005), pp. 1026–1041.
[43] X. Y. Zheng and K. F. Ng, *Metric subregularity and constraint qualifications for convex generalized equations in Banach spaces*, SIAM J. Optim., 18 (2007), pp. 437–460.
[44] X. Y. Zheng and K. F. Ng, *Linear regularity for a collection of subsmooth sets in Banach spaces*, SIAM J. Optim., 19 (2008), pp. 62–76.
[45] X. Y. Zheng and X. Q. Yang, *Weak sharp minima for semi-infinite optimization problems with applications*, SIAM J. Optim., 18 (2007), pp. 573–588.
[46] X. Y. Zheng, X. M. Yang, and K. L. Teo, *Super efficiency of vector optimization in Banach spaces*, J. Math. Anal. Appl., 327 (2007), pp. 453–460.

SIAM J. OPTIM.
Vol. 19, No. 4, pp. 1674–1693

ADAPTIVE BARRIER UPDATE STRATEGIES FOR NONLINEAR INTERIOR METHODS*

JORGE NOCEDAL†, ANDREAS WÄCHTER‡, AND RICHARD A. WALTZ†

Abstract. This paper considers strategies for selecting the barrier parameter at every iteration of an interior-point method for nonlinear programming. Numerical experiments suggest that heuristic adaptive choices, such as Mehrotra's probing procedure, outperform monotone strategies that hold the barrier parameter fixed until a barrier optimality test is satisfied. A new adaptive strategy is proposed based on the minimization of a *quality function*. The paper also proposes a globalization framework that ensures the convergence of adaptive interior methods, and examines convergence failures of the Mehrotra predictor-corrector algorithm. The barrier update strategies proposed in this paper are applicable to a wide class of interior methods and are tested in the two distinct algorithmic frameworks provided by the IPOPT and KNITRO software packages.

Key words. interior-point methods, barrier methods, nonlinear programming, constrained optimization

AMS subject classifications. 49M37, 65K05, 90C06, 90C30, 90C51

DOI. 10.1137/060649513

1. Introduction.

In this paper we describe interior methods for nonlinear programming that update the barrier parameter adaptively, as the iteration progresses. The goal is to design algorithms that are both efficient in practice and that enjoy global convergence guarantees. The adaptive strategies studied in this paper allow the barrier parameter to increase or decrease at every iteration and provide an alternative to the so-called Fiacco–McCormick approach that fixes the barrier parameter until an approximate solution of the barrier problem is computed. Our motivation for this work stems from our belief that robust interior methods for nonlinear programming must be able to react swiftly to changes of scale in the problem and to correct overly aggressive decreases in the barrier parameter.

Adaptive barrier update strategies are well established in interior methods for linear and convex quadratic programming. The most popular approach of this type is Mehrotra's predictor-corrector (MPC) method [22]. It computes, at every iteration, a probing (affine scaling) step that determines a target value of the barrier parameter, and then takes a primal-dual step using this target value. A corrector step is added to better follow the trajectory of the central path to the solution. Mehrotra's method has proved to be very effective for linear and convex quadratic programming, but is not supported by global convergence guarantees. Indeed, as we show in section 7, its reliability is heavily dependent upon an appropriate choice of the starting point.

When solving nonlinear nonconvex programming problems, much caution must be exercised to prevent the iteration from failing. Nonminimizing stationary points can attract the iteration, and aggressive decreases in the barrier parameter can lead to failure. Our numerical experience shows that the direct extension of Mehrotra's

*Received by the editors January 9, 2006; accepted for publication (in revised form) August 12, 2008; published electronically January 28, 2009.
http://www.siam.org/journals/siopt/19-4/64951.html

†Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, 60208-3118 (nocedal@eecs.northwestern.edu, rwaltz@usc.edu). The work of these authors was supported by National Science Foundation grants CCR-0219438 and DMI-0422132, and Department of Energy grant DE-FG02-87ER25047-A004.

‡Department of Mathematical Sciences, IBM T.J. Watson Research Center, Yorktown Heights, NY 10598 (andreasw@watson.ibm.com).

predictor-corrector method to nonlinear programming does not result in a robust method. As we discuss below, the main source of instability is the corrector step. Further adaptive barrier update strategies designed specifically for nonlinear programming include [2, 11, 15, 23, 24, 25].

The global convergence properties of interior methods for nonlinear programming have recently received much attention [4, 8, 11, 17, 20, 23, 24, 26, 32]. Some of these studies focus on the effects of merit functions or filters, and on regularization techniques. With the exception of [11, 23, 24], however, these papers do not consider the numerical or theoretical properties of adaptive barrier update techniques.

The organization of this paper is as follows. After stating the basic nonlinear interior method in section 2, we start our exploration of adaptive barrier updates by examining several established techniques in section 3. In this initial investigation, we do not impose a rigorous globalization scheme on the methods, but simply compare their practical behavior on a standard test set. Motivated by the initial observations of these experiments, we then

- propose (in section 4) a new strategy for choosing the barrier parameter that, in contrast to previously proposed update rules, is not based on heuristics but follows a clear-cut objective, namely, the minimization of a "quality function";
- present two simple frameworks that ensure global convergence for interior methods that use *any* update rule for the barrier parameter (section 5);
- explore the numerical performance of the proposed strategy on standard test sets (section 6); and
- discuss the shortcomings of the Mehrotra corrector step (which can be observed even in the linear case) and propose a remedy (section 7).

To show the generality of our quality function approach, we implement it in the two different algorithmic contexts provided by the IPOPT [27] and KNITRO [6, 28] software packages.

Notation. For any vector z, we denote by Z the diagonal matrix whose diagonal entries are given by z. We let e denote the vector of ones, of appropriate dimension, that is, $e = (1, 1, \ldots, 1)^T$.

2. Primal-dual nonlinear interior methods. The problem under consideration will be written as

$$\min_x \ f(x), \tag{2.1a}$$

$$\text{s.t. } c(x) = 0, \tag{2.1b}$$

$$x \geq 0, \tag{2.1c}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $c : \mathbb{R}^n \to \mathbb{R}^m$ are twice continuously differentiable functions. For conciseness we will refer to interior-point methods for nonlinear programming as "nonlinear interior methods." A variety of these methods have been proposed in the last 10 years; they differ mainly in some aspects of the step computation and in the globalization scheme. Most of the nonlinear interior methods are related to the simple primal-dual iteration described next; our discussion of barrier parameter choices will be phrased in the context of this iteration.

We associate with the nonlinear program (2.1) the barrier problem

$$\min_x \ \varphi_\mu(x) \equiv f(x) - \mu \sum_{i=1}^{n} \ln x^{(i)}, \tag{2.2a}$$

$$\text{s.t. } c(x) = 0, \tag{2.2b}$$

where $\mu > 0$ is the barrier parameter. As is well known, the KKT conditions of the barrier problem (2.2) can be written as

$$\nabla f(x) - A(x)^T y - z = 0, \tag{2.3a}$$

$$Xz - \mu e = 0, \tag{2.3b}$$

$$c(x) = 0, \tag{2.3c}$$

where $A(x)$ denotes the Jacobian matrix of the constraint function $c(x)$. Condition (2.3b), the positivity of μ, and the requirement that the log function be well-defined in (2.2a) implicitly requires that

$$x > 0, \qquad z > 0. \tag{2.4}$$

Applying Newton's method to (2.3), in the variables (x, y, z), gives the *primal-dual* system

$$\begin{bmatrix} \nabla^2_{xx}\mathcal{L} & -A(x)^T & -I \\ Z & 0 & X \\ A(x) & 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta z \end{bmatrix} = - \begin{bmatrix} \nabla f(x) - A(x)^T y - z \\ Xz - \mu e \\ c(x) \end{bmatrix}, \tag{2.5}$$

where $\mathcal{L}$ denotes the Lagrangian of the nonlinear program, that is,

$$\mathcal{L}(x, y, z) = f(x) - y^T c(x) - z^T x. \tag{2.6}$$

After the step $\Delta = (\Delta x, \Delta y, \Delta z)$ has been determined, we compute primal and dual steplengths, α_p and α_d, and define the new iterate (x^+, y^+, z^+) as

$$x^+ = x + \alpha_p \Delta x, \quad y^+ = y + \alpha_d \Delta y, \quad z^+ = z + \alpha_d \Delta z. \tag{2.7}$$

The steplengths are computed in two stages. First we compute

$$\alpha_x^{\max} = \max\{\alpha \in (0, 1] : x + \alpha \Delta x \geq (1 - \tau)x\}, \tag{2.8a}$$

$$\alpha_z^{\max} = \max\{\alpha \in (0, 1] : z + \alpha \Delta z \geq (1 - \tau)z\}, \tag{2.8b}$$

with $\tau \in (0, 1)$ (e.g., $\tau = 0.995$). Next, we perform a backtracking line search to compute the final steplengths

$$\alpha_p \in (0, \alpha_x^{\max}], \qquad \alpha_d \in (0, \alpha_z^{\max}], \tag{2.9}$$

which provide sufficient decrease of a merit function or ensure acceptability by a filter.

The other major ingredient in this simple primal-dual iteration is the procedure for choosing the barrier parameter μ. Two types of barrier update strategies have been studied in the literature: adaptive and monotone. Adaptive strategies [11, 15, 24, 25] allow changes in the barrier parameter at every iteration, and are often efficient in practice, but as already mentioned, they generally do not enjoy global convergence properties. (The analyses presented in [11, 24] provide certain convergence results to stationary points, but these methods do not explicitly aim to decrease the objective function—they only enforce reduction of a measure of stationarity.)

The most important monotone strategy is the so-called Fiacco–McCormick approach that fixes the barrier parameter until an approximate solution of the barrier problem is computed. It has been employed in various nonlinear interior algorithms [3, 5, 14, 16, 27, 29, 31] and has been implemented, for example, in the IPOPT and

KNITRO software packages. The Fiacco–McCormick strategy provides a framework for establishing global convergence [4, 26], but suffers from important limitations. It can be very sensitive to the choice of the initial point, the initial value of the barrier parameter, and the scaling of the problem, and it is often unable to recover quickly when the iterates approach the boundary of the feasible region prematurely. The numerical experience with IPOPT and KNITRO reported below suggests that more dynamic update strategies are needed to improve the efficiency of nonlinear interior methods.

The algorithms considered in this paper guarantee only convergence to first-order stationary points; enforcing convergence to second-order points would require an estimation of the smallest eigenvalue of the reduced Hessian, which is too expensive in the large-scale case. However, the algorithms presented here generate steps that promote convergence to minimizers by ensuring descent properties for the barrier problem.

3. Choosing the barrier parameter. In this section we discuss two adaptive barrier strategies proposed in the literature and compare them numerically with the monotone Fiacco–McCormick approach. These numerical results motivate the techniques presented in the following sections.

Given an iterate (x, y, z), consider an interior method that computes primal-dual search directions by (2.5). The most common approach for choosing the barrier parameter μ is to make it proportional to the current complementarity value, that is,

$$\mu = \sigma \frac{x^T z}{n}, \tag{3.1}$$

where $\sigma > 0$ is a *centering parameter* and n denotes the number of variables. Mehrotra's predictor-corrector (MPC) method [22] for linear programming determines the value of σ using a preliminary step computation (an affine scaling step). We now describe a direct extension of Mehrotra's strategy to the nonlinear programming case.

First, we calculate an affine scaling step

$$\left(\Delta x^{\text{aff}}, \Delta y^{\text{aff}}, \Delta z^{\text{aff}}\right) \tag{3.2}$$

by setting $\mu = 0$ in (2.5), that is,

$$\begin{bmatrix} \nabla^2_{xx}\mathcal{L} & -A(x)^T & -I \\ Z & 0 & X \\ A(x) & 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta x^{\text{aff}} \\ \Delta y^{\text{aff}} \\ \Delta z^{\text{aff}} \end{bmatrix} = - \begin{bmatrix} \nabla f(x) - A(x)^T y - z \\ Xz \\ c(x) \end{bmatrix}. \tag{3.3}$$

We then compute α_x^{aff} and α_z^{aff} to be the largest steplengths in $(0, 1]$ that can be taken along the direction (3.2) before violating the nonnegativity conditions $(x, z) \geq 0$. Explicit formulae for these values are given by (2.8) with $\tau = 1$.

Next, we define μ^{aff} to be the value of complementarity that would be obtained by a full step to the boundary, that is,

$$\mu^{\text{aff}} = \left(x + \alpha_x^{\text{aff}} \Delta x^{\text{aff}}\right)^T \left(z + \alpha_z^{\text{aff}} \Delta z^{\text{aff}}\right) / n, \tag{3.4}$$

and set the centering parameter to be

$$\sigma = \left(\frac{\mu^{\text{aff}}}{x^T z / n}\right)^3. \tag{3.5}$$

This heuristic choice of σ is based on experimentation with linear programming problems and has proved to be effective for convex quadratic programming as well. Note that when good progress is made along the affine scaling direction, we have $\mu^{\text{aff}} \ll x^T z/n$, so the σ obtained from this formula is small. In other cases, σ may be chosen to be greater than 1.

Mehrotra's algorithm also computes a *corrector* step, but we take the view that the corrector is not part of the selection of the barrier parameter, and is simply a mechanism for improving the quality of the step. In section 7 we study the complete MPC algorithm including the corrector step.

Other adaptive procedures of the form (3.1) have been proposed specifically for nonlinear interior methods [11, 15, 24, 25]. The strategy employed in the LOQO software package [25] is particularly noteworthy because of its success in practice. It defines σ as

$$\sigma = 0.1 \min\left(0.05\frac{1-\xi}{\xi}, 2\right)^3, \quad \text{where} \quad \xi = \frac{\min_i \{x^{(i)} z^{(i)}\}}{x^T z/n}. \tag{3.6}$$

Note that ξ measures the deviation of the smallest complementarity product $x^{(i)} z^{(i)}$ from the average. When $\xi = 1$ (all individual products are equal to the average) we have that $\sigma = 0$ and the algorithm takes an aggressive step. The rule (3.6) always chooses $\sigma \leq 0.8$, so that even though the value of μ may increase from one iteration to the next, it will never be chosen to be larger than the current complementarity value $x^T z/n$.

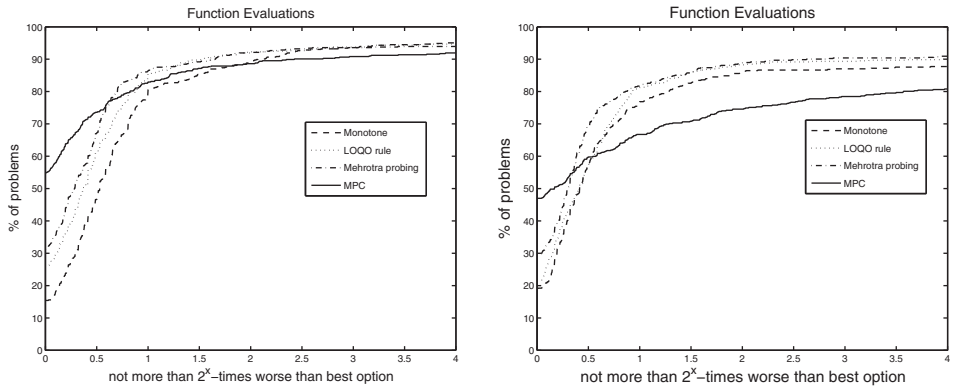
Our first set of numerical experiments compares the effectiveness of the two adaptive strategies mentioned above with the monotone Fiacco–McCormick approach. For these experiments, we use the IPOPT and KNITRO software packages, which have a globalization mechanism for the monotone variant but none for adaptive barrier parameter choices. These codes implement significantly different variations of the simple primal-dual iteration (2.5).

The experiments with KNITRO were done using the default Interior/Direct option (we will refer to this version as KNITRO-DIRECT henceforth), which implements a line search approach that is occasionally safeguarded by a trust region iteration [29]. The trust-region safeguard is needed, for example, to handle negative curvature directions. A merit function is used to promote global convergence, and when an adaptive barrier update rule is used, the penalty parameter associated with the merit function is reset at every iteration.

In our experiments with IPOPT, we go a step further and disable the line search within each iteration and always accept the full fraction-to-the-boundary step with step sizes from (2.8). In this way, we can examine the performance of pure primal-dual steps generated with the adaptive barrier schemes.

The barrier parameter strategies tested in our first set of experiments are as follows:

- LOQO *rule*. The barrier parameter is chosen by (3.1) and (3.6).
- *Mehrotra probing*. At every iteration, the barrier parameter μ is given by (3.1) and (3.5). Since this requires the computation of the affine scaling step (3.2), this strategy is more expensive than the LOQO rule. For KNITRO-DIRECT, in the iterations in which the safeguarding trust region algorithm is invoked (e.g., when the reduced Hessian is not positive definite), the barrier parameter is computed by the LOQO rule instead of Mehrotra probing. This is done because Mehrotra probing is expensive to implement in the trust region algorithm, which uses a conjugate gradient iteration.



(a) KNITRO (b) IPOPT

FIG. 1. *Results for four barrier parameter updating strategies.*

- *MPC.* The complete Mehrotra predictor-corrector algorithm as described later in section 7. As in the Mehrotra probing rule, when KNITRO-DIRECT falls back on the safeguarded trust region algorithm, the barrier parameter is computed using the LOQO rule for efficiency, and no corrector step is used.
- *Monotone.* (Also known as the Fiacco–McCormick approach.) The barrier parameter is fixed, and a series of primal-dual steps is computed, until the optimality conditions for the barrier problem are satisfied to some accuracy. At this point the barrier parameter is decreased. IPOPT and KNITRO implement somewhat different variations of this monotone approach; see [27, 29] for details about the initial value of μ, the rule for decreasing μ, and the form of the barrier stop tests.

For the numerical comparison, we select all the nonlinear programming problems in the CUTEr test set from January 2005 that contain at least one general inequality or bound constraint. We exclude those problems that seem infeasible, unbounded, or are given with initial points at which the model functions cannot be evaluated; see [27]. This gives a total of 599 problems. For all scalable models we use default sizes. Figure 1 reports the number of function evaluations for IPOPT and KNITRO, comparing the performance of the four barrier strategies. All the plots in the paper use the logarithmic performance profiles proposed by Dolan and Moré [10]. To account for the fact that different local solutions might be computed, problems with significantly different final objective function values for successful runs were excluded (for example, for the results in Figure 1, 37 problems we excluded for IPOPT, and 54 for KNITRO).

The results given in Figure 1 indicate that the adaptive strategies outperform the monotone variant, and in particular that Mehrotra probing appears to be the most successful in terms of function evaluations, both in the KNITRO experiment (using steplength control) and in the IPOPT experiment (with no steplength control). Furthermore, the results obtained with IPOPT show that the quality of the pure (unglobalized) steps is good enough to promote convergence in most problems. This observation suggests that the globalization scheme that we propose in section 5 should interfere minimally with the iteration; it should be active only when the algorithms appear to be making no progress.

We also note from Figure 1 that the complete MPC algorithm is very fast on some problems, but is not sufficiently robust. The latter can be seen most clearly in

Figure 1(b) where full MPC steps are taken at every iteration. The reason for the lack of robustness of the MPC strategy will be discussed in section 7, together with a globalization safeguarding procedure.

4. Quality functions. The Mehrotra and LOQO rules rely on the heuristic parameters (3.5) and (3.6). We now consider an approach in which μ is selected using a clear-cut objective, formulated in terms of a *quality function* to be minimized. As before, we assume that $\mu = \sigma \frac{x^T z}{n}$, where the centering parameter $\sigma \geq 0$ is to be determined, and define $\Delta(\sigma)$ to be the solution of the primal-dual equations (2.5) as a function of σ. We also let $\alpha_x^{\max}(\sigma), \alpha_z^{\max}(\sigma)$ denote the steplengths satisfying the fraction to the boundary rule (2.8) for the step $\Delta = \Delta(\sigma)$, and we define the *probing points*

$$x(\sigma) = x + \alpha_x^{\max}(\sigma)\Delta x(\sigma),$$

$$y(\sigma) = y + \alpha_z^{\max}(\sigma)\Delta y(\sigma), \quad z(\sigma) = z + \alpha_z^{\max}(\sigma)\Delta z(\sigma).$$

Our goal is to choose the value of σ that provides significant improvement toward the solution of the nonlinear program (2.1). For example, we could choose σ so as to minimize the following nonlinear quality function based on the KKT error:

$$\begin{aligned} q_N(\sigma) = &\left\|\nabla f(x(\sigma)) - A(x(\sigma))^T y(\sigma) - z(\sigma)\right\|^2 + \|c(x(\sigma))\|^2 \\ &+ \|Z(\sigma)X(\sigma)e\|^2. \end{aligned} \tag{4.1}$$

The evaluation of q_N is, however, expensive since it requires the evaluation of the problem functions and derivatives for every value of σ. We can avoid this expense by using a *linear quality function*. If we assume that f and c are linear functions, we have that (4.1) can be expressed as

$$\begin{aligned} q_L(\sigma) = &(1 - \alpha_z^{\max}(\sigma))^2 \left\|\nabla f(x) - A(x)^T y - z\right\|^2 + (1 - \alpha_x^{\max}(\sigma))^2 \|c(x)\|^2 \\ &+ \|(X + \alpha_x^{\max}(\sigma)\Delta X(\sigma))(Z + \alpha_z^{\max}(\sigma)\Delta Z(\sigma))e\|^2, \end{aligned} \tag{4.2}$$

where $\Delta X(\sigma)$ is the diagonal matrix with $\Delta x(\sigma)$ on the diagonal, and similarly for $\Delta Z(\sigma)$. We point out that by design the function q_L measures the KKT error exactly at the probing points $(x(\sigma), y(\sigma), z(\sigma))$ for linear programming problems.

Note that $\Delta(\sigma) = \Delta(0) + \sigma(\Delta(1) - \Delta(0))$. Therefore, $\Delta(\sigma)$ can be computed easily for any value of σ once the linear system (2.5) has been solved twice to obtain $\Delta(0)$ and $\Delta(1)$. Having computed $\Delta(\sigma)$, the dominant cost in the evaluation of q_L lies in the computation of the maximal steplengths $\alpha_x^{\max}(\sigma), \alpha_z^{\max}(\sigma)$ and the last term in (4.2), which requires a few vector operations.

We have defined the quality function q_L using squared norms to severely penalize any large components in the KKT error. Note that $q_L(\sigma)$ is not a convex function of σ, in general. Moreover, due to the complicated dependence of the steplengths $\alpha_x^{\max}(\sigma), \alpha_z^{\max}(\sigma)$ on the parameter σ, it does not seem possible to obtain an analytic expression for the minimizers of q_L. Nevertheless, we have observed that, in practice, this function is usually unimodal.

Therefore, we implement a one-dimensional search scheme to compute an approximate minimizer of q_L. It uses a golden trisection procedure (see, e.g., [21]), and ignores the fact that q_L may not necessarily be unimodal. We first choose $\sigma^{\min}$ and $\sigma^{\max}$, which define a minimum and maximum limit on the σ value, and

define the two intervals $[\sigma^{\min}, 1]$ and $[1, \sigma^{\max}]$. In our implementation, the value $\sigma^{\min} = \max(\gamma, \mu^{\min} n/x^T z)$, where $\mu^{\min}$ ($= 10^{-9}$ in our implementation) defines a minimal permissible value of the barrier parameter, γ is some small number (say, 10^{-6} or 10^{-8}), and $\sigma^{\max} = 1000$. We first evaluate the quality function for $\sigma = 1$ and for some σ value slightly less than 1 (say 0.99). If $q_L(0.99) \leq q_L(1)$, then we perform our golden trisection procedure in the interval $[\sigma^{\min}, 1]$, otherwise we search in the interval $[1, \sigma^{\max}]$. (It is important that σ be allowed to take on values greater than one so that the algorithm can recover from overly aggressive reductions of the barrier parameter.) Our trisection procedure terminates if either 12 evaluations of the quality functions are performed, or if the search interval $[a, b]$ becomes smaller than $b \times 10^{-2}$.

The expected advantages of the quality function approach are twofold. First, we have defined a procedure that ties the choice of the barrier parameter to a measurable and achievable decrease in the (linearized) KKT error. Therefore, we expect this approach to converge in fewer iterations compared with previously proposed approaches based on heuristic formulas. Second, our choice of the barrier parameter takes into account the fraction to the boundary steplengths (2.8) (the functions (4.1) and (4.2) are based on the steps *after* applying the fraction to the boundary rule). Thus the implicit constraints (2.4) are taken into account in choosing μ. This is similar to the Mehrotra update formulas and should discourage choices of the barrier parameter that generate steps which quickly violate the bounds (2.4) and need to be truncated.

More implementation details are given in section 6. Before presenting our numerical results with the quality function, we study how to guarantee the global convergence of nonlinear interior methods that choose the barrier parameter adaptively.

5. A globalization framework. The adaptive strategies described in section 3 can be seen from the numerical results in that section to be quite robust, even without a rigorous globalization scheme. (We show in the next section this is also the case with the quality function approach.) Yet, since the barrier parameter is allowed to change at every iteration in these algorithms, there is no mechanism that indeed enforces global convergence of the iterates in all cases. In contrast, the monotone barrier strategy employed in the Fiacco–McCormick approach allows us to establish global convergence results by combining two mechanisms. First, the algorithms that minimize a given barrier problem (2.2) use a line search or trust region to enforce a decrease in a merit function (as in KNITRO) or to guarantee acceptability by a filter (as in IPOPT). This ensures that an optimality test for the barrier function is eventually satisfied to some tolerance ϵ. Second, by repeating this minimization process for decreasing values of μ and ϵ that converge to zero, one can establish global convergence results [4, 12] to stationary points of the nonlinear programming problem (2.1).

We now propose two globalization frameworks that monitor the performance of the iterations in reference to a mechanism that enforces global convergence. As long as the adaptive primal-dual steps make sufficient progress towards the solution, the algorithm is free to choose a new value for the barrier parameter at every iteration; here, the barrier parameter can be chosen by *any* desired rule. We call this the *free mode.* However, if the iteration fails to maintain progress, then the algorithm reverts to a *monotone mode*, in which a Fiacco–McCormick strategy is applied. Here, the value of the barrier parameter remains fixed, and a robust globalization technique (e.g., based on a merit function or a filter) is employed to ensure progress for the corresponding barrier problem. Once the barrier problem is approximately minimized,

the barrier parameter is decreased. The monotone mode continues until an iterate is generated that makes sufficient progress for the original problem, at which point the free mode resumes.

We stress that also in the free mode we might want to choose steplengths α_p, α_d that are shorter than the maximal step sizes $\alpha_x^{\max}, \alpha_z^{\max}$, in order to promote convergence to minimizers. In our implementations, we make sure that the steps have descent properties with respect to the barrier problem (2.2) corresponding to the current value of μ, and we perform a line search to enforce progress in a merit function or a filter (without history), both of which are defined with respect to this barrier problem. In this way, we force the algorithm to consider the objective function when determining a new trial point, and not only the norm of the optimality conditions, so that convergence to stationary points that are not minimizers is less likely.

There are various ways to measure whether steps in the free mode make sustained progress toward the solution of the nonlinear program (2.1). We have developed two mechanisms, one based on a measure of KKT error, and the other using a filter based on the value of the objective (2.1a) and a measure of the constraint violation. Both aim to interfere with adaptive steps as little as possible so as not to slow down convergence.

5.1. Nonmonotone decrease of the KKT error. In our first globalization framework, we monitor the KKT error of the original nonlinear program,

$$\Phi(x, y, z) = \left\|\nabla f(x) - A(x)^T y - z\right\|^2 + \|c(x)\|^2 + \|ZXe\|^2. \tag{5.1}$$

We require that this measure be reduced by a factor of $\kappa \in (0, 1)$ over at most a fixed number $l^{\max}$ of iterations, when the algorithm is in the free mode.

Algorithm A: KKT-Error-Based Globalization Framework

Given (x_0, y_0, z_0) with $(x_0, z_0) > 0$, a constant $\kappa \in (0, 1)$ and an integer $l^{\max} \geq 0$.
Set $k \leftarrow 0$.
Repeat
  Choose a target value of the barrier parameter μ_k, based on any rule.
  Compute the primal dual search direction Δ from (2.5).
  Determine step sizes $\alpha_p \in (0, \alpha_x^{\max}]$ and $\alpha_d \in (0, \alpha_z^{\max}]$.
  Compute the new trial iterate $(\tilde{x}_{k+1}, \tilde{y}_{k+1}, \tilde{z}_{k+1})$ from (2.7).
  Compute the KKT error $\tilde{\Phi}_{k+1} \equiv \Phi(\tilde{x}_{k+1}, \tilde{y}_{k+1}, \tilde{z}_{k+1})$.
  Set $M_k = \max\{\Phi_{k-l}, \Phi_{k-l+1}, \ldots, \Phi_k\}$ with $l = \min\{k, l^{\max}\}$.
  If $\tilde{\Phi}_{k+1} \leq \kappa M_k$
    Accept $(\tilde{x}_{k+1}, \tilde{y}_{k+1}, \tilde{z}_{k+1})$ as the new iterate, and set $\Phi_{k+1} \leftarrow \tilde{\Phi}_{k+1}$.
    Set $k \leftarrow k + 1$ and return to the beginning of the loop.
  else
    Start Monotone Mode:
    Starting from $(\tilde{x}_{k+1}, \tilde{y}_{k+1}, \tilde{z}_{k+1})$, and for an initial value $\bar{\mu}$, solve a sequence of barrier problems with a monotonically decreasing sequence of barrier parameters to obtain a new iterate $(x_{k+1}, y_{k+1}, z_{k+1})$ such that
      $\Phi_{k+1} \equiv \Phi(x_{k+1}, y_{k+1}, z_{k+1}) \leq \kappa M_k$.
    Set $k \leftarrow k + 1$ and resume the free mode at the beginning of the loop.
  end if
End (repeat).

In the monotone mode, it is not required to solve each barrier problem to the specified tolerance before checking whether the method can revert to the free mode. Instead, we compute the optimality error $\Phi(x, y, z)$ for all intermediate iterates in the monotone mode, and return to the free mode, as soon as $\Phi(x, y, z) \leq \kappa M_k$.

Typical values for the algorithmic parameters are $\kappa = 0.9999$ and $l^{\max} = 5$. An important issue when switching to the monotone mode is the initialization of the barrier parameter $\bar{\mu}$. This can be chosen, for example, to be some fraction of the current complementarity value. The rule used in our implementations is $\bar{\mu} = 0.8(x_k^T z_k)/n$.

5.2. Two-dimensional filter. In the second globalization method, we make use of a filter that accepts a trial point if it provides sufficient progress in terms of the constraint violation $\theta(x) = \|c(x)\|$ *or* the objective function $f(x)$, compared to the previous iterates generated in the free mode. We let $\mathcal{F}_k \subseteq \{(f, \theta) \in \mathbb{R}^2 : \theta \geq 0\}$ denote the (f, θ) pairs that are not acceptable at the current iteration k. The concept of acceptability by the filter is made precise below.

Algorithm B: Filter-Based Globalization Framework

Given (x_0, y_0, z_0) with $(x_0, z_0) > 0$, and constants $\kappa_1, \kappa_2 > 0$; initialize the filter $\mathcal{F}_0 = \emptyset$.

Set $k \leftarrow 0$.

Repeat

Choose a target value of the barrier parameter μ_k, based on any rule.
Compute the primal dual search direction Δ from (2.5).
Determine step sizes $\alpha_p \in (0, \alpha_x^{\max}]$ and $\alpha_d \in (0, \alpha_z^{\max}]$.
Compute the new trial iterate $(\tilde{x}_{k+1}, \tilde{y}_{k+1}, \tilde{z}_{k+1})$ from (2.7).
Compute the filter margin $\delta_k = \kappa_1 \min\{\kappa_2, \Phi(x_k, y_k, z_k)\}$.
If $(f(\tilde{x}_{k+1}) + \delta_k, \|c(\tilde{x}_{k+1})\| + \delta_k) \notin \mathcal{F}_k$
Accept $(\tilde{x}_{k+1}, \tilde{y}_{k+1}, \tilde{z}_{k+1})$ as the new iterate.
Update the filter $\mathcal{F}_{k+1} = \mathcal{F}_k \cup \{(f, \theta) : f \geq f(\tilde{x}_{k+1}) \text{ and } \theta \geq \|c(\tilde{x}_{k+1})\|\}$.
Set $k \leftarrow k + 1$ and return to the beginning of the loop.
else
Start Monotone Mode:
Starting from $(\tilde{x}_{k+1}, \tilde{y}_{k+1}, \tilde{z}_{k+1})$, and for an initial value $\bar{\mu}$, solve a sequence of barrier problems with a monotonically decreasing sequence of barrier parameters to obtain a new iterate $(x_{k+1}, y_{k+1}, z_{k+1})$ such that
$$(f(x_{k+1}) + \delta_k, \|c(x_{k+1})\| + \delta_k) \notin \mathcal{F}_k.$$
Augment the filter:
$$\mathcal{F}_{k+1} = \mathcal{F}_k \cup \{(f, \theta) : f \geq f(x_{k+1}) \text{ and } \theta \geq \|c(x_{k+1})\|\}.$$
Set $k \leftarrow k + 1$ and resume the free mode at the beginning of the loop.
end if

End (repeat).

Similar to the KKT-error based globalization framework, the monotone mode is terminated as soon as an iterate is encountered that is acceptable to the filter.

We have tested the KKT and filter globalization approaches using IPOPT and KNITRO, and found both to be effective in practice. For the sake of brevity, we report results only for the filter globalization framework in the next section. We set $\kappa_1 = 10^{-5}$ and $\kappa_2 = 1$ for these tests, and we choose $\bar{\mu} = 0.8(x_k^T z_k)/n$ for the barrier parameter when entering the monotone mode.

5.3. Global convergence results. In the following we summarize the theoretical convergence guarantees for the two globalization frameworks presented above.

THEOREM 5.1. *Let $\{(x_k, y_k, z_k)\}$ be the sequence generated by either Algorithm A or Algorithm B, and assume that the monotone mode always terminates successfully. For Algorithm B, further assume that $\{f(x_k)\}$ is bounded below and that $\{\|c(x_k)\|\}$ is bounded above. Then, the KKT error $\Phi(x_k, y_k, z_k)$ converges to zero.*

Proof. Algorithm A: Since this framework ensures that the optimality measure $\Phi_k = \Phi(x_k, y_k z_k)$ is reduced by a factor of $\kappa \in (0, 1)$ in at most every $l^{\max}$ iterations, it is clear that $\Phi_k \to 0$.

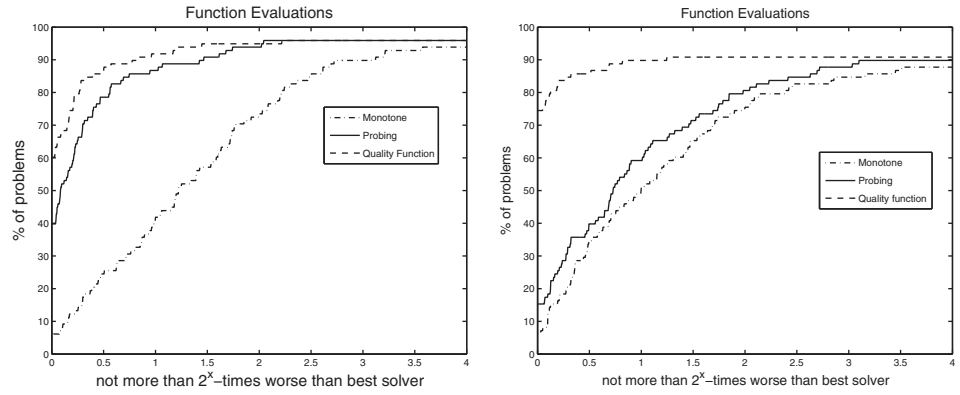
Algorithm B: The proof is by contradiction and is similar to the proof of Lemma 3.3 in [13]. Suppose that there is a subsequence $\{k_j\}$ of iterations in which $\delta_{k_j-1} \geq \epsilon > 0$. Then $f(x_{k_j})$ has to be bounded above, say by f_U, since otherwise we could find a subsequence $\{k_{j_l}\}$ of $\{k_j\}$ with $f(x_{k_{j_l}}) \leq f(x_{k_{j_{l+1}}})$ and $f(x_{k_{j_l}}) \to \infty$ so that the filter update rule would yield $\|c(x_{k_{j_{l+1}}})\| < \|c(x_{k_{j_l}})\| - \delta_{k_{j_l}-1} \leq \|c(x_{k_{j_l}})\| - \epsilon \to -\infty$. Therefore, for each k_j, the area of the region $\mathcal{F}_{k_j} \setminus \mathcal{F}_{k_j-1}$ added to $\mathcal{F}_{k_j-1}$ includes a square of size $\delta^2_{k_j-1} \geq \epsilon^2$ within the set $\bar{\mathcal{F}} = \{(f, \theta) : f_L \leq f \leq f_U \text{ and } 0 \leq \theta \leq \theta_U\}$. Here, f_L denotes a lower bound of $\{f(x_k)\}$ and θ_U an upper bound of $\{\|c(x_k)\|\}$, which exist by assumption. Because of the monotonicity $\mathcal{F}_k \subseteq \mathcal{F}_{k+1}$ of the filter, this leads to a contradiction, since $\bar{\mathcal{F}}$ is finite. ∎

The above result pertains to the cases where the algorithm does not eventually stay in the monotone mode. If it does, the iteration inherits the convergence properties from the underlying Fiacco–McCormick algorithm. In particular, if the nonlinear program (2.1) is infeasible the KKT error cannot converge to zero, and therefore Theorem 5.1 shows that both algorithms must eventually remain in the monotone mode. In that mode, there will be a value of the barrier parameter, say $\bar{\mu}$, for which the corresponding barrier problem is infeasible and cannot be solved to the required convergence tolerance. For KNITRO, it has been shown that the algorithm then generates an infeasible limit point that is a stationary point for the ℓ_2-norm of the constraint violation [4]. For IPOPT, the filter line-search algorithm for that barrier problem will eventually stay in the restoration phase [26]; the current implementation of the restoration phase then minimizes the ℓ_1-norm of the constraint violation. Therefore, if the nonlinear program is infeasible, both algorithms will generate a message indicating that the problem is locally infeasible.

6. Numerical results. We first discuss the choice of norms in the quality function (4.2) and in the optimality measure (5.1). (For consistency, we use the same norms and scaling factors for the individual terms in (4.2) and (5.1).) In IPOPT we use the 2-norm, and each of the three terms is divided by the number of elements in the vectors whose norms are being computed. In KNITRO, we choose the norm and scaling factors to be similar to the terms used in the KNITRO termination test: The first two terms in (4.2) and (5.1) use the infinity-norm, the complementarity term uses the 1-norm divided by n, and we scale these terms using the factors described in [29].

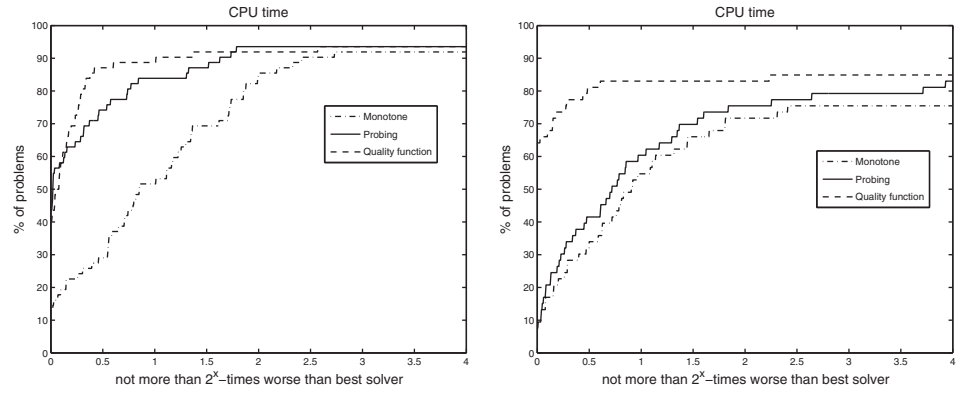
The tests involving IPOPT were run on a Dual-Pentium III, 1GHz machine running Linux. The KNITRO tests were run on a machine with an AMD Athlon XP 3200+ 2.2GHz processor running Linux. For both codes, the maximum number of iterations was set to 3000 and the time limit was set to 1800 CPU seconds. The tests were run using the development versions of IPOPT and KNITRO as of October 2005.

The first results we present are for the linear programming problems in the NETLIB collection, as specified in the CUTEr test set [18]. No preprocessing was



(a) IPOPT (b) KNITRO

FIG. 2. *Function evaluation comparison for the NETLIB test set.*

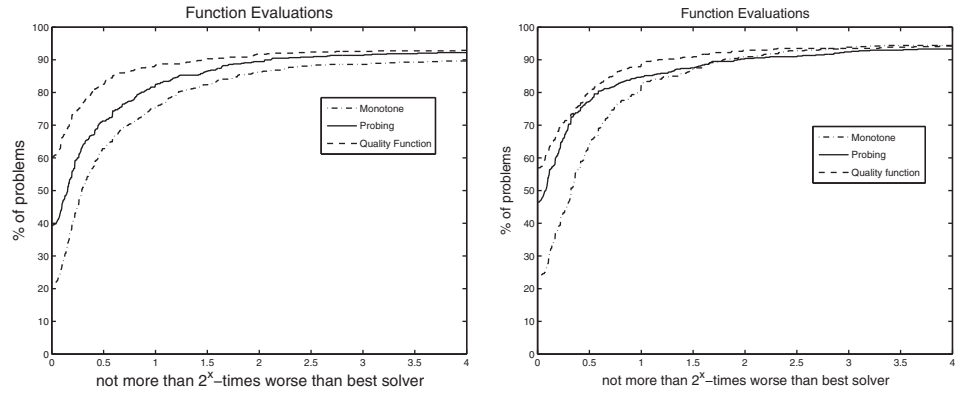


(a) IPOPT (b) KNITRO

FIG. 3. *CPU time comparison for the NETLIB test set.*

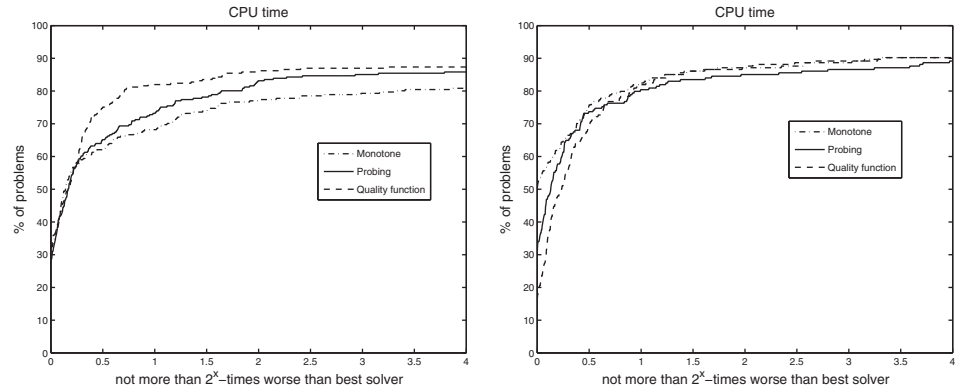
performed, and no initial point strategy was employed (i.e., the default starting point $x_0 = (0, \ldots, 0)$ was used). Figure 2 compares the performance of the quality function approach, in terms of function evaluation count, with two of the strategies described in section 3, namely, the monotone method and the Mehrotra probing heuristic. Figure 3 compares the algorithms in terms of CPU time. Since we are primarily interested in methods with globally convergent frameworks we use the filter based globalization framework described in section 5 for both the Mehrotra probing approach and the quality function approach. The monotone approach is globally convergent on its own. Even though our focus is on nonlinear optimization, linear programming problems are of interest since they allow us to assess the effectiveness of the quality function in a context in which it exactly predicts the KKT error. It is apparent from Figure 2 that the quality function approach is very effective on the NETLIB test set.

The performance (in terms of function evaluations) of the three barrier update strategies on nonlinear programming problems with at least one inequality or bound constraint from the CUTEr collection is reported in Figure 4. The quality function approach again performs significantly better than the monotone method; it also



(a) IPOPT (b) KNITRO

FIG. 4. *Function evaluation comparison for CUTEr test set.*



(a) IPOPT (b) KNITRO

FIG. 5. *CPU time comparison for the CUTEr test set.*

outperforms the Mehrotra probing strategy, which had given the best results in the experiments reported in section 3. The improvements are less pronounced when comparing CPU performance; see Figure 5.

To give more insight into the behavior of the quality function approach, we present in Table 1 data about the value of the centering parameter σ_k chosen by the quality function approach, together with statistics about the globalization strategy employed. The data was collected from the results produced by IPOPT. We compare the probing and quality function approaches. The third column gives the percentage of iterations spent in the monotone mode. The rest of the columns report the percentage of iterations in which σ_k lie in the intervals $[\sigma^{\max}, 10]$, $(10, 1]$, $(1, 10^{-1}]$, etc. The percentage numbers in Table 1 were obtained by computing the average of the percentages for each successfully solved problem.

As we can see, only a small percentage of iterations is spent in the monotone mode, showing that the adaptive mode is the main driving force. Note also that the quality function strategy tends to produce larger values of σ_k than the probing approach.

TABLE 1
Average percentage of iterations in monotone and free mode, producing σ_k in certain ranges.

Method	Testset	%mono	free mode, with σ_k in range							
		mode	≥ 10	≥ 1	$\geq 10^{-1}$	$\geq 10^{-2}$	$\geq 10^{-3}$	$\geq 10^{-4}$	$\geq 10^{-5}$	$< 10^{-5}$
Probing	NETLIB	4.77%	0.00	2.15	72.35	10.18	3.29	1.33	1.44	4.50
Qual. fctn.	NETLIB	3.70%	4.90	11.39	53.41	14.69	3.56	0.63	0.65	7.07
Probing	CUTEr	6.92%	0.49	2.25	26.59	19.99	12.08	5.90	4.27	21.50
Qual. fctn.	CUTEr	8.39%	4.53	6.98	32.89	22.50	10.57	2.08	2.46	9.59

The performance profiles in Figures 4 and 5 indicate that there are a number of problems that cannot be solved by the quality function approach. We now make some observations about the behavior of the two codes on these problems.

Considering both the NETLIB and CUTEr test sets, there were 38 failures for the KNITRO implementation of the globalized quality function adaptive barrier rule. Of these, 18 problems were solved when the default time or iteration limits were increased; 5 problems (BRANDY, PALMER2, PALMER7A, SAWPATH, SINEALI) terminated at near optimal approximate solutions, but KNITRO could not get enough accuracy in the dual feasibility measure; 2 problems (CRESC132, QCNEW) terminated because of evaluation errors resulting from IEEE exceptions (`NaN`) in the function evaluation; and 1 problem (HS110) terminated with a message of unboundedness at a feasible point with a very large negative value of the objective function. The optimal objective for HS110 is $-9.960\text{e}+39$,[1] whereas by default KNITRO declares unboundedness for a feasible objective value less than the cutoff limit value $-1.0\text{e}+20$. When this limit is changed, KNITRO solves the problem in 4 iterations. In addition, the problem KTMODEL was discovered to have incorrect gradients which caused KNITRO to terminate at an infeasible point. The remaining 11 problems (COSHFUN, DITTERT, DRUGDISE, GREENBEA, GREENBEB, MANNE, NUFFIELD, PILOT-JA, TENBARS2, TENBARS3, ZIGZAG) constitute unresolved failures.

We give some more information about these 11 problems. GREENBEA, GREENBEB, PILOT-JA, NUFFIELD, and ZIGZAG were not solved even when the time limit was increased to three hours. The problems GREENBEA, GREENBEB, and PILOT-JA are linear programs for which KNITRO experiences numerical difficulties (from rank-deficient Jacobians) that cause it to often fallback on steepest descent like steps and converge slowly. KNITRO appears to be very close to the solution in PILOT-JA when it reaches the time limit. The problems COSHFUN, DRUGDISE, TENBARS2, and TENBARS3 were not solved even when the iteration limit was raised to 100,000 (although it appears in all cases that slow progress is still being made when the iteration limit is reached). For the problem DITTERT, KNITRO terminates at an infeasible point but the code is unable to verify whether or not it is an infeasible stationary point. Finally, for the problem MANNE, KNITRO terminated at a feasible point but with a large dual feasibility error.

For the IPOPT code, there were 39 failures for the globalized quality function adaptive barrier rule. Of these, 13 problems were solved if more iterations (100,000) or CPU time (3 hours) were allowed. In 6 problems (A2NNDNIL, A5NNDNIL, CRESC100, EG3, POLAK3, SPIRAL), IPOPT terminated at a point satisfying the local infeasibility criterion; in 2 problems (A2NSDSIL, BRAINPC9) it failed during the restoration phase. For COSHFUN, the memory requirement in the linear solver was exceeded in iteration 10681, and the problems EQC and ROBOT terminated because the search

[1] This is true for problem size N=200 which was the default size for this SIF model at the time of testing.

direction became too small, but in both cases the problem was almost solved. For problem LIN, the objective function value at the (modified) starting point resulted in an IEEE exception (`NaN`).

In the remaining failures, the maximum iteration count or CPU time were exceeded. Problems AVION2, PALMER5E, YORKNET terminated after 100,000 iterations; IPOPT appeared to be cycling with small primal and dual feasibility errors for two of these problems (AVION2 and PALMER5E). For problem PALMER7E, IPOPT still made very small progress after 100,000 iterations, while for problem KT-MODEL the code seemed to diverge (as a result of incorrect gradient information). Finally, the time limit was exceeded for the linear program QAP15 from NETLIB, and for the CUTEr models A5NSDSIL, CRESC132, GAUSSELM, GLIDER, MANNE, NUFFIELD, ORTHREGE, READING8.

7. Corrector steps. The numerical results of section 3 indicate that, when solving nonlinear problems, including the corrector step in Mehrotra's method (the MPC method) is often not beneficial. This is in stark contrast with the experience in linear programming and convex quadratic programming, where the corrector step is known to accelerate the interior-point iteration without degrading its robustness. In this section we study the effect of the corrector step and find that it can also be harmful in the linear programming and quadratic programming cases *if* an initial point strategy is not used. These observations are relevant because in nonlinear programming it is much more difficult to find a good starting point.

Let us begin by considering the linear programming case. There are several ways of viewing the MPC method in this context. One is to consider the step computation as taking place in three stages (see, e.g., [30]). First, the algorithm computes the affine scaling step (3.2) and uses it to determine the target value of the barrier parameter $\mu = \sigma \frac{x^T z}{n}$, where σ is given by (3.5). Next, the algorithm computes a primal-dual step, say Δ^{pd}, from (2.5) using that value of μ. Finally, a corrector step Δ^{corr} is computed by solving (2.5) with the right-hand side given by

$$-\left(0, \Delta X^{\mathrm{aff}} \Delta Z^{\mathrm{aff}} e, 0\right)^T, \tag{7.1}$$

where ΔX^{aff} is the diagonal matrix with diagonal entries given by Δx^{aff}, and similarly for ΔZ^{aff}. The complete MPC step is the sum of the primal-dual and corrector steps. We can compute it by adding the right-hand sides and solving the following system:

$$\begin{bmatrix} \nabla^2_{xx}\mathcal{L} & -A^T(x) & -I \\ Z & 0 & X \\ A(x) & 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta x^{\mathrm{mpc}} \\ \Delta y^{\mathrm{mpc}} \\ \Delta z^{\mathrm{mpc}} \end{bmatrix} = - \begin{bmatrix} \nabla f(x) - A^T(x) y - z \\ Xz - \mu e + \Delta X^{\mathrm{aff}} \Delta Z^{\mathrm{aff}} e \\ c(x) \end{bmatrix}. \tag{7.2}$$

The new iterate (x^+, y^+, z^+) of the MPC method is given by (2.7)–(2.8) with $\Delta = (\Delta x^{\mathrm{mpc}}, \Delta y^{\mathrm{mpc}}, \Delta z^{\mathrm{mpc}})$.

Alternative views of the MPC method are possible by the linearity of the step computation: We can group the right-hand side in (7.2) in different ways and thereby interpret the step as the sum of different components. Yet all these views point out the following inconsistency in the MPC approach.

In the linear programming case, primal and dual feasibility are linear functions and hence vanish at the full affine scaling point, defined by

$$(x, y, z) + \left(\Delta x^{\mathrm{aff}}, \Delta y^{\mathrm{aff}}, \Delta z^{\mathrm{aff}}\right). \tag{7.3}$$

The complementarity term takes on the value

$$\left(X + \Delta X^{\mathrm{aff}}\right)\left(Z + \Delta Z^{\mathrm{aff}}\right) = \Delta X^{\mathrm{aff}} \Delta Z^{\mathrm{aff}}.$$

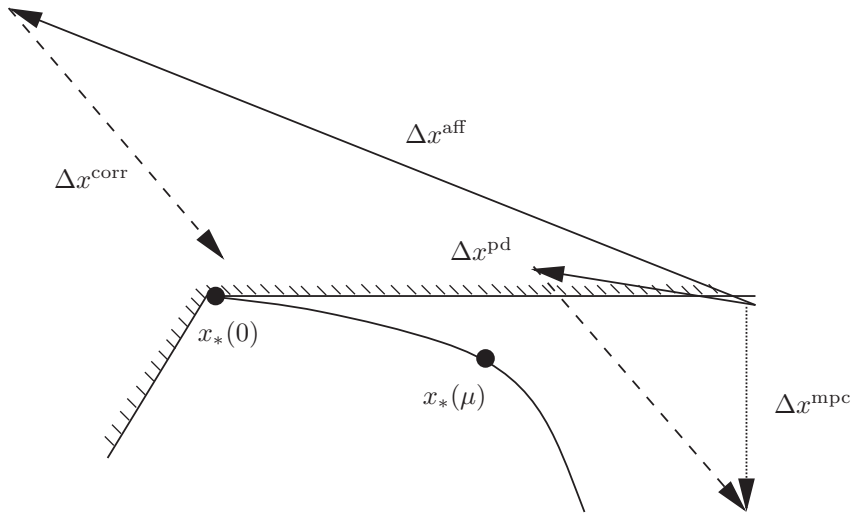


FIG. 6. *An unfavorable corrector step.*

TABLE 2
Output for NETLIB problem FORPLAN *for default PCx with bad starting point.*

Iter	Primal Obj	Dual Obj	PriInf	DualInf	$\alpha_x^{\max}$	$\alpha_z^{\max}$	$\log\left(\frac{x^T z}{n}\right)$	$\|\Delta^{\text{aff}}\|$	$\|\Delta^{\text{mpc}}\|$
0	9.0515e + 01	−4.8813e + 06	1.0e − 00	1.9e + 00	0.0e + 00	0.0e + 00	0.00	0.0e + 00	0.0e + 00
1	9.0216e + 01	−1.3664e + 08	1.0e − 00	1.9e + 00	8.6e − 13	5.0e − 12	0.08	5.6e + 06	1.2e + 13
2	9.0403e + 01	−3.3916e + 08	1.0e − 00	1.9e + 00	7.3e − 13	4.8e − 13	0.16	9.1e + 07	1.9e + 14
3	9.0769e + 01	−1.1343e + 10	1.0e − 00	1.9e + 00	4.0e − 12	1.2e − 11	1.18	2.2e + 08	3.9e + 14
4	9.0860e + 01	−1.8010e + 11	1.0e − 00	1.9e + 00	1.5e − 12	5.0e − 12	2.35	8.0e + 09	1.4e + 16
5	9.1312e + 01	−2.9307e + 12	1.0e − 00	1.9e + 00	4.3e − 12	5.1e − 12	3.56	1.3e + 11	2.2e + 17
6	9.1710e + 01	−8.2787e + 13	1.0e − 00	1.9e + 00	6.0e − 12	9.1e − 12	5.01	2.1e + 12	3.6e + 18
7	9.2036e + 01	−1.5505e + 15	1.0e − 00	1.9e + 00	7.5e − 12	6.0e − 12	6.28	5.9e + 13	1.0e + 20
8	9.2282e + 01	−6.8149e + 16	1.0e − 00	1.9e + 00	7.0e − 12	1.4e − 11	7.93	1.1e + 15	1.9e + 21
9	9.2279e + 01	−4.4155e + 18	1.0e − 00	1.9e + 00	9.2e − 12	2.1e − 11	9.74	4.8e + 16	8.3e + 22
10	9.2244e + 01	−2.8697e + 20	1.0e − 00	1.9e + 00	6.8e − 12	2.1e − 11	11.55	3.1e + 18	5.4e + 24
11	9.2381e + 01	−3.1118e + 22	1.0e − 00	1.9e + 00	1.1e − 11	3.6e − 11	13.58	2.0e + 20	3.5e + 26
12	9.2462e + 01	−7.0471e + 24	1.0e − 00	2.2e + 01	6.2e − 12	7.6e − 11	15.94	2.2e + 22	3.8e + 28
13	9.2523e + 01	−9.9820e + 26	1.0e − 00	2.8e + 03	1.4e − 11	4.7e − 11	18.09	5.0e + 24	8.6e + 30
14	9.2605e + 01	−1.1959e + 30	1.0e − 00	2.2e + 01	2.1e − 11	4.0e − 10	21.17	7.1e + 26	1.2e + 33

Therefore the value of the right-hand side vector in (2.5) at the full affine scaling step (7.3) is given by (7.1). Thus the corrector step can be viewed as a modified Newton step taken from the point (7.3) and using the primal-dual matrix evaluated at the current iterate (x, y, z).

The inconsistency in the MPC approach arises because the corrector step, which is designed to improve the full affine scaling step, is applied at the primal-dual point; see Figure 6. In some circumstances, this mismatch can cause poor steps. In particular, we have observed that if the affine scaling step is very long, in the sense that the steplengths (2.8) are very small, and if the corrector step is even larger, then the addition of the corrector step to the primal-dual step (2.7) can significantly increase the complementarity value $x^T z$. This behavior can be sustained and lead to very slow convergence or failure, as shown in Table 2. The results in this table were obtained using PCx [9], an interior-point code for linear programming that implements the MPC method, applied to problem FORPLAN from the NETLIB collection. Practical implementations of the MPC use a procedure for choosing a favorable starting point

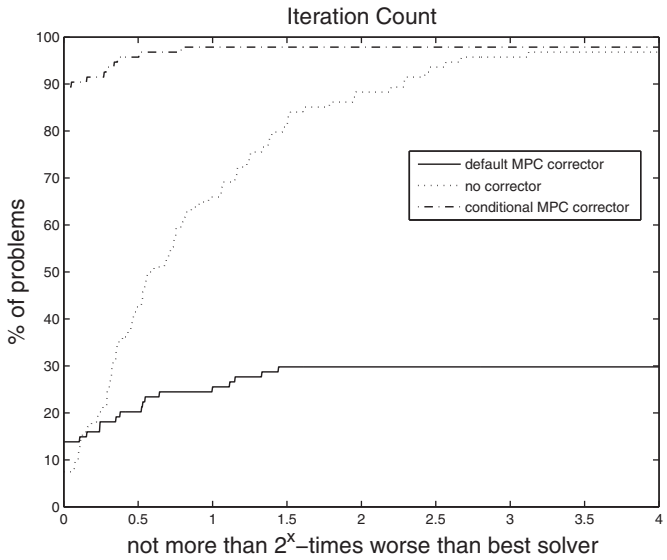


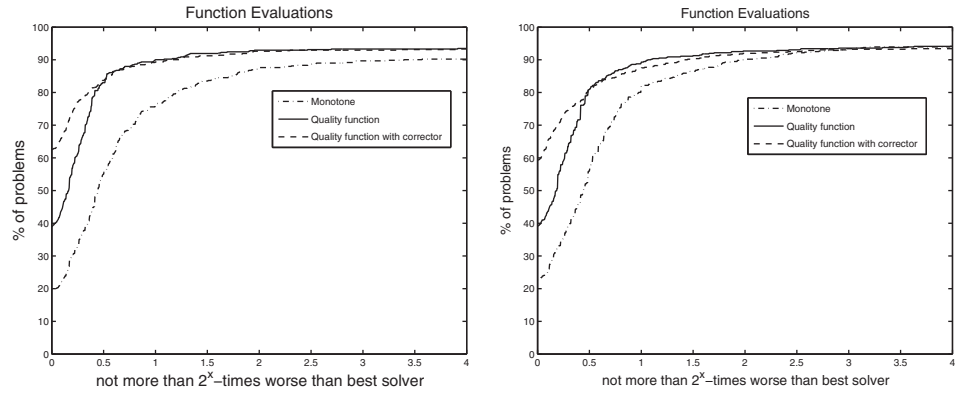
FIG. 7. *Results on the NETLIB test set for three corrector step strategies implemented in PCx. The initial point was set to* $x = e, z = e$.

described by Mehrotra [22]. We disabled this initial point strategy for PCx and set the initial point to $x = e, z = e$. Note from Table 2 that the affine scaling and corrector steps appear to grow without bound, and examination of the results shows that the dual variables diverge.

To provide further support to the claim that the corrector step can be harmful, we ran the complete set of test problems (94 in all) in the NETLIB collection. Using the default settings, which includes a strategy for computing a good starting point, PCx solved 90 problems, and terminated very close to the solution in the remaining 4 cases. Next we disabled the initial point strategy and set the initial point to $x = e, z = e$. PCx was now able to solve only 28 problems (and in only 3 additional cases terminated very close to the solution).

We repeated the experiment, using the initial point $x = e, z = e$, but this time removing the corrector step; this corresponds to the algorithm called *Mehrotra probing* in section 3. We also tested a variant that we call *conditional MPC* in which the corrector step is employed in the MPC method only if it does not result in an increase of complementarity by a factor larger than 2. The results, in terms of iterations, are reported in Figure 7. Note the dramatic increase in robustness of both strategies, compared with the MPC algorithm. The conditional MPC strategy is motivated by the observation that harmful effects of the corrector steps manifest themselves in a significant increase in complementarity. The failure of convergence of the MPC method has also been analyzed by Cartis [7].

Finally we compare the monotone and quality function approaches described in section 3 with the conditional MPC approach on the nonlinear programming problems used in that section. The conditional MPC method is now implemented so as to reject corrector steps that increase complementarity (this more conservative approach appears to be more suitable in the nonlinear case). Furthermore, if the conditional MPC step does not pass the merit function or filter acceptance test for the current barrier problem, the corrector step is also rejected, and the backtracking line search



(a) IPOPT (b) KNITRO

FIG. 8. *Results for safeguarded corrector steps.*

for the regular primal-dual step is executed. Finally, no corrector step is computed while the algorithm is in the monotone mode. The results, given in Figure 8, indicate that this conditional MPC method requires fewer function evaluations, and is not less robust, than the other strategies.

8. Conclusion. We have seen in this paper that, both for linear and nonlinear programming, classical barrier update strategies with global convergence guarantees are overly conservative, while strategies that are often fastest in practice are based on heuristic formulas and are not globally convergent. We have presented a new update strategy and shown that it is efficient in practice. Instead of basing the update on a heuristic formula, our approach follows a clearly defined objective, namely, the minimization of a quality function. We further proposed a simple globalization framework that makes use of any nonmonotone barrier parameter strategy.

We have also presented results that show the corrector steps employed in MPC can have harmful effects, even in the linear and quadratic programming cases. These observations were unexpected since the MPC method has become widely used in linear and quadratic programming; our tests show that the reliability of the MPC method depends crucially on heuristics, such as the choice of the initial point. We have shown, however, that the selective use of corrector steps can have beneficial effects in interior point methods.

A question we have not addressed is whether the approach presented in this paper enjoys fast local convergence. We have not specifically introduced features that guarantee superlinear convergence. This could be done by using various techniques proposed in the literature for controlling the asymptotic behavior of the barrier parameter; see, e.g., [19] and the references therein. In particular, we could implement the strategies recently proposed by Armand et al. [1, 2] in conjunction with the quality function approach.

We have not done so because the asymptotic behavior of the method proposed in this paper has proved to be acceptable in practice. In fact, the quality function approach promotes fast local convergence because it chooses the barrier parameter so as to (approximately) minimize the quality function in the region defined by (2.8). One can design a superlinearly convergent algorithm by choosing the barrier parameter so that a step to the region defined by (2.8) decreases the KKT error superlinearly.

Since the quality function is an approximation to the KKT error, and it attempts to minimize it, it is not surprising that the quality function approach tends to yield fast local convergence.

Acknowledgments. We would like to thank Richard Byrd for many valuable suggestions during the course of this work, and the referees for suggesting ways of improving the paper.

REFERENCES

[1] P. Armand and J. Benoist, *A local convergence property of primal-dual methods for nonlinear programming,* Math. Program., Ser. A, 115 (2008), pp. 199–222.

[2] P. Armand, J. Benoist, and D. Orban, *Dynamic updates of the barrier parameter in primal-dual methods for for nonlinear programming,* Comput. Optim. Appl., 41 (2008), pp. 1–25.

[3] J. Betts, S. K. Eldersveld, P. D. Frank, and J. G. Lewis, *An interior-point nonlinear programming algorithm for large scale optimization,* in Large-Scale PDE-Constrained Optimization, O. Ghattas, M. Heinkenschloss, D. Keyes, L. T. Biegler, and B. van Bloemen Waanders, eds., Lecture Notes Comput. Sci. Eng., Springer Verlag, 2003, pp. 184–198.

[4] R. H. Byrd, J.-Ch. Gilbert, and J. Nocedal, *A trust region method based on interior point techniques for nonlinear programming,* Math. Program., 89 (2000), pp. 149–185.

[5] R. H. Byrd, M. E. Hribar, and J. Nocedal, *An interior point algorithm for large scale nonlinear programming,* SIAM J. Optim., 9 (1999), pp. 877–900.

[6] R. H. Byrd, J. Nocedal, and R. A. Waltz, *KNITRO: An integrated package for nonlinear optimization,* in Large-Scale Nonlinear Optimization, G. di Pillo and M. Roma, eds., Springer, Berlin, 2006, pp. 35–59.

[7] C. Cartis, *Some disadvantages of a Mehrotra-type primal-dual corector interior point algorithms for linear programming,* Appl. Numer. Math., to appear.

[8] L. Chen and D. Goldfarb, *Interior-point ℓ_2 penalty methods for nonlinear programming with strong global convergence properties,* Math. Program., 108 (2006), pp. 1–36.

[9] J. Czyzyk, S. Mehrotra, and S. J. Wright, *PCx User Guide,* Technical report, Argonne National Laboratory, Argonne, IL, 1996.

[10] E. D. Dolan and J. J. Moré, *Benchmarking optimization software with performance profiles,* Math. Program., Ser. A, 91 (2002), pp. 201–213.

[11] A. S. El-Bakry, R. A. Tapia, T. Tsuchiya, and Y. Zhang, *On the formulation and theory of the Newton interior-point method for nonlinear programming,* J. Optim. Theory Appl., 89 (1996), pp. 507–541.

[12] A. V. Fiacco and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques.* J. Wiley and Sons, Chichester, England, 1968. Reprinted as *Classics in Applied Mathematics 4*, SIAM, Philadelphia, 1990.

[13] R. Fletcher, N. I. M. Gould, S. Leyffer, P. L. Toint, and A. Wächter, *Global convergence of a trust-region SQP-filter algorithms for general nonlinear programming,* SIAM J. Optim., 13 (2002), pp. 635–659.

[14] A. Forsgren and P. E. Gill, *Primal-dual interior methods for nonconvex nonlinear programming,* SIAM J. Optim., 8 (1998), pp. 1132–1152.

[15] D. M. Gay, M. L. Overton, and M. H. Wright, *A primal-dual interior method for nonconvex nonlinear programming,* in Advances in Nonlinear Programming (Beijing, 1996), Y. Yuan, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 31–56,

[16] E. M. Gertz and P. E. Gill, *A primal-dual trust region algortihm for nonlinear programming,* Math. Program., 100 (2004), pp. 49–94.

[17] N. I. M. Gould, D. Orban, and P. Toint, *An interior-point L1-penalty method for nonlinear optimization,* Technical report RAL-TR-2003-022, Rutherford Appleton Laboratory Chilton, Oxfordshire, UK, 2003.

[18] N. I. M. Gould, D. Orban, and P. L. Toint, CUTEr *and* sifdec*: A constrained and unconstrained testing environment, revisited,* ACM Trans. Math. Software, 29 (2003), pp. 373–394.

[19] N. I. M. Gould, D. Orban, and P. L. Toint, *Numerical methods for large-scale nonlinear optimization,* Acta Numer., (2005), pp. 299–361.

[20] I. Griva, D. F. Shanno, and R. J. Vanderbei, *Convergence Analysis of a Primal-Dual Method for Nonlinear Programming,* Report, Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ, 2004.

[21] D. G. Luenberger, *Linear and Nonlinear Programming,* 2nd ed., Addison-Wesley Publishing Company, Reading, MA, 1984.
[22] S. Mehrotra, *On the implementation of a primal-dual interior point method,* SIAM J. Optim., 2 (1992), pp. 575–601.
[23] A. L. Tits, A. Wächter, S. Bakhtiari, T. J. Urban, and C. T. Lawrence, *A primal-dual interior-point method for nonlinear programming with strong global and local convergence properties,* SIAM J. Optim., 14 (2003), pp. 173–199.
[24] M. Ulbrich, S. Ulbrich, and L. Vicente, *A globally convergent primal-dual interior point filter method for nonconvex nonlinear programming,* Math. Program., 2 (2004), pp. 379–410.
[25] R. J. Vanderbei and D. F. Shanno, *An interior point algorithm for nonconvex nonlinear programming,* Comput. Optim. Appl., 13 (1999), pp. 231–252.
[26] A. Wächter and L. T. Biegler, *Line search filter methods for nonlinear programming: Motivation and global convergence,* SIAM J. Optim., 16 (2005), pp. 1–31.
[27] A. Wächter and L. T. Biegler, *On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming,* Math. Program., 106 (2006), pp. 25–57.
[28] R. A. Waltz, Knitro 4.0 *User's Manual,* Technical report, Ziena Optimization, Inc., Evanston, IL, USA, 2004.
[29] R. A. Waltz, J. L. Morales, J. Nocedal, and D. Orban, *An interior algorithm for nonlinear optimization that combines line search and trust region steps,* Math. Program., Ser. A, 107 (2006), pp. 391–408.
[30] S. Wright, *Primal-dual interior-point methods,* SIAM, Philadelphia, 1997.
[31] H. Yamashita, *A globally convergent primal-dual interior-point method for constrained optimization,* Optim. Methods Software, 10 (1998), pp. 443–469.
[32] H. Yamashita, H. Yabe, and T. Tanabe, *A globally and superlinearly convergent primal-dual interor point trust region method for large scale constrained optimization,* Math. Program., Ser. A, 102 (2005), pp. 111–120.

SIAM J. OPTIM.
Vol. 19, No. 4, pp. 1694–1718

AN ADAPTIVE SCALARIZATION METHOD IN MULTIOBJECTIVE OPTIMIZATION*

GABRIELE EICHFELDER†

Abstract. This paper presents a new method for the numerical solution of nonlinear multiobjective optimization problems with an arbitrary partial ordering in the objective space induced by a closed pointed convex cone. This algorithm is based on the well-known scalarization approach by Pascoletti and Serafini and adaptively controls the scalarization parameters using new sensitivity results. The computed image points give a nearly equidistant approximation of the whole Pareto surface. The effectiveness of this new method is demonstrated with various test problems and an applied problem from medicine.

Key words. multicriteria optimization, vector optimization, approximation, sensitivity, scalarization approaches

AMS subject classifications. 90C29, 90C31, 90C59

DOI. 10.1137/060672029

1. Introduction.

The optimization problems arising nowadays in application areas like engineering, economics, or the sciences are often multiobjective; i.e., several competing objective functions have to be minimized all at once. Those optimization problems have in general not only one best solution but the solution set is very large.

In the last decades the main focus was on finding one minimal solution, e.g., by interactive methods, whereas objective numerical calculations alternate with subjective decisions done by a so-called decision maker (d.m.). Based on much better computer performance it is now possible to represent the whole efficient set. Having an approximation of the whole efficient set available, the d.m. gets a useful insight in the problem structure. Especially in engineering tasks it is interesting to have all design alternatives available [28]. So in this paper we aim to generate an approximation of the efficient set as it is done in many other works, e.g., in [9, 18, 19, 24, 42, 48].

However, the information provided by this approximation depends mainly on the quality of the approximation. Many points are related to a high numerical effort and to too many points which have to be interpreted by the d.m. A sparse approximation neglects large parts of the efficient set. According to different quality criteria as discussed in [47], an approximation is good in the sense of a stinted but representative presentation if the approximation points are evenly spread with equal distances over the whole image of the solution set (see also [19]). Thus our aim is to generate equidistant points in the value space. For this we use a parameter dependent scalarization approach by Pascoletti and Serafini [45] and control the choice of the parameters adaptively.

A common concept for minimality in the multiobjective context is Edgeworth–Pareto- (EP-) minimality based on the natural ordering defined by the ordering cone $\boldsymbol{R}^m_+ := \{x \in \boldsymbol{R}^m \mid x_i \geq 0,\ i = 1, \ldots, m\}$. Using arbitrary partial orderings minimality is defined similarly (see, e.g., [4, 29, 31, 49, 58]). Allowing this, preference structures can be mapped which cannot be formulated explicitly as an objective function; see [55,

*Received by the editors October 11, 2006; accepted for publication (in revised form) August 26, 2008; published electronically January 28, 2009.
http://www.siam.org/journals/siopt/19-4/67202.html
†Department of Mathematics, University of Erlangen-Nürnberg, Martensstr. 3, 91058 Erlangen, Germany (Gabriele.Eichfelder@am.uni-erlangen.de).

Example 4.1]. In decision theory and in economics, arbitrary partial orderings are a well-known tool to model the relative importance of several criteria or to incorporate groups of decision makers, as promoted, for instance, by Wiecek [56]. For example, cones being a superset of the positive orthant can be defined by allowable trade-offs between the objectives or by grouping objectives according to their importance. This provides a more useful representation of the decision makers' preferences than the standard cone because the set of efficient points is reduced by undesired solutions.

For example, in [26, 27] convex polyhedral cones are used for modeling the preferences of a d.m. based on trade-off information facilitating multicriteria decision making. In portfolio optimization [2] polyhedral cones as well as nonfinitely generated cones are considered. Besides, orderings, other than the natural ordering, are important in [20] where a scalar bilevel optimization problem is reformulated as a multiobjective problem. There a nonconvex cone that is the union of two convex cones is used. In [13] a multiobjective optimization problem w.r.t. a cone $K = \boldsymbol{R}^m_+ \times \{0_n\}$ is considered for solving multiobjective bilevel optimization problems. Helbig [24] constructs various cones as a tool for finding EP-minimal points; see also [37, 51]. In addition to that, Wu [57] considers convex cones for a solution concept in fuzzy multiobjective optimization. Hence, multiobjective optimization problems w.r.t. arbitrary partial orderings are essential in decision making and are further an important tool in other areas. Therefore we develop our results w.r.t. more general partial orderings defined by closed pointed convex cones.

In the remainder we proceed as follows: in section 2 we recall the basic concepts in multiobjective optimization. In section 3 we discuss the well-known scalarization approach by Pascoletti and Serafini and we give some properties of this approach. We choose this scalarization because it is very general in the sense that many other scalarizations can be seen as a special case of it; see section 7 and [15]. In section 4 we present our main sensitivity theorem on which we base our new adaptive method in section 5. In section 6 this is applied to some test problems and to a problem in intensity modulated radiotherapy in medicine. We conclude with some remarks on the presented scalarization approach and on the transferability of the given procedure to other scalarization approaches in section 7.

2. Basic notations and concepts. We consider multiobjective optimization problems formally defined by

$$\begin{aligned} &\min_K f(x) = (f_1(x), \dots, f_m(x))^\top \\ &\quad \text{subject to the constraint} \\ &\quad x \in \Omega \subset \boldsymbol{R}^n. \end{aligned} \tag{2.1}$$

Here, K represents the considered partial ordering defined later.

We assume the following:

Assumption 1. Let C be a closed convex cone in $\boldsymbol{R}^p$, $\hat{S} \subset \boldsymbol{R}^n$ a nonempty open subset, and $S \subset \hat{S}$ closed and convex. Let the functions $f\colon \hat{S} \to \boldsymbol{R}^m$, $g\colon \hat{S} \to \boldsymbol{R}^p$, and $h\colon \hat{S} \to \boldsymbol{R}^q$ $(m, n \in \boldsymbol{N},\ p, q \in \boldsymbol{N}_0,\ m \geq 2)$ be continuously differentiable on $\hat{S}$. Let the set $\Omega \subset \boldsymbol{R}^n$, given by $\Omega = \{x \in S \mid g(x) \in C,\ h(x) = 0_q\}$, be compact.

A *convex cone* $C \subset \boldsymbol{R}^p$ is a subset of $\boldsymbol{R}^p$ with the property $\lambda(x + y) \in C$ for all $\lambda \geq 0,\ x, y \in C$. For defining minimality we need a partial ordering "$\leq$" in the image space $\boldsymbol{R}^m$. Here we mean by a partial ordering a binary relation which is reflexive, transitive, and compatible with addition and with nonnegative scalar multiplication. Any partial ordering $\leq$ defines a convex cone K by $K := \{x \in \boldsymbol{R}^m \mid 0_m \leq x\}$ and any convex cone $K \subset \boldsymbol{R}^m$, then called *ordering cone*, defines a partial ordering by

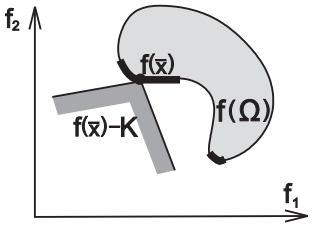


FIG. 2.1. *Illustration of Definition* 2.1 *for* $m = 2$. *The thick part of the boundary of* $f(\Omega)$ *denotes the set* $\mathcal{E}(f(\Omega), K)$.

$\leq_K := \{(x, y) \in \boldsymbol{R}^m \times \boldsymbol{R}^m \mid y - x \in K\}$. For example, the natural ordering is defined by the cone $\boldsymbol{R}^m_+$. The partial ordering is antisymmetric if the related ordering cone is *pointed*, i.e., $K \cap (-K) = \{0_m\}$. Here we consider only partial orderings defined by closed pointed convex cones.

Assumption 2. Let Assumption 1 hold. In addition let $K \subset \boldsymbol{R}^m$ be a closed pointed convex cone.

Minimality is then defined by (see, among others, [29, 31, 49, 58]):

DEFINITION 2.1. *Let* K *be a closed pointed convex cone. A point* $\bar{x} \in \Omega$ *is called* K*-minimal of* (2.1) *if* $(f(\bar{x}) - K) \cap f(\Omega) = \{f(\bar{x})\}$. *Additionally for* $int(K) \neq \emptyset$ *a point* $\bar{x} \in \Omega$ *is called weakly* K*-minimal of* (2.1) *if* $(f(\bar{x}) - int(K)) \cap f(\Omega) = \emptyset$.

For an illustration of this definition see Figure 2.1.

We denote the set of all K-minimal points by $\mathcal{M}(f(\Omega), K)$ and the set of all weakly K-minimal points by $\mathcal{M}_w(f(\Omega), K)$. The set $\mathcal{E}(f(\Omega), K) := \{f(x) \in \boldsymbol{R}^m \mid x \in \mathcal{M}(f(\Omega), K)\}$ is called *efficient set* (see Figure 2.1) and the set $\mathcal{E}_w(f(\Omega), K) := \{f(x) \in \boldsymbol{R}^m \mid x \in \mathcal{M}_w(f(\Omega), K)\}$ *weakly efficient set*. For $K = \boldsymbol{R}^m_+$ the K-minimal points are denoted as *Edgeworth–Pareto (EP)-minimal* points, too.

Later on we need that in the bicriteria case ($m = 2$) every ordering cone is finitely generated. This is stated in the following lemma. The proof is omitted here and the interested reader is referred to [13]. For the definition of a finitely generated cone (or polyhedral cone) see [46, Definition 2.17, 2.18].

LEMMA 2.2. *Let* $K \subset \boldsymbol{R}^2$ *be a closed pointed convex cone with* $K \neq \{0_2\}$. *Then* K *is polyhedral and there is either a* $k \in \boldsymbol{R}^2 \setminus \{0_2\}$ *with* $K = \{\lambda k \mid \lambda \geq 0\}$ *or there are* $l^1, l^2 \in \boldsymbol{R}^2 \setminus \{0_2\}$, l^1, l^2 *linearly independent, and* $\tilde{l}^1, \tilde{l}^2 \in \boldsymbol{R}^2 \setminus \{0_2\}$, $\tilde{l}^1, \tilde{l}^2$ *linearly independent, with*

(2.2)
$$K = \left\{y \in \boldsymbol{R}^2 \mid {l^1}^\top y \geq 0,\ {l^2}^\top y \geq 0\right\} = \left\{y \in \boldsymbol{R}^2 \mid y = \lambda^1 \tilde{l}^1 + \lambda^2 \tilde{l}^2,\ \lambda^1, \lambda^2 \geq 0\right\}.$$

In general, Lemma 2.2 is not true for $m \geq 3$. This is illustrated by the ice-cream cone $\{(x_1, x_2, x_3) \in \boldsymbol{R}^3 \mid x_1 \geq \sqrt{x_2^2 + x_3^2}\}$ which is not polyhedral.

3. Scalarization approach. For determining minimal solutions of the multiobjective optimization problem (2.1) a common approach is the scalarization of the problem. We examine the scalarization problem by Pascoletti and Serafini [45] named (SP(a, r)) which is defined by

(SP(a, r))
$$\begin{array}{c} \min t \\ \text{subject to the constraints} \\ a + t\, r - f(x) \in K, \\ x \in \Omega, \quad t \in \boldsymbol{R} \end{array}$$

for parameters $a, r \in \boldsymbol{R}^m$. Properties of this scalarization approach can be found in [13, 45]. Here we concentrate on some major results only.

THEOREM 3.1.

(a) *Let $\bar{x}$ be a K-minimal point of* (2.1)*, then $(0, \bar{x})$ is a global minimal solution of* (SP(a, r)) *with $a = f(\bar{x})$, $r \in K \setminus \{0_m\}$.*

(b) *Let $(\bar{t}, \bar{x})$ be a global minimal solution of* (SP(a, r))*, then $\bar{x}$ is a weakly K-minimal solution of* (2.1)*.*

Theorem 3.1 is given for global minimal solutions, but the statements can easily be adapted to local minimal solutions (for a continuous function f) too; see [13, 45]. Theorem 3.1 (a) yields that any K-minimal solution of the multiobjective optimization problem can be found by solving the scalar problem (SP(a, r)) for appropriate parameters, even for nonconvex problems. This is not the case for any scalarization approach as, e.g., the weighted sum method [59] shows. In the following Theorems 3.2, 3.3, and 3.5 we even see that it is sufficient to consider only variations of the parameter a in (a subset of) a hyperplane for a fixed chosen parameter $r \in K \setminus \{0_m\}$. Then it is still possible that any minimal solution of the multiobjective optimization can be recovered.

THEOREM 3.2. *Let $\bar{x} \in \mathcal{M}(f(\Omega), K)$ and $r \in K$ be given. We define a hyperplane H by $H = \{y \in \boldsymbol{R}^m \mid b^\top y = \beta\}$ with $b \in \boldsymbol{R}^m \setminus \{0_m\}$, $b^\top r \neq 0$, $\beta \in \boldsymbol{R}$. Then there is a parameter $a \in H$ and some $\bar{t} \in \boldsymbol{R}$ so that $(\bar{t}, \bar{x})$ is a minimal solution of* (SP(a, r))*.*

Proof. We set $\bar{t} = (b^\top f(\bar{x}) - \beta)/(b^\top r)$ and $a = f(\bar{x}) - \bar{t}\, r$. Then $a \in H$ and $(\bar{t}, \bar{x})$ is feasible for (SP(a, r)). We assume $(\bar{t}, \bar{x})$ is not a minimal solution of (SP(a, r)). Then there are $t' \in \boldsymbol{R}$, $t' < \bar{t}$, $x' \in \Omega$, and $k' \in K$ so that $a + t'\, r - f(x') = k'$. Using the definition of the parameter a, this results in $f(\bar{x}) = f(x') + k' + (\bar{t} - t')r$. Since K is a pointed convex cone, $r \in K \setminus \{0_m\}$ and $\bar{t} - t' > 0$ we have $f(\bar{x}) \in f(x') + K \setminus \{0_m\}$ for $x' \in \Omega$ which is a contradiction to $\bar{x}$ K-minimal. ∎

We can restrict the parameter set even further. First we consider the bicriteria case ($m = 2$). Then, according to Lemma 2.2, every closed pointed convex cone is finitely generated. We suppose the cone K is given by (2.2). (The more trivial case $K = \{\lambda k \mid \lambda \geq 0\}$ for some $k \in \boldsymbol{R}^2 \setminus \{0_2\}$ can be handled similarly, but we will not consider this less interesting case here. For more details see [13, pp. 73f]). Next we solve the scalar optimization problems

$$\min_{x \in \Omega} l^{i\top} f(x), \qquad i = 1, 2, \tag{3.1}$$

with minimal solutions $\bar{x}^i$, $i = 1, 2$. Then the points $\bar{x}^i$, $i = 1, 2$, are weakly K-minimal and we can easily show that for every K-minimal point x of (2.1) we have

$$l^{1\top} f\left(\bar{x}^1\right) \leq l^{1\top} f(x) \leq l^{1\top} f\left(\bar{x}^2\right) \quad \text{and} \quad l^{2\top} f\left(\bar{x}^2\right) \leq l^{2\top} f(x) \leq l^{2\top} f\left(\bar{x}^1\right). \tag{3.2}$$

Using (3.2), $l^{1\top} f(\bar{x}^1) = l^{1\top} f(\bar{x}^2)$ implies $l^{1\top} f(x) = l^{1\top} f(\bar{x}^2)$ and, consequently, $f(x) \in f(\bar{x}^2) + K$ for all $x \in \mathcal{M}(f(\Omega), K)$ resulting in $\mathcal{E}(f(\Omega), K) = \{f(\bar{x}^2)\}$. The same, $l^{2\top} f(\bar{x}^2) = l^{2\top} f(\bar{x}^1)$, leads to $\mathcal{E}(f(\Omega), K) = \{f(\bar{x}^1)\}$. Thus, assuming the efficient set does not consist of one point only, we have

$$l^{1\top} f\left(\bar{x}^1\right) < l^{1\top} f\left(\bar{x}^2\right) \quad \text{and} \quad l^{2\top} f\left(\bar{x}^2\right) < l^{2\top} f\left(\bar{x}^1\right). \tag{3.3}$$

If we define the points $\bar{a}^i$, $i = 1, 2$, by a projection of the points $f(\bar{x}^i)$, $i = 1, 2$, in direction $r \in K$ on the hyperplane H, then we get

$$\bar{a}^i := f\left(\bar{x}^i\right) - \bar{t}^i\, r \in H \quad \text{with} \quad \bar{t}^i := \frac{b^\top f\left(\bar{x}^i\right) - \beta}{b^\top r}, \qquad i = 1, 2. \tag{3.4}$$

In the next theorem we see that we can restrict ourselves to the set $H^a := \{y \in \boldsymbol{R}^2 \mid y = \lambda\bar{a}^1 + (1-\lambda)\bar{a}^2,\ \lambda \in [0,1]\}$ for choosing the parameter a.

THEOREM 3.3. *We consider the multiobjective optimization problem* (2.1) *with* $m = 2$ *and* K *as in* (2.2). *Further let* $\bar{a}^i,\ i = 1,2$, *be defined as in* (3.4) *with* $\bar{x}^i$, $i = 1,2$, *minimal solutions of* (3.1) *and assume* $\bar{x} \in \mathcal{M}(f(\Omega), K)$. *Then there is a parameter* $a \in H^a \subset H$ *and some* $\bar{t} \in \boldsymbol{R}$ *so that* $(\bar{t}, \bar{x})$ *is a minimal solution of* (SP(a,r)).

Proof. Notice that $\bar{a}^1, \bar{a}^2 \in H$ and, hence, $H^a \subset H$. According to Theorem 3.2 we have for any $\bar{x} \in \mathcal{M}(f(\Omega), K)$ a parameter $a \in H$ and some $\bar{t} \in \boldsymbol{R}$ so that $(\bar{t}, \bar{x})$ is a minimal solution of (SP(a,r)). This is achieved by $\bar{t} = (b^\top f(\bar{x}) - \beta)/(b^\top r)$ and $a = f(\bar{x}) - \bar{t}\, r$. Hence, it suffices to show $a = \lambda\bar{a}^1 + (1-\lambda)\bar{a}^2$ for some $\lambda \in [0,1]$. Using the definitions of a, $\bar{a}^1$, and $\bar{a}^2$, this equation can be written as

$$f(\bar{x}) - \bar{t}\, r = \lambda \left(f\left(\bar{x}^1\right) - \bar{t}^1\, r\right) + (1-\lambda)\left(f\left(\bar{x}^2\right) - \bar{t}^2\, r\right). \tag{3.5}$$

If the efficient set consists of one point only (and then this point is $f(\bar{x}^1)$ or $f(\bar{x}^2)$) (3.5) is fulfilled for $\lambda = 1$ or $\lambda = 0$. Otherwise, the strict inequalities (3.3) hold. Equation (3.5) can further be written as

$$f(\bar{x}) = \lambda\, f\left(\bar{x}^1\right) + (1-\lambda)\, f\left(\bar{x}^2\right) + \left(\bar{t} - \lambda\, \bar{t}^1 - (1-\lambda)\, \bar{t}^2\right) r, \tag{3.6}$$

and we differentiate the following two cases: $(\bar{t} - \lambda\, \bar{t}^1 - (1-\lambda)\, \bar{t}^2) = \frac{1}{b^\top r}(b^\top(f(\bar{x}) - \lambda\, f(\bar{x}^1) - (1-\lambda)\, f(\bar{x}^2)) \geq 0$ and $(\bar{t} - \lambda\, \bar{t}^1 - (1-\lambda)\, \bar{t}^2) < 0$.

For $\bar{t} - \lambda\, \bar{t}^1 - (1-\lambda)\, \bar{t}^2 \geq 0$ we suppose (3.6) is satisfied for $\lambda < 0$. Applying the linear map l^1 on (3.6) results, together with $r \in K$ and (3.3), in the following:

$$\begin{aligned} l^{1\top} f(\bar{x}) &= \lambda\, l^{1\top} f\left(\bar{x}^1\right) + (1-\lambda)\, l^{1\top} f\left(\bar{x}^2\right) + \underbrace{\left(\bar{t} - \lambda\, \bar{t}^1 - (1-\lambda)\, \bar{t}^2\right)}_{\geq 0} \underbrace{l^{1\top} r}_{\geq 0} \\ &\geq \underbrace{\lambda}_{<0} \underbrace{l^{1\top} f\left(\bar{x}^1\right)}_{<l^{1\top} f(\bar{x}^2)} + (1-\lambda) l^{1\top} f\left(\bar{x}^2\right) \\ &> \lambda\, l^{1\top} f\left(\bar{x}^2\right) + (1-\lambda)\, l^{1\top} f\left(\bar{x}^2\right) = l^{1\top} f\left(\bar{x}^2\right), \end{aligned}$$

which is a contradiction to (3.2).

Instead, assuming (3.6) is satisfied for $\lambda > 1$, we get by applying l^2 on (3.6) together with (3.3)

$$l^{2\top} f(\bar{x}) \geq \lambda\, l^{2\top} f\left(\bar{x}^1\right) + \underbrace{(1-\lambda)}_{<0} \underbrace{l^{2\top} f\left(\bar{x}^2\right)}_{<l^{2\top} f(\bar{x}^1)} > l^{2\top} f\left(\bar{x}^1\right)$$

in contradiction to (3.2). Hence, we have $\lambda \in [0,1]$ for $\bar{t} - \lambda\, \bar{t}^1 - (1-\lambda)\, \bar{t}^2 \geq 0$.

In the same way we show $\lambda \in [0,1]$ for $\bar{t} - \lambda\, \bar{t}^1 - (1-\lambda)\, \bar{t}^2 < 0$, too, and the assertion of the theorem is proven. ∎

A generalization to the case $m \geq 3$ is not possible because then a cone need not be finitely generated as we have seen. Even if the cone is finitely generated, even if it is the positive orthant, the previous results are not true in general for more than two objectives as the following example shows.

Example 3.4. We consider the function $f\colon \boldsymbol{R}^3 \to \boldsymbol{R}^3$ with $f(x) = x$ for all $x \in \boldsymbol{R}^3$ and the set $\Omega = \{x \in \boldsymbol{R}^3 \mid x_1^2 + x_2^2 + x_3^2 \leq 1\}$ representing the unit ball in $\boldsymbol{R}^3$. Let the ordering cone be the natural ordering cone $K = \boldsymbol{R}^3_+$ finitely generated by $l^1 = (1,0,0)^\top$, $l^2 = (0,1,0)^\top$, and $l^3 = (0,0,1)^\top$. The multiobjective optimiz-

ation problem $\min_{x\in\Omega} f(x)$ has the minimal solution set $\mathcal{M}(f(\Omega), \boldsymbol{R}^3_+) = \{x \in \boldsymbol{R}^3 \mid x_1^2 + x_2^2 + x_3^2 = 1,\ x_i \le 0,\ i = 1,2,3\}$. By solving the scalar optimization problems $\min_{x\in\Omega} l^{i\top} f(x),\ i = 1,2,3$ (compare (3.1)), we get the minimal solutions $\bar{x}^1 = (-1,0,0)^\top,\ \bar{x}^2 = (0,-1,0)^\top$, and $\bar{x}^3 = (0,0,-1)^\top$. If we consider now the hyperplane $H = \{y \in \boldsymbol{R}^3 \mid (-1,-1,-1)\cdot y = 1\}$ we have $f(\bar{x}^i) = \bar{x}^i \in H$ for $i = 1,2,3$.

Choosing $r = (1,1,1)^\top$ and determining the points $\bar{a}^i \in H$, $i = 1,2,3$, as in (3.4) results in $\bar{a}^1 = (-1,0,0)^\top,\ \bar{a}^2 = (0,-1,0)^\top,\ \bar{a}^3 = (0,0,-1)^\top$ and, hence, in the set H^a consisting of all convex combinations of the points $\bar{a}^i,\ i = 1,2,3$. In this example the point $\bar{x} = (-1/\sqrt{2}, -1/\sqrt{2}, 0)^\top$ is EP-minimal, but there is no parameter $\bar{a} \in H^a$ such that $\bar{x}$ is a minimal solution of $(\mathrm{SP}(\bar{a}, r))$. For $\bar{a} = -1/(3\sqrt{2})\cdot(1+\sqrt{2}, 1+\sqrt{2}, \sqrt{2}-2)^\top$ and $\bar{t} = (1-\sqrt{2})/3$ the point $(\bar{t}, \bar{x})$ is a minimal solution of $(\mathrm{SP}(\bar{a}, r))$ but it is $\bar{a} \notin H^a$.

Similar considerations have been done in [9, pp. 635f] in connection with the normal boundary intersection method. Nevertheless we want to restrain the set H from which we choose the parameters a for the case of more than two objectives. For this aim we project the set $f(\Omega)$ in the direction r into the set H and determine the set $\tilde{H} := \{y \in H \mid y + t\,r = f(x),\ t \in \boldsymbol{R},\ x \in \Omega\} \subset H$. Of course we would get a stricter limitation by projecting the set $\mathcal{E}(f(\Omega), K)$ instead of $f(\Omega)$, but in general the efficient set is not known in advance. Because the set $\tilde{H} \subset H$ has usually an irregular shape, which complicates a methodic procedure, we embed the set $\tilde{H}$ in the image of an $(m-1)$-dimensional hyperplane under a linear transformation $H^0 \subset \boldsymbol{R}^m$, which we attempt to choose minimally. For doing this we first determine $m-1$ orthogonal vectors $v^1, \dots, v^{m-1}$ spanning the hyperplane H with $\tilde{H} \subset H$. Hence, we have

$$H = \left\{ y \in \boldsymbol{R}^m \mid y = \sum_{i=1}^{m-1} s_i v^i,\ s \in \boldsymbol{R}^{m-1} \right\}. \tag{3.7}$$

Next we solve the following $2(m-1)$ scalar optimization problems

$$\begin{array}{c} \min s_j \\ \text{subject to the constraints} \\ \sum_{i=1}^{m-1} s_i v^i + t\,r = f(x), \\ t \in \boldsymbol{R},\ x \in \Omega,\ s \in \boldsymbol{R}^{m-1}, \end{array} \quad \text{and} \quad \begin{array}{c} \min -s_j \\ \text{subject to the constraints} \\ \sum_{i=1}^{m-1} s_i v^i + t\,r = f(x), \\ t \in \boldsymbol{R},\ x \in \Omega,\ s \in \boldsymbol{R}^{m-1} \end{array} \tag{3.8}$$

for $j \in \{1, \dots, m-1\}$ with minimal solutions $(t^{\min,j}, x^{\min,j}, s^{\min,j})$ and minimal values $s_j^{\min,j}$ and $(t^{\max,j}, x^{\max,j}, s^{\max,j})$ and minimal values $-s_j^{\max,j}$, respectively. These optimization problems are generally nonconvex even if the related multiobjective optimization problem is convex. However, note that it suffices to provide lower bounds for the optimal function values for the following results. Then we obtain the set H^0 by $H^0 := \{y \in \boldsymbol{R}^m \mid y = \sum_{i=1}^{m-1} s_i v^i,\ s_i \in [s_i^{\min,i}, s_i^{\max,i}],\ i = 1, \dots, m-1\}$. The set H^0 includes the set $\tilde{H}$ and is calculated numerically as small as possible.

THEOREM 3.5. *Let $\bar{x} \in \mathcal{M}(f(\Omega), K)$. Then there is a parameter $\bar{a} \in H^0$ and some $\bar{t} \in \boldsymbol{R}$ so that $(\bar{t}, \bar{x})$ is a minimal solution of* $(\mathrm{SP}(\bar{a}, r))$.

Proof. According to the proof of Theorem 3.2 we have for $\bar{t} = (b^\top f(\bar{x}) - \beta)/(b^\top r)$ and $\bar{a} = f(\bar{x}) - \bar{t}\,r$ that $(\bar{t}, \bar{x})$ is a minimal solution of $(\mathrm{SP}(\bar{a}, r))$ with $\bar{a} \in H$. Because $H^0 \subset H$ it suffices to show $\bar{a} \in H^0$. Because $\bar{a} \in H$ there is, according to the representation in (3.7), a vector $\bar{s} \in \boldsymbol{R}^{m-1}$ with $\bar{a} = \sum_{i=1}^{m-1} \bar{s}_i v^i$. Because of $\bar{a} + \bar{t}\,r =$

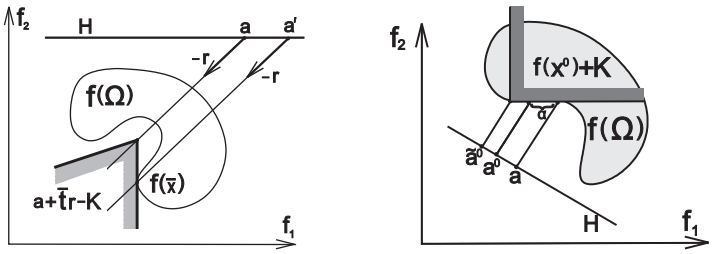


FIG. 3.1. (a) *Visualization of Theorem* 3.6. (b) *Choosing the new parameter in the case* $a^0 + t^0\, r - f(x^0) \neq 0_2$ *and* $\tilde{a}_1^0 < a_1^0$.

$f(\bar{x})$, the point $(\bar{t}, \bar{x}, \bar{s})$ is feasible for the optimization problems (3.8) and thus we have $s_i^{\min,i} \leq \bar{s}_i \leq s_i^{\max,i}$ for $i = 1, \ldots, m-1$. Hence, it follows $\bar{a} \in H^0$. ☐

Thus, it is sufficient to consider only parameters $a \in H^0$ for a fixed $r \in K \setminus \{0_m\}$ to find all K-minimal solutions of the multiobjective optimization problem (2.1).

For our considerations in the next section it will be important that in the minimal solution $(\bar{t}, \bar{x})$ the constraint $a + t\, r - f(x) \in K$ is active; i.e., $a + \bar{t}\, r - f(\bar{x}) = 0_m$. If this is not the case for the choice of a parameter a, we can easily define a new parameter a' for which this property is satisfied (see Figure 3.1(a)).

THEOREM 3.6. *Let the hyperplane* $H = \{y \in \boldsymbol{R}^m \mid b^\top y = \beta\}$ *with* $b \in \boldsymbol{R}^m \setminus \{0_m\}$, $\beta \in \boldsymbol{R}$ *be given. Suppose* $(\bar{t}, \bar{x})$ *is a minimal solution of* $(\mathrm{SP}(a, r))$ *for* $a \in H$, $r \in K$, $b^\top r \neq 0$. *Then there is a* $\bar{k} \in K$ *with* $a + \bar{t}\, r - f(\bar{x}) = \bar{k}$. *However, then there is a point* $a' \in H$ *and some* $t' \in \boldsymbol{R}$ *so that* $(t', \bar{x})$ *is also a minimal solution of* $(\mathrm{SP}(a', r))$ *with* $a' + t'\, r - f(\bar{x}) = 0_m$.

Proof. We set $t' := (b^\top f(\bar{x}) - \beta)/(b^\top r)$ and $a' := a + (\bar{t} - t')\, r - \bar{k} = f(\bar{x}) - t'\, r$. Then $a' \in H$ and $a' + t'\, r - f(\bar{x}) = 0_m$. Next we show that $(t', \bar{x})$ is a minimal solution of $(\mathrm{SP}(a', r))$. Otherwise there is a feasible point $(\hat{t}, \hat{x}) \in \boldsymbol{R} \times \Omega$ with $\hat{t} < t'$ and there is a $\hat{k} \in K$ with $a' + \hat{t}\, r - f(\hat{x}) = \hat{k}$. Together with the definition of a' this leads to $a + (\bar{t} - t' + \hat{t})\, r - f(\hat{x}) = \hat{k} + \bar{k} \in K$. However, then $(\bar{t} - t' + \hat{t}, \hat{x})$ is feasible for $(\mathrm{SP}(a, r))$ with $\bar{t} - t' + \hat{t} < \bar{t}$ in contradiction to $(\bar{t}, \bar{x})$ a minimal solution of $(\mathrm{SP}(a, r))$. ☐

It can also be shown that there is no K-minimal point $\tilde{x}$ for which $\tilde{a} = a + \lambda\,(a' - a)$ for some $\lambda \in]0, 1[$ with $\tilde{a} := f(\tilde{x}) - \tilde{t}\, r$ and $\tilde{t} := (b^\top f(\tilde{x}) - \beta)/(b^\top r)$ [13, Theorem 4.2.12]. Thus the section on the hyperplane H between the parameters a and a' can be neglected. According to Theorem 3.6 we have for the parameter a' not only $a' \in H$ but also $a' + t'\, r = f(\bar{x})$. Hence, for the set H^0 as in Theorem 3.5, it holds $a' \in H^0$.

Under the Assumption 2 it can be shown that if a minimal solution $(\bar{t}, \bar{x})$ of $(\mathrm{SP}(a, r))$ has Lagrange multipliers $(\mu, \nu, \xi) \in K^* \times C^* \times \boldsymbol{R}^q$, then these are also Lagrange multipliers to the point $(t', \bar{x})$ for $(\mathrm{SP}(a', r))$ under the transformation of Theorem 3.6. For the definition of the Lagrange function and the Lagrange multipliers see, for instance, [32, p. 115]. Here, $\mu \in K^*$ is the Lagrange multiplier to the constraint $a + t\, r - f(x) \in K$ (and thus also to $a' + t\, r - f(x) \in K$), $\nu \in C^*$ corresponds to the constraint $g(x) \in C$, and $\xi \in \boldsymbol{R}^q$ to the equality constraint $h(x) = 0_q$.

4. Sensitivity results. The main point of our adaptive parameter control for the parameter dependent scalarization approach $(\mathrm{SP}(a, r))$ is a sensitivity theorem based on theorems by Alt [3, Theorem 5.3, 6.1]. For two reasons we apply these

theorems on a modified version of problem $(\mathrm{SP}(a,r))$ called $(\overline{\mathrm{SP}}(a,r))$:

$$(\overline{\mathrm{SP}}(a,r)) \qquad \begin{array}{c} \min t \\ \text{subject to the constraints} \\ a + t\,r - f(x) = 0_m, \\ t \in \boldsymbol{R},\ x \in \Omega \end{array}$$

with the constraint set $\overline{\Sigma}(a,r) := \{(t,x) \in \boldsymbol{R} \times \boldsymbol{R}^n \mid a + t\,r - f(x) = 0_m,\ x \in \Omega\}$.

First, by getting sensitivity results for the minimal value $t = t(a,r)$ of the problem $(\overline{\mathrm{SP}}(a,r))$ we can at once conclude how a variation of the parameters influences the generated points $f(x(a,r))$ by using the equality constraint $a + t(a,r)\,r - f(x(a,r)) = 0_m$. Second, for applying the sensitivity theorems by Alt directly to the problem $(\mathrm{SP}(a,r))$, for the Hessian of the Lagrange function $\mathcal{L}(t,x,\mu,\nu,\xi,a,r) = t - \mu^\top(a + t\,r - f(x)) - \nu^\top g(x) - \xi^\top h(x)$ w.r.t. the variables (t,x)

$$\nabla^2_{(t,x)}\mathcal{L}(t,x,\mu,\nu,\xi,a,r) = \begin{pmatrix} 0 & 0 \\ 0 & W(x,\mu,\nu,\xi) \end{pmatrix}$$

with $W(x,\mu,\nu,\xi) = \sum_{i=1}^m \mu_i \nabla^2 f_i(x) - \sum_{j=1}^p \nu_j \nabla^2 g_j(x) - \sum_{k=1}^q \xi_k \nabla^2 h_k(x)$ the assumption

$$\left(t, x^\top\right) \nabla^2_{(t,x)}\mathcal{L}(t,x,\mu,\nu,\xi,a,r) \begin{pmatrix} t \\ x \end{pmatrix} \geq \alpha \left\| \begin{pmatrix} t \\ x \end{pmatrix} \right\|^2$$

has to be satisfied for some constant $\alpha > 0$ for all $(t,x) \in \boldsymbol{R}^{n+1}$ with $\nabla h(x)x = 0_q$. This is always contradicted by the points $(t,x) = (t,0_n)$ with $t \neq 0$.

Problem $(\overline{\mathrm{SP}}(a,r))$ is in general nonconvex, even if the original multiobjective optimization problem is convex. However, note that the problem $(\overline{\mathrm{SP}}(a,r))$ is never actually solved. We solve only the problems $(\mathrm{SP}(a,r))$ (which are convex if the original multiobjective optimization problem is convex) and we use the problem $(\overline{\mathrm{SP}}(a,r))$ only for approximating the minimal value and the points $f(x(a,r))$ of the problem $(\mathrm{SP}(a,r))$.

We obtain the connection between the problems $(\mathrm{SP}(a,r))$ and $(\overline{\mathrm{SP}}(a,r))$ by Theorem 3.6: if we solve the problem $(\mathrm{SP}(a,r))$ with a minimal solution $(\bar{t},\bar{x})$ and with $a + \bar{t}\,r - f(\bar{x}) = \bar{k} \in K$, then there always exists a parameter a' and some t' so that $(t',\bar{x})$ is minimal for $(\mathrm{SP}(a',r))$ with $a' + t'\,r - f(\bar{x}) = 0_m$ and then $(t',\bar{x})$ is a minimal solution of $(\overline{\mathrm{SP}}(a',r))$ too. We examine now the dependence of the minimal values of the problem $(\overline{\mathrm{SP}}(a',r))$ on the parameter a' (in Theorem 4.2) and from that we conclude on the (approximated) dependence of the minimal value of the problem $(\mathrm{SP}(a',r))$ on the parameter a' (in section 5.1). Before we come to the main sensitivity result we need the following assumption and the following lemma.

Assumption 3. Let Assumption 2 hold. In addition let the functions f, g, and h be twice continuously differentiable on $\hat{S}$.

LEMMA 4.1. *Let Assumption* 3 *hold. Let* (t^0,x^0) *be a local minimal solution of* $(\overline{\mathrm{SP}}(a^0,r^0))$ *with Lagrange multipliers* $(\mu^0,\nu^0,\xi^0) \in \boldsymbol{R}^m \times C^* \times \boldsymbol{R}^q$. *Assume there exists some constant* $\tilde{\alpha} > 0$ *such that for the matrix* $W(x^0,\mu^0,\nu^0,\xi^0) = \mu^{0\top}\nabla^2 f(x^0) - \nu^{0\top}\nabla^2 g(x^0) - \xi^{0\top}\nabla^2 h(x^0)$ *we have*

$$x^\top W\left(x^0,\mu^0,\nu^0,\xi^0\right) x \geq \tilde{\alpha}\,\|x\|^2 \tag{4.1}$$

for all $x \in \{x \in \boldsymbol{R}^n \mid \nabla h(x^0)x = 0_q,\ \nabla f(x^0)x = r^0\,t \ \text{ for a } t \in \boldsymbol{R}\}$. *Then there exists*

some constant $\alpha > 0$ such that for the Lagrange function $\overline{\mathcal{L}}$ to $(\overline{\mathrm{SP}}(a,r))$ we have

$$\left(t, x^\top\right) \nabla^2_{(t,x)} \overline{\mathcal{L}}\left(t^0, x^0, \mu^0, \nu^0, \xi^0, a^0, r^0\right) \begin{pmatrix} t \\ x \end{pmatrix} \geq \alpha \left\| \begin{pmatrix} t \\ x \end{pmatrix} \right\|^2 \tag{4.2}$$

for all $(t,x) \in \{(t,x) \in \boldsymbol{R} \times \boldsymbol{R}^n \mid \nabla h(x^0)x = 0_q,\ \nabla f(x^0)x = r^0\, t\}$.

Proof. Because (t^0, x^0) is a local minimal solution of $(\overline{\mathrm{SP}}(a^0, r^0))$ with Lagrange multipliers (μ^0, ν^0, ξ^0) we have for the associated Lagrange function

$$\nabla_{(t,x)} \overline{\mathcal{L}}\left(t^0, x^0, \mu^0, \nu^0, \xi^0, a^0, r^0\right)^\top \begin{pmatrix} t - t^0 \\ x - x^0 \end{pmatrix} \geq 0 \text{ for all } t \in \boldsymbol{R},\ x \in S. \tag{4.3}$$

With $\frac{\partial \overline{\mathcal{L}}(t^0, x^0, \mu^0, \nu^0, \xi^0, a^0, r^0)}{\partial t} = 1 - \mu^{0\top} r^0$ and because (4.3) has to be fulfilled for all $t \in \boldsymbol{R}$ we have

$$\mu^{0\top} r^0 = 1 \tag{4.4}$$

and, therefore, $\mu^0 \neq 0_m$, $r^0 \neq 0_m$. Because in $\boldsymbol{R}^n$ and $\boldsymbol{R}^m$, respectively, all norms are equivalent, there exist positive constants M^l, $M^u \in \boldsymbol{R}_+$ and $\tilde{M}^l$, $\tilde{M}^u \in \boldsymbol{R}_+$, respectively, with $M^l \|x\|_2 \leq \|x\| \leq M^u \|x\|_2$ and

$$\tilde{M}^l \left\| \begin{pmatrix} t \\ x \end{pmatrix} \right\|_2 \leq \left\| \begin{pmatrix} t \\ x \end{pmatrix} \right\| \leq \tilde{M}^u \left\| \begin{pmatrix} t \\ x \end{pmatrix} \right\|_2$$

for all $(t,x) \in \boldsymbol{R} \times \boldsymbol{R}^n$. For all $(t,x) \in \boldsymbol{R} \times \boldsymbol{R}^n$ with $\nabla f(x^0)x = r^0\, t$ we have together with (4.4) the equation $\mu^{0\top} \nabla f(x^0)x = t$ and then we get the estimation $|t|^2 = |\mu^{0\top} \nabla f(x^0)x|^2 \leq \|\mu^0\|_2^2 \, \|\nabla f(x^0)\|_2^2 \, \|x\|_2^2$. If we set now

$$\alpha := \frac{\tilde{\alpha}\ \left(M^l\right)^2}{\left(\tilde{M}^u\right)^2\ \left(1 + \|\mu^0\|_2^2 \, \|\nabla f\left(x^0\right)\|_2^2\right)} > 0,$$

we conclude from (4.1) for all $(t,x) \in \{(t,x) \in \boldsymbol{R} \times \boldsymbol{R}^n \mid \nabla h(x^0)x = 0_q,\ \nabla f(x^0)x = r^0\, t\}$

$$\begin{aligned} x^\top W\left(x^0, \mu^0, \nu^0, \xi^0\right) x &\geq \tilde{\alpha}\, \|x\|^2 \geq \tilde{a}\ \left(M^l\right)^2 \|x\|_2^2 \\ &= \alpha\ \left(\tilde{M}^u\right)^2\ \left(1 + \|\mu^0\|_2^2 \, \|\nabla f(x^0)\|_2^2\right)\ \|x\|_2^2 \\ &\geq \alpha\ \left(\tilde{M}^u\right)^2\ \left(\|x\|_2^2 + |t|^2\right) \\ &= \alpha\ \left(\tilde{M}^u\right)^2\ \left\| \begin{pmatrix} t \\ x \end{pmatrix} \right\|_2^2 \geq \alpha \left\| \begin{pmatrix} t \\ x \end{pmatrix} \right\|^2. \end{aligned}$$

With

$$\nabla^2_{(t,x)} \overline{\mathcal{L}}\left(t^0, x^0, \mu^0, \nu^0, \xi^0, a^0, r^0\right) = \begin{pmatrix} 0 & 0 \\ 0 & W\left(x^0, \mu^0, \nu^0, \xi^0\right) \end{pmatrix}$$

the assertion is proven. ∎

The condition (4.2) for all (t,x) of the given set is called strict second-order sufficient condition. If this condition is fulfilled for a regular point, then this is sufficient for strict local minimality of the considered point [41, Theorem 5.2].

THEOREM 4.2. *Let Assumption* 3 *and the assumptions of Lemma* 4.1 *hold. We consider the parametric optimization problem* $(\overline{\mathrm{SP}}(a,r))$ *with the constraint set* $\overline{\Sigma}(a,r)$ *starting with a reference problem* $(\overline{\mathrm{SP}}(a^0,r^0))$ *with a local minimal solution* (t^0,x^0) *and with Lagrange multipliers* $(\mu^0,\nu^0,\xi^0)\in \boldsymbol{R}^m\times C^*\times \boldsymbol{R}^q$.

(i) *Suppose the point* (t^0,x^0) *is regular for the set* $\overline{\Sigma}(a^0,r^0)$, *i.e., we have*

$$0_{m+p+q}\in \mathit{int}\left\{\left(\begin{array}{rcl} r^0\ \left(t-t^0\right) & - & \nabla f\left(x^0\right)\left(x-x^0\right)\\ g\left(x^0\right) & + & \nabla g\left(x^0\right)\left(x-x^0\right)-c\\ & & \nabla h\left(x^0\right)\left(x-x^0\right)\end{array}\right)\ \left|\ \begin{array}{l} c\in C,\\ x\in S,\\ t\in \boldsymbol{R}\end{array}\right.\right\}.$$

(ii) *Assume there exists some* $\zeta>0$ *such that the following holds for all* $p^1,\ p^2\in \zeta\tilde{B}$ *(with* $\tilde{B}$ *the closed unit ball in* $\boldsymbol{R}^{1+n+m+p+q}$*) with* $p^i=(t^{*i},x^{*i},u^i,v^i,w^i)$, $i=1,\,2$: *if* (t^1,x^1) *and* (t^2,x^2), *respectively, are solutions of the quadratic optimization problem*

$$\begin{array}{c}\min J\left(t,x,p^i\right)\\ \textit{subject to the constraints}\\ r^0\ \left(t-t^0\right)-\nabla f\left(x^0\right)\left(x-x^0\right)-u^i=0_m,\\ g\left(x^0\right)+\nabla g\left(x^0\right)\left(x-x^0\right)-v^i\in C,\\ \nabla h\left(x^0\right)\left(x-x^0\right)-w^i=0_q,\\ t\in \boldsymbol{R},\ x\in S,\end{array}$$

$(i=1,2)$ *with* $J(t,x,p^i):=\frac{1}{2}(x-x^0)^\top W(x^0,\mu^0,\nu^0,\xi^0)\,(x-x^0)+(t-t^0)-t^{*i}\,(t-t^0)-(x^{*i})^\top(x-x^0)$, *then the Lagrange multipliers* $(\mu_q^i,\nu_q^i,\xi_q^i)$ *to the solutions* (t^i,x^i), $i=1,\,2$, *are uniquely determined and*

$$\left\|\left(\mu_q^1,\nu_q^1,\xi_q^1\right)-\left(\mu_q^2,\nu_q^2,\xi_q^2\right)\right\|\le c_M\left(\left\|\left(t^1,x^1\right)-\left(t^2,x^2\right)\right\|+\left\|p^1-p^2\right\|\right)$$

with some constant c_M.

Then there exists some $\delta>0$ *and a neighborhood* $N(a^0,r^0)$ *of* (a^0,r^0) *so that the local minimal value function* $\overline{\tau}^\delta(a,r):=\inf\{t\mid (t,x)\in\overline{\Sigma}(a,r)\cap B_\delta(t^0,x^0)\}$ *is differentiable on* $N(a^0,r^0)$ *with the derivative*

$$\nabla_{(a,r)}\overline{\tau}^\delta(a,r)=\nabla_{(a,r)}\overline{\mathcal{L}}\left(\bar{t}(a,r),\bar{x}(a,r),\mu(a,r),\nu(a,r),\xi(a,r),a,r\right).$$

Here $(\bar{t}(a,r),\bar{x}(a,r))$ *denotes the strict local minimal solution of* $(\overline{\mathrm{SP}}(a,r))$ *for* $(a,r)\in N(a^0,r^0)$ *with the unique Lagrange multipliers* $(\mu(a,r),\nu(a,r),\xi(a,r))$. *In addition to that the mapping* $\phi\colon N(a^0,r^0)\to B_\delta(t^0,x^0)\times B_\delta(\mu^0,\nu^0,\xi^0)$ *defined by* $\phi(a,r)=(\bar{t}(a,r),\bar{x}(a,r),\mu(a,r),\nu(a,r),\xi(a,r))$ *is Lipschitzian on* $N(a^0,r^0)$.

Proof. By using Lemma 4.1 it can easily be shown that all premises for applying the Theorems 5.3 and 6.1 by Alt [3] are met. □

Remark 4.3. The condition (ii) of the preceding theorem is always satisfied if we have only equality constraints [3, Theorem 7.1] or, in the case of the natural ordering $C=\boldsymbol{R}^n_+$, if the gradients of the active constraints are linearly independent; compare [16, Theorem 2.1] and [35, Theorem 2].

LEMMA 4.4. *Let the assumptions of Theorem* 4.2 *be satisfied with* $S=\boldsymbol{R}^n$. *Then there is some* $\delta>0$ *and a neighborhood* $N(a^0,r^0)$ *of* (a^0,r^0) *so that for all* $(a,r)\in N(a^0,r^0)$ *the derivatives of the local minimal value function are given by*

$$\begin{aligned}&\nabla_a\overline{\tau}^\delta(a,r)=-\mu(a,r)-\nabla_a\nu(a,r)^\top g(\bar{x}(a,r))\\ \textit{and}\quad &\nabla_r\overline{\tau}^\delta(a,r)=-\bar{t}(a,r)\mu(a,r)-\nabla_r\nu(a,r)^\top g(\bar{x}(a,r)).\end{aligned}$$

Proof. According to Theorem 4.2 there is a neighborhood $N(a^0,r^0)$ of (a^0,r^0) such that for all $(a,r)\in N(a^0,r^0)$ there is a strict minimal solution $(\bar t(a,r),\bar x(a,r))$ with unique Lagrange multipliers $(\mu(a,r),\nu(a,r),\xi(a,r))$. Because of $S=\boldsymbol{R}^n$ we have for the derivative of the Lagrangian $\nabla_{(t,x)}\overline{\mathcal{L}}(\bar t(a,r),\bar x(a,r),\mu(a,r),\nu(a,r),\xi(a,r),a,r)=0_{n+1}$. Then it follows

$$\begin{aligned}0_m &= \nabla_a\begin{pmatrix}\bar t(a,r)\\ \bar x(a,r)\end{pmatrix}^\top \nabla_{(t,x)}\overline{\mathcal{L}}(\bar t(a,r),\bar x(a,r),\mu(a,r),\nu(a,r),\xi(a,r),a,r)\\
&= \nabla_a\bar t(a,r)-\sum_{i=1}^m\mu_i(a,r)\Big(\nabla_a\bar t(a,r)\,r_i-\nabla_a\bar x(a,r)^\top\nabla_x f_i(\bar x(a,r))\Big)\\
&\quad-\sum_{j=1}^p\nu_j(a,r)\nabla_a\bar x(a,r)^\top\nabla_x g_j(\bar x(a,r))-\sum_{k=1}^q\xi_k(a,r)\nabla_a\bar x(a,r)^\top\nabla_x h_k(\bar x(a,r)).\end{aligned}\tag{4.5}$$

According to Theorem 4.2 there exists some $\delta>0$ so that the derivative of the local minimal value function is given by $\nabla_{(a,r)}\overline{\tau}^\delta(a,r)=\nabla_{(a,r)}\overline{\mathcal{L}}(\bar t(a,r),\bar x(a,r),\ \mu(a,r),\ \nu(a,r),\ \xi(a,r),a,r)$. Applying standard rules of differentiation and together with (4.5) we conclude

$$\begin{aligned}\nabla_a\overline{\tau}^\delta(a,r) &= \nabla_a\bar t(a,r)-\sum_{i=1}^m\mu_i(a,r)\Big(e_i+\nabla_a\bar t(a,r)r_i-\ \nabla_a\bar x(a,r)^\top\nabla_x f_i(\bar x(a,r))\Big)\\
&\quad-\sum_{i=1}^m\nabla_a\mu_i(a,r)\underbrace{(a_i+\bar t(a,r)\,r_i-f_i(\bar x(a,r)))}_{=0}\\
&\quad-\sum_{j=1}^p\nu_j(a,r)\nabla_a\bar x(a,r)^\top\nabla_x g_j(\bar x(a,r))-\sum_{j=1}^p\nabla_a\nu_j(a,r)g_j(\bar x(a,r))\\
&\quad-\sum_{k=1}^q\xi_k(a,r)\nabla_a\bar x(a,r)^\top\nabla_x h_k(\bar x(a,r))-\sum_{k=1}^q\nabla_a\xi_k(a,r)\underbrace{h_k(\bar x(a,r))}_{=0}\\
&= -\mu(a,r)-\nabla_a\nu(a,r)^\top g(\bar x(a,r)).\end{aligned}$$

The same for $\nabla_r\overline{\tau}^\delta(a,r)$. ∎

We can use that inactive constraints remain inactive for small parameter changes and then in the case $C=\boldsymbol{R}^p_+$ we conclude (using the arguments in [17, Theorem 3.2.2, Proof of Theorem 3.4.1]) $\nabla_{(a,r)}\nu(a^0,r^0)^\top g(\bar x(a^0,r^0))=0_{2m}$. This results in the following.

COROLLARY 4.5. *Under the assumptions of Lemma* 4.4 *and with* $C=\boldsymbol{R}^p_+$ *it follows*

$$\nabla_{(a,r)}\overline{\tau}^\delta\left(a^0,r^0\right)=-\begin{pmatrix}\mu^0\\ t^0\,\mu^0\end{pmatrix}.$$

Hence, we get in this special case the derivative information via the Lagrange multipliers without additional effort just by solving the problems (SP(a,r)). Otherwise, the derivative of the local minimal value function, being equivalent to the derivative of the Lagrange function, has to be approximated. Under some special additional assumptions as $C=\boldsymbol{R}^p_+,\ K=\boldsymbol{R}^m_+,\ \hat S=S=\boldsymbol{R}^n$, and nondegeneracy the second-order information $\nabla^2_a\tau^\delta(a^0,r^0)=-\nabla_a\mu(a^0,r^0)$ and $\nabla^2_r\tau^\delta(a^0,r^0)=t^0\mu^0(\mu^0)^\top-t^0\nabla_r\mu(a^0,r^0)$ is available [13, Theorem 3.2.4], too.

5. Parameter control and algorithm. In the literature several quality criteria have been discussed for approximations of the efficient set (see [7, 12, 39, 47, 54], and others). Most of them have been developed for evaluating evolutionary algorithms, as, e.g., the quality criteria of measuring the distance of the approximation set to the efficient set. As our approximation points are determined by solving the problems $(\mathrm{SP}(a,r))$ they are at least weakly K-minimal. Here we suppose that a numerical solver is at our disposal which allows us to find global minimal solutions of the considered scalar optimization problems. However, generally numerical methods generate only approximations of a minimal solution if not even only local minimal solutions. Then the distance to the efficient set depends on the numerical solvers used and not on the adaptive parameter control in which we are interested here. Therefore, quality criteria as the distance to the efficient set are not considered in this context.

The most interesting criteria, in the case of scalarization approaches, are the three proposed by Sayin [47] called coverage error, uniformity, and cardinality. In our opinion and with respect to these targets, an approximation possesses a high quality in the sense of a concise but representative approximation if it consists of almost equidistant approximation points.

We want to use the sensitivity results from section 4 to reach the aim of an equidistant approximation, at least locally. We first apply these results for developing a method for determining the parameters a such that we can control the distance between the generated approximation points. This will be used in section 5.2 for locally refining coarse approximations of the efficient set with equidistant points. In section 5.3 we specialize our results to the bicriteria case ($m = 2$), because for two objective functions we do not have to determine a coarse approximation first which is then refined. Instead we can adaptively control the parameter a from the beginning to generate an equidistant approximation of the whole efficient set.

5.1. Parameter control. We start by assuming that we have already solved a so-called reference problem $(\mathrm{SP}(a^0,r))$ with minimal solution (t^0,x^0) with Lagrange multipliers (μ^0,ν^0,ξ^0) and with $a^0 + t^0\, r - f(x^0) = 0_m$. (Otherwise, for $a^0 + t^0\, r - f(x^0) = k \neq 0_m$, we can apply Theorem 3.6 and determine a scalar t' and a parameter a' with $a' + t'\, r - f(x^0) = 0_m$. Then we take the problem $(\mathrm{SP}(a',r))$ as reference problem.) Then (t^0,x^0) is a minimal solution of the modified problem $(\overline{\mathrm{SP}}(a^0,r))$. We further assume that the derivative $\nabla_a \overline{\tau}^\delta(a^0,r)$ of the local minimal value function is known as a consequence of Theorem 4.2.

We will concentrate on a variation of the parameter a. We use the derivative information for a first-order Taylor series approximation (assuming this is possible) of the minimal value t of problem $(\overline{\mathrm{SP}}(a,r))$ depending on the parameter a given by $\bar{t}(a,r) \approx t^0 + \nabla_a \overline{\tau}^\delta(a^0,r)^\top (a - a^0)$. Using the equality constraint $f(\bar{x}(a,r)) = a + \bar{t}(a,r)\, r$ of problem $(\overline{\mathrm{SP}}(a,r))$ we get, together with $f(x^0) = a^0 + t^0\, r$,

$$\begin{aligned} f(\bar{x}(a,r)) &\approx a^0 + \left(a - a^0\right) + \left(t^0 + \nabla_a \overline{\tau}^\delta \left(a^0,r\right)^\top \left(a - a^0\right)\right)\, r \\ &= f\left(x^0\right) + \left(a - a^0\right) + \left(\nabla_a \overline{\tau}^\delta \left(a^0,r\right)^\top \left(a - a^0\right)\right)\, r. \end{aligned}$$

We now use $(\bar{t}(a,r),\bar{x}(a,r))$ (the minimal solution of $(\overline{\mathrm{SP}}(a,r))$) as an approximation of the minimal solution $(t(a,r),x(a,r))$ of $(\mathrm{SP}(a,r))$. We have at least $t(a,r) \leq \bar{t}(a,r)$ and $(\bar{t}(a,r),\bar{x}(a,r))$ feasible for $(\mathrm{SP}(a,r))$. Hence, we get the following local approximation of the generated weakly efficient points of (2.1) depending on the

parameter a:

$$f(x(a,r)) \approx f\left(x^0\right) + \left(a - a^0\right) + \left(\nabla_a \overline{\tau}^\delta \left(a^0, r\right)^\top \left(a - a^0\right)\right) r. \tag{5.1}$$

Our goal is to compute equidistant approximation points and for a predefined distance of $\alpha > 0$ we want to find a new parameter a^1 such that

$$\left\| f\left(x\left(a^0, r\right)\right) - f\left(x\left(a^1, r\right)\right)\right\| = \alpha \tag{5.2}$$

with $f(x(a^0, r)) = f(x^0)$, i.e., such that the new approximation point has a distance of α to the former. In Theorem 3.2 we have seen that it is sufficient to consider parameters a in a hyperplane H. Assuming $a^0 \in H = \{y \in \boldsymbol{R}^m \mid b^\top y = \beta\}$, we choose a direction $v \in \boldsymbol{R}^m$ with $b^\top v = 0$ such that $a^0 + s\,v \in H$ for all $s \in \boldsymbol{R}$. Because we want $a^1 \in H$ we set $a^1 = a^0 + s^1\,v$, $s^1 \in \boldsymbol{R}$, and together with (5.2) and (5.1) this results in

$$\begin{aligned}
\alpha &= \left\| f\left(x\left(a^0, r\right)\right) - f\left(x\left(a^1, r\right)\right)\right\| \\
&\approx \left\| f\left(x^0\right) - \left(f\left(x^0\right) + s^1\,v + s^1 \left(\nabla_a \overline{\tau}^\delta \left(a^0, r\right)^\top v\right) r\right)\right\| \\
&= \left|s^1\right| \left\| v + \left(\nabla_a \overline{\tau}^\delta \left(a^0, r\right)^\top v\right) r\right\|.
\end{aligned}$$

Hence, we choose

$$s^1 := \frac{\alpha}{\left\| v + \left(\nabla_a \overline{\tau}^\delta \left(a^0, r\right)^\top v\right) r\right\|}. \tag{5.3}$$

For the new parameter $a^1 := a^0 + s^1\,v$ (or $a^1 := a^0 - s^1\,v$) we now solve the problem $(\mathrm{SP}(a^1, r))$ with minimal solution (t^1, x^1), and if the quality of our approximations have been good this results in $\|f(x^0) - f(x^1)\| \approx \alpha$.

5.2. General multiobjective case. For the general multiobjective case ($m \geq 2$) we have seen in Theorem 3.5 that it is sufficient to vary the parameter a in the subset H^0 for being able to detect all K-minimal points of the multiobjective optimization problem. We use this information for determining in a first step a coarse approximation of the efficient set. Then, in a second step, this approximation is refined locally with (almost) equidistant points.

The coarse approximation is done using equidistant parameters only. The d.m. can, for instance, define the desired number of points N^i in each direction v^i, $i = 1, \ldots, m-1$, spanning the hyperplane H (see (3.7)). With a distance of $L_i := (s_i^{\max,i} - s_i^{\min,i})/N^i$ this leads to $\prod_{i=1}^{m-1} N^i$ equidistant discretization points. For any of these parameters $a \in H^0$ and for $r \in K$ constant we solve the scalar optimization problem $(\mathrm{SP}(a, r))$ with minimal solution (t^a, x^a) (if one exists) and Lagrange multiplier μ^a to the constraint $a + t\,r - f(x) \in K$.

Based on this coarse approximation the d.m. gets a first overview over the efficient set and can now choose which areas or points are of special interest for doing a refinement now using the results of section 5.1. Let $f(x^0)$ be such a chosen approximation point (with $a^0 + t^0\,r - f(x^0) = 0_m$) and assume a refinement with $\bar{n} \in \boldsymbol{N}$ additional points in every direction should be done. Thus, we search for parameters a with $\|f(x(a)) - f(x^0)\| = \alpha$ for a distance $\alpha > 0$. As the hyperplane H is spanned by

the $m-1$ vectors v^k, $k=1,\dots,m-1$, we set for the new parameters $a=a^0+s\cdot v^k$, $k\in\{1,\dots,m-1\}$, for some $s\in\boldsymbol{R}$. This leads, according to (5.3), to

$$s^k:=\frac{\alpha}{\left\|v+\left(\nabla_a\overline{\tau}^\delta\,(a^0,r)^\top v^k\right)\,r\right\|}.$$

Hence, we get the $2(m-1)$ new parameters $a:=a^0\pm s^k\,v^k$, $k\in\{1,\dots,m-1\}$.

We extend this to a consideration of all $(2\bar{n}+1)^2-1$ parameters a with $a=a^0+\sum_{k=1}^{m-1}l_k\,s^k\,v^k$ for $l_k\in\{-\bar{n},\dots,\bar{n}\}\subset\boldsymbol{Z}$, $k=1,\dots,m-1$, $l:=(l_1,\dots,l_{m-1})\neq 0_{m-1}$, for getting $\bar{n}$ new parameters in every direction. Solving $(\mathrm{SP}(a,r))$ for all these parameters results in a refined approximation with locally equidistant points (around $f(x^0)$).

For the calculation of the values s^k we need the derivative of the local minimal value function which is given in Theorem 4.2. For that, the derivative of the Lagrange function has to be approximated. In the case of $S=\mathbb{R}^n$ this is reduced to an approximation of the derivative of the function ν w.r.t. a. According to Corollary 4.5, in the case of $C=\mathbb{R}^p_+$ the derivative is even immediately given by $\nabla_a\overline{\tau}^\delta(a^0,r)=-\mu^0$. Choosing additionally $K=\mathbb{R}^n_+$ and the hyperplane $H=\{y\in\boldsymbol{R}^m\mid y_m=0\}$, i.e., $b=e_m$, with e_m the mth unit vector in $\boldsymbol{R}^m$, $\beta=0$, and $r=e_m$, then solving the problem $(\mathrm{SP}(a,r))$ for parameters $a=(a_1,\dots,a_{m-1},0)\in H$ is equivalent to solve the problem

$$\begin{gathered}\min f_m(x)\\ \text{subject to the constraints}\\ f_i(x)\le a_i,\qquad i=1,\dots,m-1,\\ x\in\Omega.\end{gathered}$$

This problem is well known as ε-constraint scalarization [21]. Choosing for v^i the unit vectors the problems (3.8) then reduce to $\min_{x\in\Omega}f_i(x)$ and $\max_{x\in\Omega}f_i(x)$ for $i=1,\dots,m-1$, and thus the calculation of the set H^0 (Step 1) is facilitated. Also, the points a', t' to a minimal solution $\bar{x}$ as in Theorem 3.6 are just given by $t'=f_m(\bar{x})$ and $a'=(f_1(\bar{x}),\dots,f_{m-1}(\bar{x}),0)$. Hence, we assume for the given algorithm:

Assumption 4. *Let Assumption* 3 *hold with* $S=\boldsymbol{R}^n$, $K=\mathbb{R}^m_+$, *and* $C=\mathbb{R}^p_+$. *To any choice of parameters* (a,r) *for which we consider the optimization problem* $(\mathrm{SP}(a,r))$ *or* $(\overline{\mathrm{SP}}(a,r))$ *let there exist a minimal solution* $(\bar{t},\bar{x})$ *with Lagrange multipliers* $(\bar{\mu},\bar{\nu},\bar{\xi})\in\boldsymbol{R}^m\times\boldsymbol{R}^p_+\times\boldsymbol{R}^q$ *and let the assumptions of Theorem* 4.2 *in* $(\bar{t},\bar{x})$ *be satisfied.*

This simplifies the algorithm considerably and allows a short representation. For multiobjective optimization problems with arbitrary ordering cones, the ε-constraint reformulation instead of $(\mathrm{SP}(a,r))$ is generally not possible and the determination of the set H^0 is more laborious. Also the calculation of the derivatives $\nabla_a\overline{\tau}^\delta(a,r)$ is more costly if C is not the natural ordering and if S does not equal the whole space (see Lemma 4.4). As the formulation of the general algorithm goes straightforward we restrain here to this common special case.

Algorithm 5.1 (Algorithm for an adaptive parameter control).

Input: Set $r=e_m$, $b=e_m$, $\beta=0$. Choose desired number of discretization points N^i in direction $v^i=e_i$ for $i=1,\dots,m-1$.

Step 1: Solve problem $\min_{x\in\Omega}f_i(x)$ with minimal solution $x^{\min,i}$ and minimal value $f(x^{\min,i})=:a_i^{\min}$ for $i=1,\dots,m-1$, and problem $\max_{x\in\Omega}f_i(x)$ with maximal solution $x^{\max,i}$ and maximal value $f(x^{\max,i})=:a_i^{\max}$ for $i=1,\dots,m-1$.

Step 2: Set $L_i := (a_i^{\max} - a_i^{\min})/N^i$ for $i = 1, \ldots, m-1$ and solve (SP(a, r)) for all $a \in E$ with $E := \{a = (a_1, \ldots, a_{m-1}, 0) \in \boldsymbol{R}^m \mid a_i = a_i^{\min} + L_i/2 + l_i \cdot L_i \text{ for } l_i = 0, \ldots, N^i - 1,\ i = 1, \ldots, m-1\}$ with minimal solution x^a and Lagrange multiplier μ^a to the constraint $a + t\,r - f(x) \in \boldsymbol{R}^m_+$. Determine the set $A := \{x^a \mid x^a$ minimal solution of (SP(a,r)) for $a \in E\}$.

Step 3: Determine the set $D := \{f(x) \mid x \in A\}$ and set $l = 0$.

Input: Choose $y \in D$ with $y = f(x^a)$ and associated Lagrange multiplier μ^a and set $a = (f_1(x^a), \ldots, f_{m-1}(x^a), 0)$. If y is a sufficient good solution, then stop. Otherwise, if additional points in the neighborhood of y are desired, then define a distance $\alpha \in \boldsymbol{R},\ \alpha > 0$, and the number $n^l \in \boldsymbol{N}$ of additional points for each direction and go to step 4.

Step 4: For all $\bar{\iota} = (i_1, \ldots, i_{m-1}) \in \{(i_1, \ldots, i_{m-1}) \in \boldsymbol{Z}^{m-1} \setminus \{(0, \ldots, 0)\} \,|\, i_j = -n^l, \ldots, n^l, \text{ for } j = 1, \ldots, m-1\}$ set

$$a^{\bar{\iota}} := a + \sum_{j=1}^{m-1} i_j \cdot \frac{\alpha}{\sqrt{1 + (\mu_j^a)^2}} \cdot e_j$$

and solve (SP$(a^{\bar{\iota}}, r)$). If there exists a solution $x^{\bar{\iota}}$ with Lagrange multiplier $\mu^{\bar{\iota}}$ set $A := A \cup \{x^{\bar{\iota}}\}$. Set $l := l + 1$ and go to Step 3.

Output: The set D is an approximation of the set of weakly efficient points.

Note that some of the problems considered in Step 2 and Step 4 may be infeasible. Thus in general not all parameters result in approximation points of the efficient set; see also test problem 4 in section 6.1. In [5, 23] conditions are given under which there exist minimal solutions of the problems (SP(a, r)). In detail, we have that for $K = \boldsymbol{R}^m_+$ and a nonempty efficient set there always exists a minimal solution for $a \in \boldsymbol{R}^m,\ r \in \text{int}(\boldsymbol{R}^m_+)$. However, here we have chosen $r = e_m \notin \text{int}(\boldsymbol{R}^m_+)$ for simplicity.

For solving the scalar optimization problems in Steps 1, 2, and 4, an appropriate numerical method has to be used as, e.g., the SQP method. However, using just a local solver can lead to only local minimal solutions of the scalar problems and thus to only locally weakly EP-minimal points of (2.1). As a starting point for a numerical method for solving problem (SP$(a^{\bar{\iota}}, r)$) in Step 4 the point $(f(x^a), x^a)$ can be used.

In Steps 2 and 3 a coarse approximation of the efficient set is calculated and in Step 4 around the special chosen points the refinement is done. Based on this algorithm it is possible to generate local equidistant approximations. With the coarse approximation in Step 2 it is ensured that all parts of the efficient set are covered and that the d.m. gets a survey of the efficient set. Then the method changes to an interactive part where the d.m. has to choose the areas in which a refinement is done.

5.3. Biobjective case. Now we come to the bicriteria case, i.e., $m = 2$. Of course the general Algorithm 5.1 presented in the previous section can be applied for $m = 2$, too. However, in the biobjective case we can use some special properties which do not hold generally for $m \geq 3$. This allows us not only to refine a coarse approximation locally but to determine equidistant approximations of the whole efficient set.

For $m = 2$ we can restrict the parameter set to a line segment H^a (Theorem 3.3). On this line segment we can easily define a total ordering, for instance increasing order w.r.t. the first coordinate. Then, points in the set H^a that are neighbors to each other are neighbors w.r.t. this order, too. This is no longer possible for the set H^0

for $m \geq 3$. As we have to know which points are neighbors to a point already found for using sensitivity information and for controlling the distance between the points, we start with a coarse approximation in the general case (section 5.2). However, since we are considering the biobjective case, we can use the special structure of H^a for adaptively determining the parameter a from the beginning without a previous coarse approximation. The result is an (almost) equidistant approximation of the efficient curve.

In the following we choose the parameters a in increasing order w.r.t. the first coordinate i.e., $a_1^0 \leq a_1^1 \leq \cdots \leq a_1^l \leq a_1^{l+1} \leq \cdots$ assuming we have $b_2 \neq 0$ for the hyperplane H. We choose $v \in \boldsymbol{R}^2$ with $b^\top v = 0$ such that we have $a^0 + s\,v \in H$ for $s \in \boldsymbol{R}$ and for $a^0 \in H$. Further, let $v_1 > 0$. We assume again we have already solved a reference problem (SP(a^0, r)) with a minimal solution (t^0, x^0). For $a^0 + t^0\,r - f(x^0) = 0_2$ we choose the next parameter a^1 by $a^1 = a^0 + s^1\,v$ with $s^1 > 0$ as in (5.3). Then, we have $a_1^1 > a_1^0$. For the case $a^0 + t^0\,r - f(x^0) = k^0 \neq 0_2$ we calculate $\tilde{a}^0$ as described in the proof of Theorem 3.6 by $\tilde{a}^0 = f(x^0) - \tilde{t}^0\,r$ and $\tilde{t}^0 = (b^\top f(x^0) - \beta)/(b^\top r)$. Then $\tilde{a}^0 + \tilde{t}^0\,r - f(x^0) = 0_2$.

In the case of $\tilde{a}_1^0 \geq a_1^0$ we set $a^1 := \tilde{a}^0 + s^1\,v$ (i.e., $a_1^1 > a_1^0$). For the ordering cone $K = \boldsymbol{R}_+^2$ we can show by an easy calculation using the fact that $a^0 + t^0\,r - f(x^0) = k^0$ with $k^0 \in \partial\boldsymbol{R}_+^2 \setminus \{0_2\}$ (see [45], but it is also a direct conclusion of the proof of Theorem 3.1 (b)) that $\tilde{a}_1^0 \geq a_1^0$ if and only if $\big(k_1^0 = 0,\ k_2^0 > 0$ and $\frac{r_1 b_2}{b^\top r} > 0\big)$ or $(k_1^0 > 0,\ k_2^0 = 0$ and $\frac{r_2 b_2}{b^\top r} < 0)$.

For the case $\tilde{a}_1^0 < a_1^0$ special considerations have to be made as it is not desirable to continue with the parameter $\tilde{a}^0$ instead of a^0 as we are looking for parameters with increasing first coordinate. In that case we still use the parameter a^0 for determining a^1. We can no longer assume $f(x(a,r)) = a + t(a,r)\,r$ as we have $f(x^0) = a^0 + t^0\,r - k^0$ with $k^0 \neq 0_2$. However, we can presume that the constraint $a + t\,r - f(x) \in K$ remains inactive and thus in view of $a^0 + t^0\,r = f(x^0) + k^0$ we set $a + t\,r = f(x^0) + k^0 + s\,k^0$ for some $s > 0$. For $s := \alpha/\|k^0\|$ we have a distance of $\alpha > 0$ between the points $a + t\,r$ and $a^0 + t^0\,r$, see Figure 3.1(b). Thus, we set the new parameter as

$$a^1 := f\left(x^0\right) + (1+s)\,k^0 - t\,r \tag{5.4}$$

with $s = \alpha/\|k^0\|$ and with some $t \in \boldsymbol{R}$. As we still demand $a \in H$ we choose $t = \frac{b^\top\left(f(x^0) + (1+s)\,k^0\right) - \beta}{b^\top r}$. Using the definition of $\tilde{a}^0$ we get $a^1 = \tilde{a}^0 + (1+s)(k^0 - \frac{b^\top k^0}{b^\top r} r)$. Because we have $b^\top a^1 = \beta$, the vector a^1 is actually an element of the hyperplane H. Again by an easy calculation, we can show that $a_1^1 \geq a_1^0$ for a_1 as in (5.4) and $s \geq 0$; i.e., the next parameter is chosen with increasing first coordinate.

By repeating the described steps for finding the next parameters $a^2, a^3, \ldots$ we can adaptively determine an almost equidistant approximation $f(x^0), f(x^1), \ldots$ of the efficient set of (2.1). However, it can happen that $\|f(x^l) - f(x^{l-1})\| \gg \alpha$ or $\|f(x^l) - f(x^{l-1})\| \ll \alpha$. This can be due to a strong varying curvature of the efficient set (see, e.g., [10, test problem CTP2]), especially if the distance α is not chosen appropriately small. It can also be due to gaps in the efficient set, i.e., nonconnected parts, see test problem 3 in section 6.1. In practice the efficient set of the examined multiobjective optimization problems is very often smooth (see, e.g., the application problem given in section 6.2 or in [6, 30, 36]). Nevertheless, if the distance between consecutive points is too large a refinement strategy as in Step 4 in Algorithm 5.1 can be applied subsequently. Too small distances can be eliminated afterwards just by reducing the approximation set.

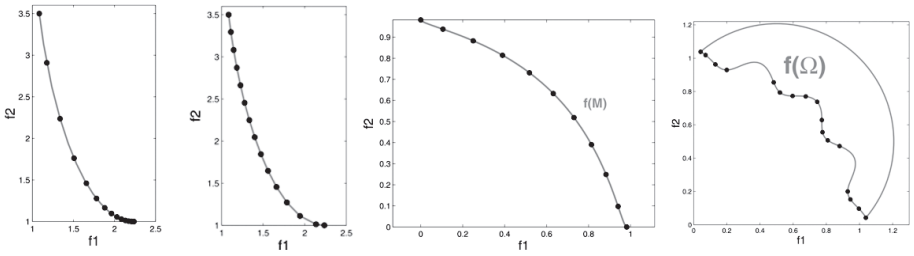


FIG. 6.1. (a) *Efficient set and approximation points of test problem* 1 *with weighted sum method and* (b) *with adaptive parameter control.* (c) *Efficient set and approximation points of test problem* 2. (d) *Image set and approximation points of test problem* 3.

6. Numerical results. In the following we apply the proposed methods on some test problems and on an application in intensity modulated radiotherapy.

6.1. Test problems. We start by considering some test problems which demonstrate the main properties of the proposed methods. It is shown that equidistant approximations are generated and that nonconvex problems as well as problems with a nonconnected efficient set or more than two objectives are covered.

Test problem 1: Comparison with the weighted sum method. The following test problem is chosen to show the advantage of the new method compared to the well-known habitual weighted sum method with which by a variation of the weights approximations of the efficient set can also be generated. For the bicriteria case the scalarization $\min_{x\in\Omega} w_1 f_1(x) + w_2 f_2(x)$ with weights $w_1, w_2 \in [0,1],\ w_1 + w_2 = 1$, is considered. Applying this scalarization to the test problem

$$\min_{\boldsymbol{R}^2_+} \begin{pmatrix} \sqrt{1+x_1^2} \\ x_1^2 - 4x_1 + x_2 + 5 \end{pmatrix}$$
$$\text{subject to the constraints}$$
$$x_1^2 - 4x_1 + x_2 + 5 \le 3.5,$$
$$x_1 \ge 0,\ x_2 \ge 0,$$

choosing uniformly distributed weights leads to the approximation shown in Figure 6.1(a). This approximation has an uneven distribution and thus a low uniformity and a high coverage error. A much better result again with 15 points is gained by applying the procedure of section 5.3 with a hyperplane $H = \{y \in \boldsymbol{R}^2 \mid (1,1)y = 2.5\}$, $r = (1,0)^\top$, and a predefined distance of $\alpha = 0.2$ between the approximation points, see Figure 6.1(b).

Test problem 2: Nonconvex set. The following example by van Veldhuizen [54, p. 545], see also [11, 33], has a nonconvex image. Letting $n \in \boldsymbol{N}$ be a parameter the problem is defined as follows:

$$\min_{\boldsymbol{R}^2_+} \begin{pmatrix} 1 - \exp\left(-\sum_{i=1}^{n}\left(x_i - \frac{1}{\sqrt{n}}\right)^2\right) \\ 1 - \exp\left(-\sum_{i=1}^{n}\left(x_i + \frac{1}{\sqrt{n}}\right)^2\right) \end{pmatrix}$$
$$\text{subject to the constraints}$$
$$x_i \in [-4,4] \quad \text{for all } i = 1,\dots,n.$$

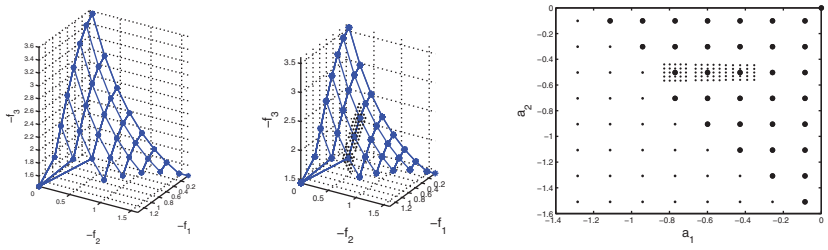


FIG. 6.2. *Test problem* 4: (a) *coarse approximation,* (b) *refined approximation, and* (c) *parameter set.*

We observe that by using the weighted sum method not all EP-minimal points can be found. For $n = 40$ we get with $r = (1,1)^\top$, $b = (1,0)^\top$, $\beta = 1.2$, $\alpha = 0.15$, and the procedure of section 5.3 the approximation shown in Figure 6.1(c). The connected line shows the efficient set of test problem 2 denoted as $f(M) := f(\mathcal{M}(f(\Omega), \boldsymbol{R}^2_+))$.

Test problem 3: Nonconnected efficient set. In the following problem by Tanaka [52] the image set $f(\Omega)$ is nonconvex, too, and additionally the efficient set is nonconnected (in a topological meaning):

$$\min_{\boldsymbol{R}^2_+} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$
$$\text{subject to the constraints}$$
$$x_1^2 + x_2^2 - 1 - 0.1\cos\left(16\arctan\left(\tfrac{x_1}{x_2}\right)\right) \geq 0,$$
$$(x_1 - 0.5)^2 + (x_2 - 0.5)^2 \leq 0.5,$$
$$x_1, x_2 \in [0, \pi].$$

We have $f(\Omega) = \Omega$. By choosing $r = (1,2)^\top$, $b = (1,1)^\top$, $\beta = 0.5$, $\alpha = 0.08$, and with the procedure of section 5.3 we get the approximation of Figure 6.1(d). Here the difficulty is that by solving the scalar optimization problems (SP(a,r)) global solutions have to be found and thus, by using only a local method as the SQP method only local EP-minimal points instead of global solutions of (2.1) are guaranteed.

Test problem 4: Three objectives. This test problem with a nonconvex image set is a modified version of a problem in [38]. For the ordering cone $K = \boldsymbol{R}^3_+$ we consider

$$\min_{\boldsymbol{R}^3_+} \begin{pmatrix} -x_1 \\ -x_2 \\ -x_3^2 \end{pmatrix}$$
$$\text{subject to the constraints}$$
$$-\cos(x_1) - \exp(-x_2) + x_3 \leq 0,$$
$$0 \leq x_1 \leq \pi,\ x_2 \geq 0,\ x_3 \geq 1.2.$$

We use Algorithm 5.1 with $N^1 = N^2 = 8$ for a coarse approximation. The result after Steps 1–3 is given in Figure 6.2(a). Note that the negative of the objective function values is drawn and that the approximation points are connected with lines.

In the input step of the algorithm we choose the three points $y \in D$ for which $y_1 \leq -0.4$ and $-0.6 \leq y_2 \leq -0.4$ holds. We do a refinement with $n^1 = 2$ and $\alpha = 0.06$. This

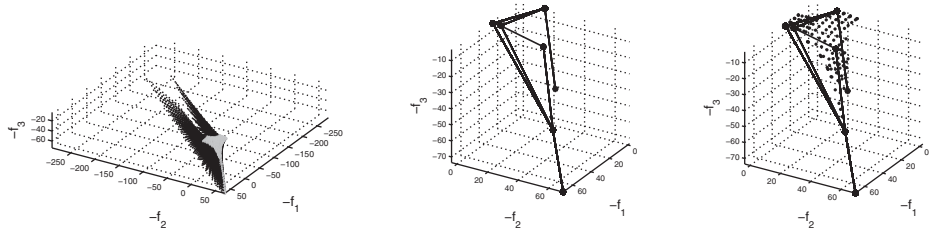


FIG. 6.3. *Test problem* 5: (a) *image set in dark color and efficient set in grey,* (b) *coarse approximation with the approximation points connected with lines, and* (c) *refined approximation.*

results in the refined approximation of the efficient set of Figure 6.2(b). In the course of the algorithm several scalar optimization problems (SP(a, r)) to parameters $a = (a_1, a_2, 0)$ are solved. These parameters are plotted as points (a_1, a_2) in Figure 6.2(c) as dots. To the parameters represented by the smallest dots, no minimal solution of the problem (SP(a, r)) exists. Around three parameters one can see the parameters of the refinement step (Step 4 of the algorithm), and it can well be seen that the distance between the refinement parameters differs depending on the sensitivity information delivered by the Lagrange multipliers (corresponding to the steepness of the efficient set).

Test problem 5: Comet problem. This problem [11, p. 9]

$$\min_{\boldsymbol{R}^3_+} \begin{pmatrix} (1+x_3)\left(x_1^3x_2^2 - 10x_1 - 4x_2\right) \\ (1+x_3)\left(x_1^3x_2^2 - 10x_1 + 4x_2\right) \\ 3(1+x_3)x_1^2 \end{pmatrix}$$
$$\text{subject to the constraints}$$
$$1 \le x_1 \le 3.5,\ -2 \le x_2 \le 2,\ 0 \le x_3 \le 1$$

with the set of K-minimal points $\mathcal{M}(f(\Omega), \boldsymbol{R}^3_+) = \{x \in \boldsymbol{R}^3 \mid 1 \le x_1 \le 3.5,\ -2 \le x_1^3x_2 \le 2,\ x_3 = 0\}$ has its name because of the image of the efficient set with a short broad and a long small area (see Figure 6.3(a)). A first coarse approximation for $N^1 = N^2 = 12$ delivers only a few approximation points (Figure 6.3(b)), but by doing a refinement according to Step 4 of Algorithm 5.1 with $n^1 = 3$, $\alpha = 4$ for all points with no other point with a distance of less than 5 in the neighborhood, an approximation with a high quality is finally achieved (Figure 6.3(c)).

6.2. Application in IMRT. We have also examined a problem in intensity modulated radiotherapy (IMRT) in medical engineering. Here, a patient with, e.g., a prostate tumor has to be irradiated to destroy the tumor. An optimal irradiation plan which is represented by an optimal intensity profile $x \in \mathbb{R}^{400}$ to 400 separate controllable beamlets B_i, $i = 1, \ldots, 400$ has to be found. We assume the beam geometry to be fixed. The problem is that the healthy surrounding organs should be damaged as little as possible while in each cell of the tumor a minimal curative dose has to be reached [1, 8, 40, 43].

Hence, this problem is a multiobjective optimization problem which has formerly been solved by just summing up the objective functions to one single scalar-valued objective using a weighted sum approach. Thereby, the difficulty is that the weights have no medical interpretation and that the physician has to find a good irradiation plan by a laborious trial and error process [40]. Using instead a multiobjective

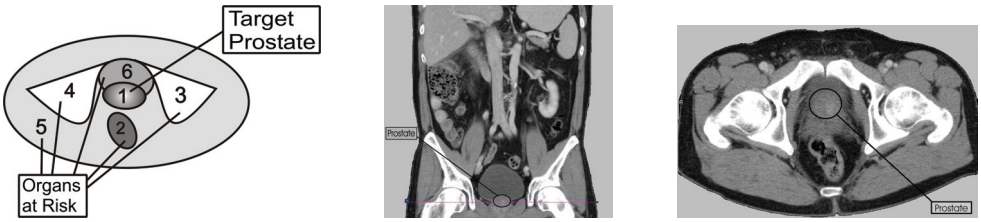


FIG. 6.4. (a) *Schematic axial body cut.* (b) *Coronar and axial CT-cut.*[1]

optimization approach with an approximation of the whole efficient set [22, 40, 43] simplifies this procedure and improves the results significantly and is, therefore, now actually applied. Thereby, a high quality approximation with equidistant points is demanded [40].

The number of objective functions depends on the number of different healthy tissues surrounding the tumor. In our example, the treatment planning for a prostate tumor, one is especially concerned about the bladder (V_6) and the rectum (V_2), while it has to be ensured that the doses in the remaining tissues as left (V_3) and right (V_4) hip bone and surrounding unspecified tissue (V_5) remain below an upper level (for the location of the organs see Figure 6.4(a)).

For evaluating and comparing the radiation stress in the several organs the concept of the equivalent uniform dose by Niemierko [44] based on p-norms is used. Therefore, the relevant part of the patient's body is mapped with the help of a computer tomography (CT), see Figure 6.4(b), and according to the thickness of the slices dissected in cubes, the so-called voxels. Then, using a clustering method [40, 50], whereas voxels with equal radiation exposure are collected, the high number of 435 501 voxels is reduced to 11 877 clusters c_j, $j = 1, \ldots, 11\,877$, which are allocated to the seven volumes $V_0, \ldots, V_6$ by a physician.

Volumes V_0 and V_1 describe the tumor (the so-called target-tissue) while V_1 is the boost-tissue, which is tumor tissue that has to be irradiated especially high. Thus, depending on the volume V_k, the number of voxels $N(V_k)$ in this organ, the number of voxels $N(c_j)$ in cluster c_j, and the dose limit U_k, the radiation stress is evaluated by

$$\mathrm{EUD}_k(x) := \frac{1}{U_k}\left(\frac{1}{N(V_k)}\sum_{\{j|c_j\in V_k\}} N(c_j)\cdot(P_j x)^{p_k}\right)^{\frac{1}{p_k}} - 1, \qquad k = 2, \ldots, 6.$$

The vector P_j denotes the jth row of the matrix $P = (P_{ji})_{j=1,\ldots,11\,877, i=1,\ldots,400}$ which describes the emission by the beamlet B_i ($i = 1, \ldots, 400$) in the cluster c_j ($j = 1, \ldots, 11\,877$) at one radiation unit.[2] For the intensity profile $x \in \boldsymbol{R}^{400}$, $P_j x$ denotes the irradiation dose in the cluster c_j caused by the beamlets B_i, $i = 1, \ldots, 400$. The parameter $p_k \in [1, \infty[$, which represents the physiology of the organ, is determined statistically and is given, like the other parameters, in Table 6.1.

The irradiation stress should remain below a critical value which results in the constraints $U_k(\mathrm{EUD}_k(x) + 1) \le Q_k$, $k = 2, \ldots, 6$, which can be restated as

$$\sum_{\{j|c_j\in V_k\}} N(c_j)(P_j x)^{p_k} \le Q_k^{p_k}\, N(V_k), \quad k = 2, \ldots, 6.$$

[1] By courtesy of Dr. R. Janka, Institute of Diagnostic Radiology, Univ. Erlangen-Nürnberg.
[2] The data are available on request by sending an email to the author.

TABLE 6.1
Critical values for the organs at risk.

	number of organ (k)	p_k	U_k	Q_k	$N(V_k)$
rectum	2	3.0	30	36	6 459
left hip-bone	3	2.0	35	42	3 749
right hip-bone	4	2.0	35	42	4 177
remaining tissue	5	1.1	25	35	400 291
bladder	6	3.0	35	42	4 901

TABLE 6.2
Critical values for the tumor tissues.

	number of organ (k)	L_k	δ_k	ε_k
target-tissue	0	67	0.11	0.11
boost-tissue	1	72	0.07	0.07

The radiation stress in the tumor tissues V_0 and V_1 is considered w.r.t. each single cluster, as it is important to destroy each single cancer cell. For homogeneity reasons this results in the constraints

$$\begin{aligned} L_0(1-\varepsilon_0) \le P_j x \le L_0(1+\delta_0), \qquad & \forall j \text{ with } c_j \in V_0 \\ \text{and } \; L_1(1-\varepsilon_1) \le P_j x \le L_1(1+\delta_1), \qquad & \forall j \text{ with } c_j \in V_1, \end{aligned}$$

with constants $L_0,\ L_1,\ \varepsilon_0, \varepsilon_1, \delta_0$ and δ_1 given in Table 6.2. Volume V_0 consists of 8 593 clusters while V_1 has 302 clusters. Including nonnegativity constraints for the beamlet intensity, this results in the feasible set

$$\begin{aligned} \Omega = \{x \in \mathbb{R}^{400}_+ \mid \; & U_k(\mathrm{EUD}_k(x)+1) \le Q_k, \qquad k = 2,\dots,6, \\ & L_0(1-\varepsilon_0) \le P_j x \le L_0(1+\delta_0), \qquad \forall j \text{ with } c_j \in V_0, \\ & L_1(1-\varepsilon_1) \le P_j x \le L_1(1+\delta_1), \qquad \forall j \text{ with } c_j \in V_1\} \end{aligned}$$

with 17 795 constraints.

The objectives are a minimization of the dose stress in the rectum (V_2) and in the bladder (V_6) as these two healthy organs always have the highest irradiation stress and a stress reduction for the rectum deteriorates the level for the bladder and vice versa. This leads to the biobjective optimization problem

$$\begin{gathered} \min_{\boldsymbol{R}^2_+} \begin{pmatrix} f_1(x) \\ f_2(x) \end{pmatrix} = \begin{pmatrix} \mathrm{EUD}_6(x) \\ \mathrm{EUD}_2(x) \end{pmatrix} \\ \text{subject to the constraint} \\ x \in \Omega. \end{gathered}$$

We apply the procedure of section 5.3 with $r = (1,1)^\top$, $\alpha = 0.04$, and $H = \{y \in \boldsymbol{R}^2 \mid y_1 = 0\}$, and we get that only parameters $a \in H^a$ with $H^a = \{y \in \boldsymbol{R}^2 \mid y_1 = 0,\ y_2 = \lambda \cdot 0.1841 + (1-\lambda)\cdot(-0.2197),\ \lambda \in [0,1]\}$ have to be considered. The approximation given in Figure 6.5(a) with 10 approximation points (connected with lines) is generated. These points as well as the distances δ^i between consecutive approximation points are listed in Table 6.3.

Based on these results the physician can choose a treatment plan by weighting the damage to the bladder and the rectum against each other. Besides he can choose an interesting plan and refine around it by using the strategy as in Step 4 of Algorithm 5.1. Further he can choose a point y determined by interpolation between the

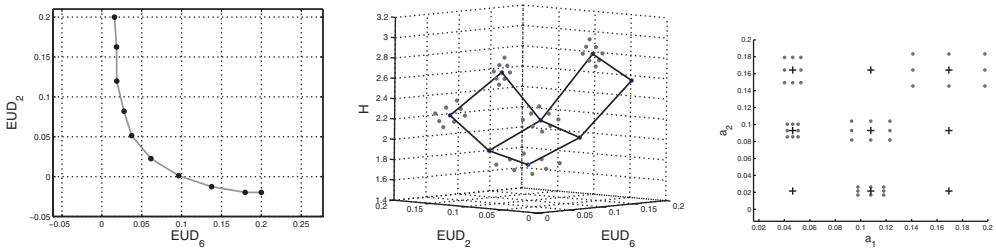


FIG. 6.5. *IMRT problem:* (a) *efficient set and approximation points of the bicriteria problem,* (b) *refined approximation of the tricriteria problem, and* (c) *parameter set of the tricriteria problem.*

TABLE 6.3
Approximation points and distances δ^i *between them for* $\alpha = 0.04$.

approximation point	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
$\mathrm{EUD}_2(\bar{x}^i)$	0.2000	0.1625	0.1197	0.0819	0.0515
$\mathrm{EUD}_6(\bar{x}^i)$	0.0159	0.0184	0.0187	0.0278	0.0374
δ_i	–	0.0375	0.0429	0.0389	0.0319
approximation point	$i=6$	$i=7$	$i=8$	$i=9$	$i=10$
$\mathrm{EUD}_2(\bar{x}^i)$	0.0228	0.0012	−0.0126	−0.0197	−0.0197
$\mathrm{EUD}_6(\bar{x}^i)$	0.0615	0.0964	0.1376	0.1796	0.2000
δ_i	0.0375	0.0411	0.0434	0.0426	0.0204

approximation points and solve problem (SP(a,r)) to the correspondent parameters, see [53], to get a new approximation point.

As it turned out that the treatment success depends also on the irradiation homogeneity, this objective can be added to the former two objective functions. Thereby the homogeneity of the irradiation is measured by

$$H(x) := \sqrt{\frac{\sum\limits_{\{j|c_j\in V_0\}} N(c_j)\,(P_j x - L_0)^2 + \sum\limits_{\{j|c_j\in V_1\}} N(c_j)\,(P_j x - L_1)^2}{N(V_0)+N(V_1)}}$$

with $N(V_0) = 13\,238$ and $N(V_1) = 2686$. This results in the multiobjective optimization problem

$$\min_{\boldsymbol{R}^3_+} \begin{pmatrix} f_1(x) \\ f_2(x) \\ f_3(x) \end{pmatrix} = \begin{pmatrix} \mathrm{EUD}_6(x) \\ \mathrm{EUD}_2(x) \\ H(x) \end{pmatrix}$$
$$\text{subject to the constraint}$$
$$x \in \Omega.$$

We have solved this problem using Algorithm 5.1 with $N^1 = N^2 = 3$. In Step 1 we get $a_1^{\min} = 0.0158$, $a_1^{\max} = 0.2000$, $a_2^{\min} = -0.0141$, and $a_2^{\max} = 0.2000$. This results in $L_1 = 0.0614$, $L_2 = 0.0714$ and thus in the parameter set $E := \{a \in \mathbb{R}^3 \mid a_1 \in \{0.0465,\ 0.1079,\ 0.1693\},\ a_2 \in \{0.0216,\ 0.0929,\ 0.1643\},\ a_3 = 0\}$. For solving the related scalar optimization problems we use the SQP procedure implemented in Matlab with 600 iterations and a restart after 150 iteration steps. We do not get a solution for the parameter $a = (0.0465, 0.0216, 0)$. We assume a physician chooses certain points and we do a refinement around these points with $n^1 = 1$ and $\alpha = 0.07$.

This results in the refined approximation shown in Figure 6.5(b). The determined parameters $a = (a_1, a_2, 0)$ according to Steps 2 and 4 are shown in Figure 6.5(c) as points (a_1, a_2).

If we choose, e.g., the approximation point $(0.0465, 0.1643, 0.2294)$ of the efficient set and calculate the distances between that point and the surrounding refinement points, we get the following 12 distances: $0.0697, 0.0801, 0.0795, 0.0814, 0.0880, 0.0679$, 0.0736, 0.0624, 0.0663, 0.0687, 0.0640, and 0.0712 with a rounded average value of 0.0727.

A more detailed and more technical description of this problem can be found in [13, 14, 40].

7. Outlook. Here we have developed an adaptive parameter control for the Pascoletti–Serafini scalarization. We have chosen this scalarization because it is not only suitable also for finding K-minimal points with $K \neq \boldsymbol{R}^m_+$, but it is also a very general method. Many other scalarization approaches such as the weighted Chebyshev norm, the ε-constraint method (see p. 1707 and [14]), the Polak method [34], or the normal boundary intersection (NBI) method [9] can be seen as a special case of this method (see [15]), and thus the presented results can be applied there too.

Acknowledgments. The author wishes to thank Prof. Dr. J. Jahn for his supervision, comments, and suggestions. Further, the author is grateful to PD Dr. K.-H. Küfer for providing the medical example and to the referees for their valuable comments and suggestions.

REFERENCES

[1] M. Alber and R. Reemtsen, *Intensity modulated radiotherapy treatment planning by use of a barrier-penalty multiplier method*, Optim. Methods Software, 22 (2007), pp. 391–411.

[2] C. Aliprantis, M. Florenzano, V. Martins-da-Rocha, and R. Tourky, *Equilibrium analysis in financial markets with countably many securities*, J. Math. Econom., 40 (2004), pp. 683–699.

[3] W. Alt, *Parametric optimization with applications to optimal control and sequential quadratic programming*, Bayreuth. Math. Schr., 35 (1991), pp. 1–37.

[4] K. Bergstresser, A. Charnes, and P. L. Yu, *Generalization of domination structures and nondominated solutions in multicriteria decision making*, J. Optim. Theory Appl., 18 (1976), pp. 3–13.

[5] H. Bernau, *Interactive methods for vector optimization*, Optimization in mathematical physics, Pap. 11th Conf. Methods Techniques Math. Phys., Oberwolfach/Ger. 1985, Methoden Verfahren Math. Phys., 34 (1987), pp. 21–36.

[6] E. Bijick, D. Diehl, and W. Renz, *Bikriterielle Optimierung des Hochfrequenzfeldes bei der Magnetresonanzbildung*, in Multicriteria Decision Making and Fuzzy Systems, Theory, Methods and Applications, K.-H. Küfer, H. Rommelfanger, C. Tammer, and K. Winkler, eds., Shaker, Aachen, 2006, pp. 85–98.

[7] Y. Collette and P. Siarry, *Three new metrics to measure the convergence of metaheuristics towards the Pareto frontier and the aesthetic of a set of solutions in biobjective optimization*, Comput. Oper. Res., 32 (2005), pp. 773–792.

[8] C. Cotrutz, C. Lahanas, and C. Kappas, *A multiobjective gradient-based dose optimization algorithm for external beam conformal radiotherapy*, Phys. Med. Biol., 46 (2001), pp. 2161–2175.

[9] I. Das and J. E. Dennis, *Normal-boundary intersection: A new method for generating the Pareto surface in nonlinear multicriteria optimization problems*, SIAM J. Optim., 8 (1998), pp. 631–657.

[10] K. Deb, A. Pratap, and T. Meyarivan, *Constrained test problems for multi-objective evolutionary optimization*, in First International Conference on Evolutionary Multi-Criterion Optimization, E. Zitzler, K. Deb, L. Thiele, C. A. Coello, and D. Corne, eds., Springer, Heidelberg, 2001, pp. 284–298.

[11] K. Deb, L. Thiele, M. Laumann, and E. Zitzler, *Scalable test problems for evolutionary multi-objective optimization*, in Evolutionary Multiobjective Optimization, A. Abraham, L. Jain, and R. Goldberg, eds., Springer, London, 2005, pp. 105–145.
[12] K. Deb and S. Jain, *Running performance metrics for evolutionary multi-objective optimization*, in Proceedings of the Fourth Asia-Pacific Conference on Simulated Evolution and Learning (SEAL'02), 2002, pp. 13–20.
[13] G. Eichfelder, *Parametergesteuerte Lösung nichtlinearer multikriterieller Optimierungsprobleme*, Ph.D. thesis, University of Erlangen-Nürnberg, Germany, 2006.
[14] G. Eichfelder, *ε-constraint method with adaptive parameter control and an application to intensity-modulated radiotherapy*, in Multicriteria Decision Making and Fuzzy Systems, Theory, Methods and Applications, K.-H. Küfer, H. Rommelfanger, C. Tammer, and K. Winkler, eds., Shaker, Aachen, 2006, pp. 25–42.
[15] G. Eichfelder, *Scalarizations for adaptively solving multi-objective optimization problems*, Comput. Optim. Appl., to appear, DOI 10.1007/s10589-007-9155-4, 2008.
[16] A. V. Fiacco, *Sensitivity analysis for nonlinear programming using penalty methods*, Math. Program., 10 (1976), pp. 287–311.
[17] A. V. Fiacco, *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*, Academic Press, London, 1983.
[18] J. Fliege, *Gap-free computation of Pareto-points by quadratic scalarizations*, Math. Methods Oper. Res., 59 (2004), pp. 69–89.
[19] J. Fliege, *An efficient interior-point method for convex multicriteria optimization problems*, Math. Oper. Res., 31 (2006), pp. 825–845.
[20] J. Fliege and L. N. Vicente, *A multicriteria optimization approach to bilevel optimization*, J. Optim. Theory Appl., 131 (2006), pp. 209–225.
[21] Y. Y. Haimes, L. S. Lasdon, and D. A. Wismer, *On a bicriterion formulation of the problems of integrated system identification and system optimization*, IEEE Trans. Syst. Man Cybern., 1 (1971), pp. 296–297.
[22] H. W. Hamacher and K.-H. Küfer, *Inverse radiation therapy planning - a multiple objective optimization approach*, Discrete Appl. Math., 118 (2002), pp. 145–161.
[23] S. Helbig, *An Interactive algorithm for nonlinear vector optimization*, Appl. Math. Optim., 22 (1990), pp. 147–151.
[24] S. Helbig, *Approximation of the efficient point set by perturbation of the ordering cone*, Z. Oper. Res., 35 (1991), pp. 197–220.
[25] C. Hillermeier and J. Jahn, *Multiobjective optimization: Survey of methods and industrial applications*, Surv. Math. Ind., 11 (2005), pp. 1–42.
[26] B. J. Hunt and M. M. Wiecek, *Cones to aid decision making in multicriteria programming*, in Multi-Objective Programming and Goal-Programming, T. Tanino, T. Tanaka, and M. Inuiguchi, eds., Springer, Berlin, 2003, pp. 153–158.
[27] B. J. Hunt, *Multiobjective Programming with Convex Cones*: *Methodology and Applications*, Ph.D. thesis, University of Clemson, 2004.
[28] C.-L. Hwang and A. S. M. Masud, *Multiple Objective Decision Making—Methods and Applications. A State-of-the-art Survey*, Springer, Berlin, 1979.
[29] J. Jahn, *Mathematical Vector Optimization in Partially Ordered Linear Spaces*, Lang, Frankfurt, 1986.
[30] J. Jahn, A. Kirsch, and C. Wagner, *Optimization of rod antennas of mobile phones*, Math. Methods Oper. Res., 59 (2004), pp. 37–51.
[31] J. Jahn, *Vector Optimization*: *Theory, Applications and Extensions*, Springer, Berlin, 2004.
[32] J. Jahn, *Introduction to the Theory of Nonlinear Optimization*, Springer, Berlin, 2007.
[33] J. Jahn, *Multiobjective search algorithm with subdivision technique*, Comput. Optim. Appl., 35 (2006), pp. 161–175.
[34] J. Jahn and A. Merkel, *Reference point approximation method for the solution of bicriterial nonlinear optimization problems*, J. Optim. Theory Appl., 74 (1992), pp. 87–103.
[35] K. Jittorntrum, *Solution point differentiability without strict complementarity in nonlinear programming*, Math. Program. Study, 21 (1984), pp. 127–138.
[36] A. Jüschke, J. Jahn, and A. Kirsch, *A bicriterial optimization problem of antenna design*, Comput. Optim. Appl., 7 (1997), pp. 261–276.
[37] I. Kaliszewski, *Quantitative Pareto Analysis by Cone Separation Technique*, Kluwer Academic Publishers, Boston, 1994.
[38] I. Y. Kim and O. de Weck, *Adaptive weighted sum method for bi-objective optimization*, Structural Multidiscip. Optim., 29 (2005), pp. 149–158.
[39] J. Knowles and D. Corne, *On metrics for comparing non-dominated sets*, in Proceedings of the World Congress on Computational Intelligence, 2002, pp. 711–716.

[40] K.-H. Küfer, A. Scherrer, M. Monz, F. Alonso, H. Trinkaus, T. Bortfeld, and C. Thieke, *Intensity-modulated radiotherapy - a large scale multi-criteria programming problem*, OR Spectrum, 25 (2003), pp. 223–249.

[41] H. Maurer and J. Zowe, *First and second-order necessary and sufficient optimality conditions for infinite-dimensional programming problems*, Math. Program., 16 (1979), pp. 98–110.

[42] A. Messac and C. A. Mattson, *Normal constraint method with guarantee of even representation of complete Pareto frontier*, AIAA Journal, 42 (2004), pp. 2101–2111.

[43] M. Monz, *Pareto Navigation*: *Interactive multiobjective optimisation and its application in radiotherapy planning*, Ph.D. thesis, University of Kaiserslautern, Germany, 2006.

[44] A. Niemierko, *Reposting and analysing dose distributions: A concept of equivalent uniform dose*, Medical Physics, 24 (1997), pp. 103–110.

[45] A. Pascoletti and P. Serafini, *Scalarizing vector optimization problems*, J. Optim. Theory Appl., 42 (1984), pp. 499–524.

[46] Y. Sawaragi, H. Nakayama, and T. Tanino, *Theory of Multiobjective Optimization*, Academic Press, London, 1985.

[47] S. Sayin, *Measuring the quality of discrete representations of efficient sets in multiple objective mathematical programming*, Math. Program., 87 A (2000), pp. 543–560.

[48] B. Schandl, K. Klamroth, and M. M. Wiecek, *Norm-based approximation in bicriteria programming*, Comput. Optim. Appl., 20 (2001), pp. 23–42.

[49] Y. Sawaragi, H. Nakayama, and T. Tanino, *Theory of Multiobjective Optimization*, Academic Press, London, 1985.

[50] A. Scherrer, K.-H. Küfer, T. Bortfeld, M. Monz, and F. Alonso, *IMRT planning on adaptive volume structures–a decisive reduction in computational complexity*, Phys. Med. Biol., 50 (2005), pp. 2033–2053.

[51] C. Tammer and K. Winkler, *A new scalarization approach and applications in multicriteria d. c. optimization*, J. Nonlinear Convex Anal., 4 (2003), pp. 365–380.

[52] M. Tanaka, *GA-based decision support system for multi-criteria optimization*, in Proceedings of the International Conference on Systems, Man and Cybernetics, 2 (1995), pp. 1556–1561.

[53] C. Thieke, *Multicriteria optimization in inverse radiotherapy planning*, Ph.D. thesis, University of Heidelberg, Germany, 2003.

[54] D. A. Van Veldhuizen, *Multiobjective evolutionary algorithms*: *Classifications, analyses, and new innovations*, Ph.D. thesis, Graduate School of Engineering, Air Force Institute of Technology, Dayton, Ohio, 1999.

[55] P. Weidner, *Dominanzmengen und Optimalitätsbegriffe in der Vektoroptimierung*, Wiss. Z. Tech. Hochsch. Ilmenau, 31 (1985), pp. 133–146.

[56] M. Wiecek, *Multi-scenario multi-objective optimization with applications in engineering design*, semi-plenary talk given at the 7th International Conference devoted to Multi-Objective Programming and Goal Programming, Tours, France (2006).

[57] H. C. Wu, *A solution concept for fuzzy multiobjective programming problems based on convex cones*, J. Optim. Theory Appl., 121 (2004), pp. 397–417.

[58] P. L. Yu, *Cone convexity, cone extreme points, and nondominated solutions in decision problems with multiple objectives*, J. Optim. Theory Appl., 14 (1974), pp. 319–377.

[59] L. Zadeh, *Optimality and non-scalared-valued performance criteria*, IEEE Trans. Automatic Control, 8 (1963), pp. 59–60.

SIAM J. OPTIM.
Vol. 19, No. 4, pp. 1719–1734

APPROXIMATIONS OF STOCHASTIC OPTIMIZATION PROBLEMS SUBJECT TO MEASURABILITY CONSTRAINTS*

PIERRE CARPENTIER†, JEAN-PHILIPPE CHANCELIER‡, AND MICHEL DE LARA‡

Abstract. Motivated by the numerical resolution of stochastic optimization problems subject to measurability constraints, we focus upon the issue of discretization. There exist indeed two components to be discretized for such problems, namely, the random variable modelling uncertainties (noise) and the σ-field modelling the knowledge (information) according to which decisions are taken. There is no reason to bind these two discretizations, which are a priori unrelated. In this setting, we present conditions under which the discretized problems converge to the original one. The focus is put on the convergence notions ensuring the quality of the approximation; we illustrate their importance by means of a counterexample based on the Monte Carlo approximation.

Key words. stochastic programming, measurability constraints, discretization

AMS subject classifications. 90C15, 49M25, 62L20

DOI. 10.1137/070692376

1. Introduction.

Let $\big(\Omega, \mathcal{A}, \mathbb{P}\big)$ be a probability space, and let $\big(\Xi, \mathcal{B}_\Xi\big)$ and $\big(U, \mathcal{B}_U\big)$ be $\mathbb{R}^n$ and $\mathbb{R}^p$, respectively, with their associated Borel σ-fields. Given a random variable $\boldsymbol{\xi}$ with values in Ξ and a subfield $\mathcal{F}$ of $\mathcal{A}$, which, respectively, represent the noise and the observation, we are concerned with the following stochastic optimization problem:

$$V\big(\boldsymbol{\xi}, \mathcal{F}\big) := \min_{\boldsymbol{u} \in L^2(\Omega, \mathcal{A}, \mathbb{P}; U)} \mathbb{E}\big[j(\boldsymbol{u}, \boldsymbol{\xi})\big] \tag{1.1a}$$

$$\text{subject to} \quad \boldsymbol{u} \text{ is } \mathcal{F}\text{-measurable}\,. \tag{1.1b}$$

Here $j : U \times \Xi \to \mathbb{R}$ (technical assumptions given later), and $\mathbb{E}$ is the mathematical expectation under probability $\mathbb{P}$. *The random variables, defined over* $\big(\Omega, \mathcal{A}, \mathbb{P}\big)$*, will be denoted using bold characters (e.g.,* $\boldsymbol{\xi} \in L^2(\Omega, \mathcal{A}, \mathbb{P}; \Xi)$*), whereas their realizations will be denoted using normal characters (e.g.,* $\xi \in \Xi$*).*

Remark. Problem (1.1) can be easily extended to the sequential control case with direct observation of the noises. Then $\boldsymbol{u} = (\boldsymbol{u_0}, \ldots, \boldsymbol{u_{T-1}})$, each $\boldsymbol{u_t}$ being measurable with respect to a subfield $\mathcal{F}_t$ of $\sigma\big(\boldsymbol{\xi}_0, \ldots, \boldsymbol{\xi}_t\big)$, which is the σ-field generated by the noises prior to t. Practical instances are multistage stochastic programming problems:

$$\min_{(\boldsymbol{u}, \boldsymbol{x})} \mathbb{E}\left[\sum_{t=0}^{T-1} L_{t+1}\big(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\xi}_{t+1}\big) + K\big(\boldsymbol{x}_T\big)\right] \tag{1.2a}$$

$$\text{subject to} \quad \begin{cases} \boldsymbol{x}_0 & = f_0\big(\boldsymbol{\xi}_0\big), \\ \boldsymbol{x}_{t+1} & = f_{t+1}\big(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\xi}_{t+1}\big), \end{cases} \tag{1.2b}$$

$$\boldsymbol{u}_t \text{ is } \mathcal{F}_t\text{-measurable.} \tag{1.2c}$$

*Received by the editors May 18, 2007; accepted for publication (in revised form) September 18, 2008; published electronically January 28, 2009.
http://www.siam.org/journals/siopt/19-4/69237.html

†École Nationale Supérieure de Techniques Avancées, 32 Boulevard Victor, 75739 Paris Cedex 15, France (Pierre.Carpentier@ensta.fr).

‡École Nationale des Ponts et Chaussées, 6 et 8 avenue Blaise Pascal, 77455 Marne la Vallée Cedex, France (chancelier@cermics.enpc.fr, delara@cermics.enpc.fr).

Formulation (1.1) has the advantage to clearly distinguish the respective roles of the noise $\boldsymbol{\xi}$ and of the information $\mathcal{F}$. However, in practical situations, the subfield $\mathcal{F}$ is often generated by an observation function h which depends on the noise

$$\mathcal{F} = \sigma\big(h(\boldsymbol{\xi})\big)$$

so that the distinction is not always obvious.

Two polar cases are worth mentioning.

- The *full information* case corresponds to $\mathcal{F} = \mathcal{A}$. In this case, under technical assumptions ensuring the interchange of minimization and expectation (see [15, Theorem 14.60]), problem (1.1) becomes

$$\mathbb{E}\big[\min_{u\in U} j(u, \boldsymbol{\xi})\big].$$

  Once the noise $\boldsymbol{\xi}$ is discretized by a random variable $\boldsymbol{\xi}_n$, the approximate solution is given by $\min_{u\in U} j(u, \boldsymbol{\xi}_n)$, which returns a $\boldsymbol{\xi}_n$-measurable solution and thus an $\mathcal{A}$-measurable one.
- The *open-loop* case arises when $\mathcal{F} = \{\emptyset, \Omega\}$. In this case, the problem is of deterministic nature provided that $\mathbb{E}\big[j(u, \boldsymbol{\xi})\big]$ and its gradient are readily available for each $u \in U$. Otherwise, a standard way to get around the difficulty of computing an expectation is to use samples of $\boldsymbol{\xi}$. A first approach, known as *sample average approximation* (SAA), consists of replacing the expectation to be minimized by a Monte Carlo approximation (see [8]). Another possibility is to use the stochastic gradient method (see [14]).

In the last two cases, one has to deal with only *one* stochastic approximation. However, in the general case, *two* different components of the problem have to be taken into account in order to discretize problem (1.1):

1. The σ-field $\mathcal{F}$ in (1.1b) must be approximated by a finite object $\mathcal{F}_n$ in order to deal with tractable constraints.
2. The expectation in (1.1a) must be approximated in order to be computable, the noise $\boldsymbol{\xi}$ being thus replaced with a finitely valued random variable $\boldsymbol{\xi}_n$.

As already mentioned, the subfield $\mathcal{F}$ is often, e.g., in the stochastic programming framework, given by $\mathcal{F} = \sigma\big(h(\boldsymbol{\xi})\big)$ so that most discretization schemes aim at deducing the discretization of $\mathcal{F}$ from the discretization of $\boldsymbol{\xi}$. In this way, the discretization $\boldsymbol{\xi}_n$ of $\boldsymbol{\xi}$ induces a discretization $\mathcal{F}_n = \sigma\big(h(\boldsymbol{\xi}_n)\big)$. This last discretization scheme may fail to satisfactorily approximate problem (1.1) so that additional conditions have to be added in order to overcome the difficulty (see Pennanen's approach in section 4 for further details). Nevertheless, in the general framework of problem (1.1), we observe that the discretizations of the noise and the information are *a priori unrelated* in the sense that there is no reason for one of these approximations to be deduced from the other. We will follow this way of proceeding in this paper, thus obtaining different perspectives both from the theoretical and the practical point of view.

The noise discretization—related to the convergence of measures and random variables—is somewhat "traditional" in probability theory, whereas the information discretization is not so well-known. Let us recall some results about the space $\mathcal{A}^\star$ of the subfields of $\mathcal{A}$ (see [11] and [7] for further details). The *strong convergence*[1] topology

[1] Although being termed as "strong," this topology actually corresponds to a pointwise convergence notion. We follow here the terminology given by Kudo [11]. Note that there exist "stronger" convergence notions for σ-fields, such as the *uniform convergence* topology given by Boylan [5].

on $\mathcal{A}^\star$ is the coarsest topology such that the conditional expectation is continuous with respect to the σ-field:

$$\mathcal{F}_n \xrightarrow[n\to+\infty]{\text{strong}} \mathcal{F} \Longleftrightarrow \forall f \in L^1(\Omega,\mathcal{A},\mathbb{P};\mathbb{R}), \lim_{n\to+\infty} \|\mathbb{E}[f \mid \mathcal{F}_n] - \mathbb{E}[f \mid \mathcal{F}]\|_{L^1} = 0.$$

Note that this definition depends on the probability $\mathbb{P}$. The main properties of $\mathcal{A}^\star$ equipped with the strong convergence topology are the following.

P_1 The strong convergence topology on $\mathcal{A}^\star$ is metrizable.

P_2 The set of σ-fields generated by finite partitions of Ω is dense in $\mathcal{A}^\star$.

P_3 If $\boldsymbol{y}_n \longrightarrow \boldsymbol{y}$ in probability and $\sigma(\boldsymbol{y}_n) \subset \sigma(\boldsymbol{y})$, then $\sigma(\boldsymbol{y}_n)$ strongly converges to $\sigma(\boldsymbol{y})$.

According to [13, Theorem 2.3.1], the notion of strong convergence of σ-fields, given using $L^1(\Omega,\mathcal{A},\mathbb{P};\mathbb{R})$, can be equivalently defined using $L^r(\Omega,\mathcal{F},\mathbb{P};U)$ for $r \in [1,+\infty)$.

PROPOSITION 1.1. *Let $r \in [1,+\infty)$. The two following statements are equivalent.*

- $\mathcal{F}_n \xrightarrow[n\to+\infty]{\text{strong}} \mathcal{F}$.
- $\forall f \in L^r(\Omega,\mathcal{A},\mathbb{P};U), \lim_{n\to+\infty} \|\mathbb{E}[f \mid \mathcal{F}_n] - \mathbb{E}[f \mid \mathcal{F}]\|_{L^r} = 0$.

In this paper, we present in section 2 a convergence result for approximations of problem (1.1). Contrary to the two polar cases of full or null information, the *functional* $J(\boldsymbol{u},\boldsymbol{\xi}) := \mathbb{E}\big[j(\boldsymbol{u},\boldsymbol{\xi})\big]$ now plays a central role. The continuity of J turns out to be crucial for convergence, and we enlighten the importance of the convergence notions related to the discretization of both the observation and the noise that are used in the approximation. In section 3, we illustrate by an example in what relaxing these notions may lead to failures. Ultimately, we review in section 4 the convergence results obtained in [17], [3], [12], and [10] about the same problem.

2. Convergence theorem. We go back to the initial problem (1.1). The framework of the study is the following.

- The underlying probability space is $(\Omega,\mathcal{A},\mathbb{P})$, and we consider $\mathcal{F}$ a subfield of $\mathcal{A}$.
- The control variable $\boldsymbol{u}$ belongs to the subset $\Delta(\mathcal{F})$ of the $\mathcal{F}$-measurable random variables and is moreover subject to pointwise constraints

  $$\Delta(\mathcal{F}) := \left\{\boldsymbol{u} \in L^r(\Omega,\mathcal{A},\mathbb{P};U),\ \boldsymbol{u} \text{ is } \mathcal{F}\text{-measurable and } \boldsymbol{u}(\omega) \in U^{\text{ad}}\ \mathbb{P}\text{ a.s.}\right\},$$

  U^{ad} being a closed convex subset of U. Here $1 \le r < +\infty$ and $L^r(\Omega,\mathcal{A},\mathbb{P};U)$ is equipped with the topology induced by the norm.
- The random variable $\boldsymbol{\xi}$ belongs to $L^q(\Omega,\mathcal{A},\mathbb{P};\Xi)$, where $1 \le q < +\infty$, equipped with the topology induced by the norm.
- The cost function J, defined on $L^r(\Omega,\mathcal{A},\mathbb{P};U) \times L^q(\Omega,\mathcal{A},\mathbb{P};\Xi)$, is given by

  $$J(\boldsymbol{u},\boldsymbol{\xi}) := \mathbb{E}\big[j(\boldsymbol{u},\boldsymbol{\xi})\big].$$

  Here j is a normal integrand on $U \times \Xi$, J being the associated integral functional (see [15, Chapter 14]).

Using these notations, we want to compute the optimal value $V(\boldsymbol{\xi},\mathcal{F})$ of problem (1.1):

$$V(\boldsymbol{\xi},\mathcal{F}) := \min_{\boldsymbol{u}\in\Delta(\mathcal{F})} J(\boldsymbol{u},\boldsymbol{\xi}). \tag{2.1}$$

Remark. There is no additional difficulty in incorporating in $\Delta(\mathcal{F})$ more general pointwise constraints such as $\boldsymbol{u}(\omega) \in U^{\text{ad}}(\omega)$ $\mathbb{P}$ a.s., U^{ad} being a measurable set-valued mapping whose values are closed and convex.

To approximate problem (2.1), we choose a sequence $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ of subfields of $\mathcal{A}$ and a sequence $\{\boldsymbol{\xi}_n\}_{n\in\mathbb{N}}$ of random variables in $L^q(\Omega,\mathcal{A},\mathbb{P};\Xi)$, and we consider the following approximated problem:

$$V\big(\boldsymbol{\xi}_n,\mathcal{F}_n\big) := \min_{\boldsymbol{u}\in\Delta(\mathcal{F}_n)} J\big(\boldsymbol{u},\boldsymbol{\xi}_n\big). \tag{2.2}$$

The next theorem emphasizes the role of adequate convergence notions, with rather strong assumptions ($\boldsymbol{H_3}$) on the criterion (weaker assumptions may be found in a companion paper [6]).

THEOREM 2.1. *Under the assumptions*

$\boldsymbol{H_1}$ *the sequence* $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ *strongly converges to* $\mathcal{F}$*, and* $\mathcal{F}_n\subset\mathcal{F}$*;*

$\boldsymbol{H_2}$ *the sequence* $\{\boldsymbol{\xi}_n\}_{n\in\mathbb{N}}$ *converges in norm to* $\boldsymbol{\xi}$ *in* $L^q(\Omega,\mathcal{A},\mathbb{P};\Xi)$*;*

$\boldsymbol{H_3}$ *the normal integrand* j *is such that*

$$\forall(u,u')\in U^2\ \forall(\xi,\xi')\in\Xi^2,\ |j(u,\xi)-j(u',\xi')|\le\alpha\,\|u-u'\|_U^r+\beta\,\|\xi-\xi'\|_\Xi^q\,,$$

the convergence of the approximated optimal costs holds true:

$$\lim_{n\to+\infty} V\big(\boldsymbol{\xi}_n,\mathcal{F}_n\big) = V\big(\boldsymbol{\xi},\mathcal{F}\big). \tag{2.3}$$

Proof.

Step 1 ($\limsup_{n\to+\infty} V\big(\boldsymbol{\xi}_n,\mathcal{F}_n\big)\le V\big(\boldsymbol{\xi},\mathcal{F}\big)$).

For any $\boldsymbol{u}\in\Delta(\mathcal{F})$, we define $\boldsymbol{u}_n=\mathbb{E}\left[\boldsymbol{u}\mid\mathcal{F}_n\right]$. Note that $\boldsymbol{u}=\mathbb{E}\left[\boldsymbol{u}\mid\mathcal{F}\right]$ $\mathbb{P}$ almost surely. Using assumption $\boldsymbol{H_1}$ and Proposition 1.1, we obtain the convergence of the sequence $\{\boldsymbol{u}_n\}_{n\in\mathbb{N}}$ to $\boldsymbol{u}$ in $L^r(\Omega,\mathcal{A},\mathbb{P};U)$. This implies that the set-valued mapping Δ is lower semicontinuous (see [2, Definition 1.4.2]). From assumption $\boldsymbol{H_3}$, we then deduce that the integral functional J is continuous and therefore upper semicontinuous. Using [2, Theorem 1.4.16], we conclude that the marginal function V is also upper semicontinuous:

$$\limsup_{n\to+\infty} V(\boldsymbol{\xi}_n,\mathcal{F}_n)\le V(\boldsymbol{\xi},\mathcal{F}). \tag{2.4}$$

Step 2 ($\liminf_{n\to+\infty} V\big(\boldsymbol{\xi}_n,\mathcal{F}_n\big)\ge V\big(\boldsymbol{\xi},\mathcal{F}\big)$).

Starting from

$$J(\boldsymbol{u},\boldsymbol{\xi}_n)=J(\boldsymbol{u},\boldsymbol{\xi})+\big(J(\boldsymbol{u},\boldsymbol{\xi}_n)-J(\boldsymbol{u},\boldsymbol{\xi})\big),$$

we obtain by minimization over $\Delta(\mathcal{F}_n)$

$$\begin{aligned}\min_{\boldsymbol{u}\in\Delta(\mathcal{F}_n)} J(\boldsymbol{u},\boldsymbol{\xi}_n) &\ge \min_{\boldsymbol{u}\in\Delta(\mathcal{F}_n)} J(\boldsymbol{u},\boldsymbol{\xi}) + \min_{\boldsymbol{u}\in\Delta(\mathcal{F}_n)} \Big(J(\boldsymbol{u},\boldsymbol{\xi}_n)-J(\boldsymbol{u},\boldsymbol{\xi})\Big)\\ &\ge \min_{\boldsymbol{u}\in\Delta(\mathcal{F})} J(\boldsymbol{u},\boldsymbol{\xi}) + \min_{\boldsymbol{u}\in\Delta(\mathcal{F})} \Big(J(\boldsymbol{u},\boldsymbol{\xi}_n)-J(\boldsymbol{u},\boldsymbol{\xi})\Big),\end{aligned}$$

the last inequality being true because $\mathcal{F}_n\subset\mathcal{F}$ implies $\Delta(\mathcal{F}_n)\subset\Delta(\mathcal{F})$. We thus obtain

$$V(\boldsymbol{\xi}_n,\mathcal{F}_n)\ge V(\boldsymbol{\xi},\mathcal{F})+\min_{\boldsymbol{u}\in\Delta(\mathcal{F})}\Big(J(\boldsymbol{u},\boldsymbol{\xi}_n)-J(\boldsymbol{u},\boldsymbol{\xi})\Big). \tag{2.5}$$

From assumptions $\boldsymbol{H_2}$ and $\boldsymbol{H_3}$, the last term in (2.5) converges to 0 as n goes to infinity, which proves that V is lower semicontinuous:

$$\liminf_{n\to+\infty} V(\boldsymbol{\xi}_n,\mathcal{F}_n)\ge V(\boldsymbol{\xi},\mathcal{F}). \tag{2.6}$$

Gathering (2.4) and (2.6) leads to the result. ∎

From the numerical point of view, problem (2.2) is a tractable approximation of problem (2.1) provided that the range of $\boldsymbol{\xi}_n$ is finite and that $\mathcal{F}_n$ is generated by a finite partition of Ω. Indeed, let

- $\big(\Omega_n^{(1)},\dots,\Omega_n^{(n)}\big)$ be a partition of Ω generating the σ-field $\mathcal{F}_n$, $u_n^{(i)}$ denoting the (constant) value of an $\mathcal{F}_n$-measurable control $\boldsymbol{u}$ on the subset $\Omega_n^{(i)}$,
- $\big(\mho_n^{(1)},\dots,\mho_n^{(n)}\big)$ be a partition of Ω generated by $\boldsymbol{\xi}_n$, $\xi_n^{(l)}$ denoting the (constant) value of the random variable $\boldsymbol{\xi}_n$ on the subset $\mho_n^{(l)}$.

Problem (2.2) is then equivalent to

$$\min_{(u_n^{(1)},\dots,u_n^{(n)})\in U^{\mathrm{ad}}\times\cdots\times U^{\mathrm{ad}}} \sum_{i=1}^{n}\sum_{l=1}^{n} \mathbb{P}\big(\Omega_n^{(i)}\cap\mho_n^{(l)}\big)\; j\big(u_n^{(i)},\xi_n^{(l)}\big), \tag{2.7}$$

whose numerical solution is obtained using classical optimization techniques.

Let us now comment on the assumptions made in Theorem 2.1.

- First of all, the *uniform continuity* assumption $\boldsymbol{H_3}$ made on j is a very strong technical one, which allows for an elementary proof. It can be alleviated using the tools of *epi-convergence*. We do not elaborate on that particular point (see [6] for a comprehensive report) and concentrate on assumptions concerning the convergence notions used in the problem approximation.
- Assumption $\boldsymbol{H_1}$ appears as a reasonable trade-off between two requirements. On the one hand, the strong convergence topology is the coarsest topology such that the conditional expectation is continuous. This is a minimal requirement for approximating $\mathbb{E}\big[j(\boldsymbol{u},\boldsymbol{\xi})\mid\mathcal{F}\big]$, so that it seems that the topology cannot be weakened. On the other hand, the subset of σ-fields generated by a finite partition of Ω is dense in the space $\mathcal{A}^\star$ for the strong convergence topology. This is a desirable feature as far as numerical approximation is concerned, which is no longer verified for more sophisticated topologies as, for instance, the uniform convergence topology defined by Boylan in [5] (see also [3] for a comparison between strong and uniform convergence topologies). Note moreover that $\mathcal{F}_n\subset\mathcal{F}$ is not obvious: When $\mathcal{F}=\sigma\big(h(\boldsymbol{\xi})\big)$, the last inclusion is not automatically satisfied for $\mathcal{F}_n=\sigma\big(h(\boldsymbol{\xi}_n)\big)$.
- Assumption $\boldsymbol{H_2}$ is again a bit strong because the convergence in probability of the sequence $\{\boldsymbol{\xi}_n\}_{n\in\mathbb{N}}$ is in fact sufficient to prove the theorem (see [6]). But the key point here is that the convergence in distribution is insufficient to ensure the result. Indeed, a continuity property on J with respect to $\boldsymbol{\xi}$ cannot be obtained by the *convergence in distribution*, as shown by the following example:
  - $\big(\Omega,\mathcal{A},\mathbb{P}\big)=\big([-1,1],\mathcal{B}_{[-1,1]},\mu\big)$, μ is the uniform distribution on $[-1,1]$ and $U=\Xi=\Omega$.
  - $j(u,\xi)=u\xi$.
  - $\boldsymbol{\xi}_n(\omega)=\begin{cases}(-1)^n & \text{if } \omega\geq 0,\\ (-1)^{n+1} & \text{otherwise,}\end{cases}\qquad \boldsymbol{u}(\omega)=\begin{cases}+1 & \text{if } \omega\geq 0,\\ -1 & \text{otherwise.}\end{cases}$
  - Being stationary in distribution, the sequence $\{\boldsymbol{\xi}_n\}_{n\in\mathbb{N}}$ is converging in distribution, whereas $J\big(\boldsymbol{u},\boldsymbol{\xi}_n\big)=(-1)^n$.

  The discretization process thus requires a stronger convergence notion than Monte Carlo. It is a major difference with the open-loop case, for which it is well-known that the SAA method requires only a notion of convergence in distribution (see, e.g., [8] and [16]).

In the next section, we focus on that particular point and illustrate, on the one hand, in what relaxing assumption $\boldsymbol{H_2}$ may alter the optimal solution and, on the other hand, in what deducing the discretization of $\mathcal{F}$ from the discretization of $\boldsymbol{\xi}$ may conflict with assumption $\boldsymbol{H_1}$.

3. Counterexample. We present here an example illustrating how convergence notions matter in order to accurately discretize problem (1.1). The example, originally designed by Systems & Optimization Working Group (SOWG),[2] has already been used several times, for instance, in [3] to illustrate why using a naive Monte Carlo fails to provide the optimal solution or in [19] (see also [18]) to show that even a sophisticated tool as the Fortet–Mourier metric between probability measures cannot by itself control the error when discretizing a stochastic optimal control problem. Although based on the same example,[3] the purpose is here different: We lower only the convergence requirement made on the approximations of $\boldsymbol{\xi}$ in Theorem 2.1, which leads to suboptimality.

3.1. Formulation and exact solution. We consider a dynamical system incorporating two time steps and only one decision variable. The initial state $\boldsymbol{x}$ is a random variable on $[-1,1]$ with uniform distribution. The final state of the system is defined as

$$\boldsymbol{z} := \boldsymbol{x} + \boldsymbol{u} + \boldsymbol{w}, \tag{3.1a}$$

$\boldsymbol{w}$ being another uniformly distributed random variable on $[-1,1]$ independent of $\boldsymbol{x}$ and the control $\boldsymbol{u}$ being a random variable measurable with respect to the initial state $\boldsymbol{x}$. Let $\epsilon > 0$, and consider the following problem:

$$\min_{\boldsymbol{u} \text{ is } \sigma(\boldsymbol{x})\text{-measurable}} \mathbb{E}\left[\epsilon \boldsymbol{u}^2 + \boldsymbol{z}^2\right]. \tag{3.1b}$$

The probability space associated with problem (3.1) is $\big([-1,1]^2, \mathcal{B}_{[-1,1]^2}, \mu\big)$, where $\mathcal{B}_{[-1,1]^2}$ is the Borel σ-field on $[-1,1]^2$ and μ is the product of two independent uniform probability distributions on $[-1,1]$. The random variables $\boldsymbol{x}$ and $\boldsymbol{w}$ are the two components of the identity application $\mathrm{Id}_{[-1,1]^2}$ on $[-1,1]^2$, the real-valued control variable $\boldsymbol{u}$ being defined on $[-1,1]^2$. Here $j(u,\xi) = \epsilon u^2 + (x+u+w)^2$ with $\xi = (x,w)$, and problem (3.1) is equivalent to

$$\min_{\boldsymbol{u} \text{ is } \sigma(\boldsymbol{x})\text{-measurable}} \int_{[-1,1]^2} \Big(\epsilon\big(\boldsymbol{u}(x,w)\big)^2 + \big(x + \boldsymbol{u}(x,w) + w\big)^2\Big)\mu\big(\mathrm{d}x\mathrm{d}w\big). \tag{3.2}$$

This problem is a Markovian stochastic optimal control problem which can be solved using dynamic programming. Introducing the Bellman functions

$$V_1(z) := z^2 \quad , \quad V_0(x) := \min_{u \in \mathbb{R}} \mathbb{E}\left[\epsilon u^2 + V_1(x + u + \boldsymbol{w})\right],$$

we obtain the optimal feedback law $u^\sharp$ and the associated optimal cost $J^\sharp := \mathbb{E}\left[V_0(\boldsymbol{x})\right]$:

$$\boldsymbol{u}^\sharp(x) = -\frac{x}{1+\epsilon}, \quad J^\sharp = \frac{1}{3}\left(1 + \frac{\epsilon}{1+\epsilon}\right). \tag{3.3}$$

[2] See Acknowledgments.

[3] Which thus provides the same conclusions.

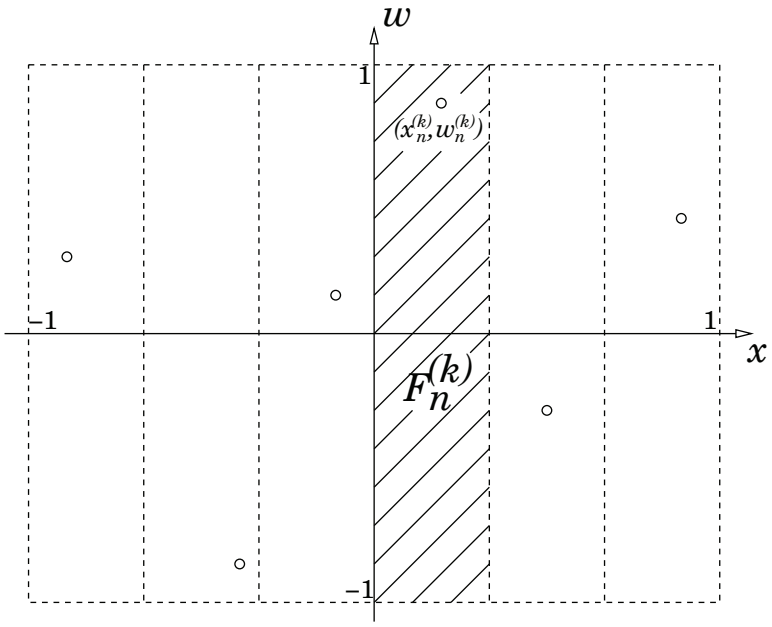


FIG. 3.1. *Partition of* $[-1,1]^2$ *and associated sample.*

3.2. Monte Carlo sampling fails.

3.2.1. Discretization. Let $\{\zeta_n\}_{n\in\mathbb{N}^*}$ be a deterministic sequence of elements in $[-1,1]^2$, with $\zeta_n = \big(\zeta_{n,1}, \zeta_{n,2}\big)$, and let μ_n be the empirical probability distribution associated with $(\zeta_1, \ldots, \zeta_n)$:

$$\mu_n := \frac{1}{n}\sum_{k=1}^{n} \delta_{\zeta_k},$$

where δ denotes the Dirac measure. We assume that the sequence $\{\mu_n\}_{n\in\mathbb{N}^*}$ of empirical probability distributions *weakly converges* to the probability measure μ (see [4]).

Remark. Such a sequence $\{\zeta_n\}_{n\in\mathbb{N}^*}$ is usually obtained as the realization of an infinite Monte carlo sample $\{\boldsymbol{\zeta}_n\}_{n\in\mathbb{N}^*}$ of i.i.d. random variables on $[-1,1]^2$ with distribution μ. The weak convergence assumption is then, almost surely, a consequence of the Glivenko–Cantelli theorem.

Let $n \in \mathbb{N}^\star$; for any $k \in \{1, \ldots, n\}$, we define

$$\big(x_n^{(k)}\ ,\ w_n^{(k)}\big) := \left(\frac{2k-1}{n} - 1 + \frac{\zeta_{k,1}}{n},\ \zeta_{k,2}\right) \tag{3.4}$$

and

$$I_n^{(k)} := \left(\frac{2k-2}{n} - 1, \frac{2k}{n} - 1\right] \quad , \quad F_n^{(k)} := I_n^{(k)} \times [-1,1]. \tag{3.5}$$

By construction, $\big(F_n^{(1)}, \ldots, F_n^{(n)}\big)$ is a partition of $[-1,1]^2$, made of vertical stripes as in Figure 3.1, and $\big(x_n^{(k)}, w_n^{(k)}\big) \in F_n^{(k)}\ \forall k \in \{1, \ldots, n\}$.

We are now ready to discretize problem (3.2).

Random variables. Let $q_n : [-1,1]^2 \to [-1,1]^2$ be the function defined by

$$q_n\big(x,w\big) := \sum_{k=1}^{n} \big(x_n^{(k)}, w_n^{(k)}\big)\mathbb{I}_{F_n^{(k)}}\big(x,w\big);$$

that is, $q_n\big(x,w\big) = \big(x_n^{(k)}, w_n^{(k)}\big)$ if $\big(x,w\big) \in F_n^{(k)}$. We define the sequence $\{\boldsymbol{x}_n, \boldsymbol{w}_n\}_{n\in\mathbb{N}^*}$ of random variables by

$$\big(\boldsymbol{x}_n, \boldsymbol{w}_n\big) := q_n\big(\boldsymbol{x}, \boldsymbol{w}\big). \tag{3.6}$$

According to this definition, the discretized random variable $\big(\boldsymbol{x}_n, \boldsymbol{w}_n\big)$ is constant over each subset $F_n^{(k)}$.

LEMMA 3.1. *The sequence* $\{\boldsymbol{x}_n, \boldsymbol{w}_n\}_{n\in\mathbb{N}^*}$ converges in distribution *to* $\big(\boldsymbol{x}, \boldsymbol{w}\big)$ *as* $n \to +\infty$.

Proof. Consider the empirical distribution function F_n of $\big(\boldsymbol{x}_n, \boldsymbol{w}_n\big)$:

$$F_n\big(x,w\big) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{I}_{[-1,x]\times[-1,w]}\big(x_n^{(k)}, w_n^{(k)}\big).$$

For a given $x \in [-1,1]$ and $n \in \mathbb{N}^*$, let k_0 be the index such that $x \in I_n^{(k_0)}$ (see (3.5)) and let ν_0 be equal to 0 if $x \leq x_n^{(k_0)}$ and equal to 1 otherwise. Then

$$\begin{aligned} F_n\big(x,w\big) &= \frac{1}{n}\sum_{k=1}^{k_0-1} \mathbb{I}_{[-1,w]}\big(w_n^{(k)}\big) + \frac{\nu_0}{n}\mathbb{I}_{[-1,w]}\big(w_n^{(k_0)}\big) \\ &= \frac{k_0-1}{n}\left(\frac{1}{k_0-1}\sum_{k=1}^{k_0-1} \mathbb{I}_{[-1,w]}\big(w_n^{(k)}\big)\right) + \frac{\nu_0}{n}\mathbb{I}_{[-1,w]}\big(w_n^{(k_0)}\big). \end{aligned}$$

The index k_0 goes to infinity as n goes to infinity (for any $x > -1$). We thus conclude that $F_n\big(x,w\big)$ converges to $F\big(x,w\big) = \frac{(1+x)(1+w)}{4}$, the distribution function of μ, the uniform probability on the square $[-1,1]^2$. ∎

Remark. Carrying on the previous remark, $\{\boldsymbol{x}_n, \boldsymbol{w}_n\}_{n\in\mathbb{N}^*}$ is usually a sequence of random variables based on the realization $\{\zeta_n\}_{n\in\mathbb{N}^*}$ of an i.i.d. sample. From Lemma 3.1, the sequence of associated probability distributions converges to μ: This is precisely the condition required in [8] in order to ensure convergence when discretizing an open-loop stochastic optimization problem.

Information. Since $\boldsymbol{x}$ is the first component of $\mathrm{Id}_{[-1,1]^2}$, the subfield $\sigma(\boldsymbol{x})$ of $\mathcal{B}_{[-1,1]^2}$ generated by the random variable $\boldsymbol{x}$ is

$$\mathcal{F} = \mathcal{B}_{[-1,1]} \otimes \{\emptyset, [-1,1]\}.$$

For a given $n \in \mathbb{N}^\star$, we approximate $\mathcal{F}$ by the σ-field $\mathcal{F}_n$ generated by the partition $\big(F_n^{(1)}, \ldots, F_n^{(n)}\big)$:

$$\mathcal{F}_n = \sigma\big(F_n^{(1)}, \ldots, F_n^{(n)}\big). \tag{3.7}$$

From the definition of the subsets $F_n^{(k)}$, the inclusion $\mathcal{F}_n \subset \mathcal{F}$ holds. Note that the approximated information constraint "$\boldsymbol{u}$ is $\mathcal{F}_n$-measurable" is equivalent to "$\boldsymbol{u}$ is

constant over each subset $F_n^{(k)}$," that is, constant on each vertical stripe of Figure 3.1. Such a control variable $\boldsymbol{u}$ is thus parameterized by the values $u_n^{(k)}$ taken on each subset $F_n^{(k)}$:

$$\boldsymbol{u}\big(x,w\big) = \sum_{k=1}^{n} u_n^{(k)} \mathbb{I}_{F_n^{(k)}}\big(x,w\big).$$

Notice that $\mathbb{I}_{F_n^{(k)}}\big(x,w\big)$ does not depend upon w, and therefore $\boldsymbol{u}\big(x,w\big)$ depends only upon x.

LEMMA 3.2. *The sequence* $\{\mathcal{F}_n\}_{n\in\mathbb{N}^*}$ strongly converges *to* $\mathcal{F}$ *as* $n \to +\infty$.

Proof. Since the values $x_n^{(k)}$, $k = 1, \ldots, n$, defined by (3.4) and taken by the random variable $\boldsymbol{x}_n$ are two-by-two distinct, then $\mathcal{F}_n = \sigma\big(\boldsymbol{x}_n\big)$. Following property $\boldsymbol{P_3}$, it is sufficient to show that $\boldsymbol{x}_n \longrightarrow \boldsymbol{x}$ in probability. This last convergence is obvious from the definition of $x_n^{(k)}$. □

3.2.2. Approximated solution. Approximating problem (3.2) consists in replacing $\mathcal{F}$ and $\big(\boldsymbol{x},\boldsymbol{w}\big)$ by their discretized versions $\mathcal{F}_n$ and $\big(\boldsymbol{x}_n,\boldsymbol{w}_n\big)$, respectively. The resulting function to be minimized is constant over each F_n^k so that the general form (2.7) for the approximated problem specializes in

$$\min_{\left(u_n^{(1)},\ldots,u_n^{(n)}\right)\in\mathbb{R}^n} \quad \sum_{k=1}^{n} \mathbb{P}\big(F_n^{(k)}\big)\left(\epsilon\big(u_n^{(k)}\big)^2 + \big(x_n^{(k)} + u_n^{(k)} + w_n^{(k)}\big)^2\right). \tag{3.8}$$

This optimization problem is of a deterministic nature and can be handled using standard optimization procedures. Since $\mathbb{P}\big(F_n^{(k)}\big) > 0$ (indeed equal to $\frac{1}{n}$), problem (3.8) splits into n independent subproblems shaped as

$$\min_{u_n^{(k)}\in\mathbb{R}} \quad \epsilon\big(u_n^{(k)}\big)^2 + \big(x_n^{(k)} + u_n^{(k)} + w_n^{(k)}\big)^2.$$

The optimal solution of this quadratic minimization problem is

$$\widehat{u}_n^{(k)} = -\frac{x_n^{(k)} + w_n^{(k)}}{1+\epsilon}, \tag{3.9}$$

and the associated optimal control variable $\widehat{\boldsymbol{u}}_{\boldsymbol{n}}$ is

$$\widehat{\boldsymbol{u}}_{\boldsymbol{n}}\big(x,w\big) = -\sum_{k=1}^{n} \frac{x_n^{(k)} + w_n^{(k)}}{1+\epsilon} \mathbb{I}_{F_n^{(k)}}\big(x,w\big)\ .$$

By construction the approximated feedback law $\widehat{\boldsymbol{u}}_{\boldsymbol{n}}$ is $\mathcal{F}_n$-measurable. From $\mathcal{F}_n \subset \mathcal{F}$, we deduce that $\widehat{\boldsymbol{u}}_{\boldsymbol{n}}$, as well as $\boldsymbol{u}^\sharp$, satisfies the measurability constraint of problem (3.2). We can compare the cost $\widehat{J}_n$ induced by $\widehat{\boldsymbol{u}}_{\boldsymbol{n}}$, namely,

$$\widehat{J}_n := \mathbb{E}\left[\epsilon\widehat{\boldsymbol{u}}_{\boldsymbol{n}}{}^2 + (\boldsymbol{x} + \widehat{\boldsymbol{u}}_{\boldsymbol{n}} + \boldsymbol{w})^2\right], \tag{3.10}$$

to the true optimal cost $J^\sharp$ in order to evaluate the quality of the approximation, that is, the optimality loss induced by $\widehat{\boldsymbol{u}}_{\boldsymbol{n}}$ with respect to $\boldsymbol{u}^\sharp$.

LEMMA 3.3. *The sequence* $\big\{\widehat{J}_n\big\}_{n\in\mathbb{N}}$ *is such that* $\lim_{n\to+\infty} \widehat{J}_n = \frac{2}{3}$.

Proof. We have

$$\widehat{J}_n = \int_{[-1,1]^2} \Big(\epsilon\big(\widehat{\boldsymbol{u}}_{\boldsymbol{n}}(x,w)\big)^2 + \big(x + \widehat{\boldsymbol{u}}_{\boldsymbol{n}}(x,w) + w\big)^2\Big)\mu\big(\mathrm{d}x\mathrm{d}w\big)$$
$$= \sum_{k=1}^{n} \int_{F_n^{(k)}} \left(\epsilon\left(\frac{x_n^{(k)} + w_n^{(k)}}{1+\epsilon}\right)^2 + \left(x + w - \frac{x_n^{(k)} + w_n^{(k)}}{1+\epsilon}\right)^2\right)\mu\big(\mathrm{d}x\mathrm{d}w\big).$$

Developing the last quadratic term in the previous expression leads to[4]

$$\widehat{J}_n = \frac{2}{3} + \frac{1}{n}\sum_{k=1}^{n} \frac{\big(x_n^{(k)} + w_n^{(k)}\big)^2}{1+\epsilon} - 2\sum_{k=1}^{n}\left(\frac{x_n^{(k)} + w_n^{(k)}}{1+\epsilon}\right)\int_{F_n^{(k)}} \big(x+w\big)\mu\big(\mathrm{d}x\mathrm{d}w\big)$$
$$= \frac{2}{3} + \frac{1}{n}\sum_{k=1}^{n} \frac{\big(x_n^{(k)} + w_n^{(k)}\big)^2}{1+\epsilon} - \frac{2}{n}\sum_{k=1}^{n}\left(\frac{2k-1}{n} - 1\right)\left(\frac{x_n^{(k)} + w_n^{(k)}}{1+\epsilon}\right).$$

Using the convergence in distribution of $\big(\boldsymbol{x}_n, \boldsymbol{w}_n\big)_{n\in\mathbb{N}^*}$ to $\big(\boldsymbol{x}, \boldsymbol{w}\big)$ as $n \to +\infty$ (see Lemma 3.1) and the inequalities $\left|x_n^{(k)} - \left(\frac{2k-1}{n} - 1\right)\right| \leq \frac{1}{n}$ for every k, we obtain that the sum of the last two terms in the previous equality goes to zero as n goes to infinity and hence the result. ∎

From the expression (3.3) of the true optimal cost $J^\sharp$, we deduce that the following inequality holds true for any $\epsilon > 0$:

$$\lim_{n\to+\infty} \widehat{J}_n > J^\sharp.$$

We thus conclude that the proposed discretization scheme fails to asymptotically give the optimal solution of the problem.

Remark. It is easy to verify that the optimal cost $\widetilde{J}_n$ of problem (3.8) is such that

$$\lim_{n\to+\infty} \widetilde{J}_n = \frac{2}{3}\left(\frac{\epsilon}{1+\epsilon}\right).$$

Once again, this limit is different from $J^\sharp$ and goes to zero as ϵ goes to zero. This is another illustration that problem (3.8) is not a valid approximation of problem (3.2). Although the *criterion* in (3.8) looks like a good approximation of the one in (3.2), the two corresponding *optimization problems* are definitely different. As a matter of fact, the "min" operator in (3.8) leads to solutions $\widehat{u}_n^{(k)} = -\frac{x_n^{(k)} + w_n^{(k)}}{1+\epsilon}$ depending on both $x_n^{(k)}$ and $w_n^{(k)}$. The computation of $\widetilde{J}_n$ therefore corresponds to the numerical integration of (3.2) using the feedback law $\widetilde{\boldsymbol{u}}(x,w) = -\frac{x+w}{1+\epsilon}$ which is not $\sigma(\boldsymbol{x})$-measurable and hence the gap with $J^\sharp$.

3.2.3. What has gone wrong. The discretization scheme we have devised in the above example is such that each subproblem derived from (3.8) is optimized using a *unique* sample of the random variable. The kth optimal control value $\widehat{u}_n^{(k)}$ depends on the corresponding first step noise value $x_n^{(k)}$ (which is in adequation with the constraint "$\boldsymbol{u}$ is $\sigma(\boldsymbol{x})$-measurable") and also depends on the second step noise

[4]Remember that $\mathbb{P}(F_n^{(k)}) = \frac{1}{n}$.

value $w_n^{(k)}$ (so that the control law is in some sense anticipative). In the following equivalent form of problem (3.1)

$$\mathbb{E}\Big[\min_{u\in\mathbb{R}} \mathbb{E}\big[\epsilon u^2 + (\boldsymbol{x}+u+\boldsymbol{w})^2 \mid \boldsymbol{x}\big]\Big],$$

our discretization scheme is close to approximating the inner conditional expectation using a single sample of $\boldsymbol{w}$, namely, a really poor way to compute such an expectation.

Let us revisit the assumptions of Theorem 2.1 in view of our example.

- Assumption $\boldsymbol{H_1}$ about the σ-fields is fulfilled.
- Assumption $\boldsymbol{H_3}$ is not exactly satisfied by the cost function j, but it is easy to check that the continuity properties of J required inside the proof are satisfied provided that the control remains bounded.
- Assumption $\boldsymbol{H_2}$ is not fulfilled because the convergence notion available in the example is significantly weaker than the one required by the theorem. More precisely, in our example, the convergence of the sequence $\{\boldsymbol{x}_n, \boldsymbol{w}_n\}_{n\in\mathbb{N}^*}$ towards $\big(\boldsymbol{x}, \boldsymbol{w}\big)$ *does not hold in probability.* Indeed, let $\tau > 0$ be given. Consider the norm $\|(x,w)\| = \sup\{|x|, |w|\}$ on $[-1,1]^2$, and let A_n be the subset of $[-1,1]^2$ defined by

$$A_n := \big\{(x,w)\in[-1,1]^2,\ \ \big\|\big(\boldsymbol{x}_n,\boldsymbol{w}_n\big)(x,w) - \big(\boldsymbol{x},\boldsymbol{w}\big)(x,w)\big\| \le \tau\big\}.$$

Since $(\boldsymbol{x},\boldsymbol{w}) = \mathrm{Id}_{[-1,1]^2}$ and since $(\boldsymbol{x}_n,\boldsymbol{w}_n)$ is constant over each $F_n^{(k)}$, the subset A_n can be expressed as the disjoint union of n subsets $A_n^{(k)}$, with

$$A_n^{(k)} := A_n \cap F_n^{(k)} = \Big\{(x,w)\in F_n^{(k)},\ \ \sup\Big\{\Big|x_n^{(k)} - x\Big|, \Big|w_n^{(k)} - w\Big|\Big\} \le \tau\Big\}.$$

From the definition of $F_n^{(k)}$ and A_n, the subset $A_n^{(k)}$ is included in a $\frac{2}{n}\times 2\tau$ rectangle. We thus obtain $\mu\big(A_n^{(k)}\big) \le \frac{\tau}{n}$ and then $\mu\big(A_n\big) \le \tau$ by summation. This demonstrates that

$$\mu\left(\big\|\big(\boldsymbol{x}_n,\boldsymbol{w}_n\big) - \big(\boldsymbol{x},\boldsymbol{w}\big)\big\| > \tau\right) \ge 1-\tau.$$

What has gone wrong in our example is now clear: Although the discretizations of the noise and the information are a priori unrelated, we have chosen to bind them in a very specific way; however, with this particular binding, one of the assumptions of the theorem, namely, $\boldsymbol{H_2}$, is not fulfilled.

Remark. Note, however, that the convergence notions used in the example may lead to a positive convergence result if the approximations of $\mathcal{F}$ and $\boldsymbol{\xi}$ are implemented in a nested manner (see Barty's approach in section 4).

3.2.4. A better discretization scheme. We ultimately illustrate in what a direct application of Theorem 2.1, and thus the use of a stronger convergence notion for the noise, leads to a positive result for our example. The information discretization previously chosen remains unchanged since it satisfies assumption $\boldsymbol{H_1}$ and leads to the vertical stripes $F_n^{(k)}$. In order to discretize the noise $\boldsymbol{\xi}$, we appeal to the theory of quantization (see, e.g., [9]) and introduce the Voronoi cells $C_n^{(k)}$ around the centroids $(x_n^{(k)}, w_n^{(k)})$ (see Figure 3.2). The discretized random variable $\big(\boldsymbol{x}_n,\boldsymbol{w}_n\big)$ is accordingly defined by

$$\big(\boldsymbol{x}_n,\boldsymbol{w}_n\big)(x,w) := \sum_{k=1}^{n} \big(x_n^{(k)}, w_n^{(k)}\big)\mathbb{I}_{C_n^{(k)}}(x,w).$$

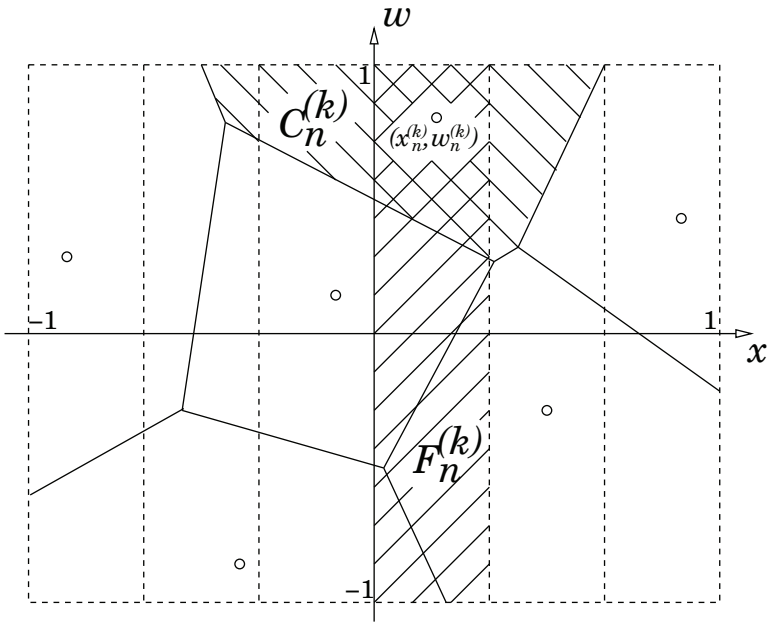


FIG. 3.2. *Noise discretization induced by Voronoi tessellation.*

The optimal Voronoi tessellation is based on the L^2 norm. We suppose that the diameter of the cells goes to zero as n goes to infinity so that assumption $\boldsymbol{H_2}$ is fulfilled. The approximated problem (2.7) to be solved again splits into n open-loop subproblems. Denoting by $\pi_{k,l} = \mathbb{P}(F_n^{(k)} \cap C_n^{(l)})$ the probability weight of the subset $F_n^{(k)} \cap C_n^{(l)}$, the kth subproblem writes

$$\min_{u_n^{(k)} \in \mathbb{R}} \sum_{l=1}^{n} \pi_{k,l} \left(\epsilon \big(u_n^{(k)}\big)^2 + \big(x_n^{(l)} + u_n^{(k)} + w_n^{(l)}\big)^2 \right). \tag{3.11}$$

It is clear that for a fixed k the number of nonempty subsets $F_n^{(k)} \cap C_n^{(l)}$ goes to infinity with n so that each optimal value $\widehat{u}_n^{(k)}$ is computed using a large (in fact asymptotically infinite) number of samples $(x_n^{(l)}, w_n^{(l)})$. This drastic difference with the approximation scheme given in section 3.2.1, where each optimal value $\widehat{u}_n^{(k)}$ is computed using one sample, explains the success of the last approximation.

Remark. A distinctive feature of the last discretization scheme is that a sample $(x_n^{(l)}, w_n^{(l)})$ may enter the computation of several control values $\widehat{u}_n^{(k)}$.

3.3. $\boldsymbol{\mathcal{F}_n = \sigma\big(h(\xi_n)\big)}$ fails. In the example given in section 3.1, the information subfield $\mathcal{F}$ is generated by a function h depending on the noise $\boldsymbol{\xi} = (\boldsymbol{x}, \boldsymbol{w})$, namely,

$$\mathcal{F} = \sigma\big(h(\boldsymbol{\xi})\big),$$

with $h(x, w) = x$. As already explained in section 1, it is possible in this case to deduce a discretization $\widetilde{\mathcal{F}}_n$ of $\mathcal{F}$ from the discretization $\boldsymbol{\xi}_n$ of $\boldsymbol{\xi}$ by

$$\widetilde{\mathcal{F}}_n = \sigma\big(h(\boldsymbol{\xi}_n)\big).$$

Going back to the discretization scheme of section 3.2.1, it is easy to figure out that the information discretization based on the vertical stripes $F_n^{(k)}$, that is,

$$\mathcal{F}_n = \sigma\big(F_n^{(1)}, \ldots, F_n^{(n)}\big),$$

is such that $\mathcal{F}_n = \widetilde{\mathcal{F}}_n$. In this case, the noise discretization induces a discretization of the information such that assumption $\boldsymbol{H_1}$ is satisfied: The pitfall is definitely the convergence notion used for approximating the noise.

Considering now the discretization scheme given in section 3.2.4, the approximated subfield $\mathcal{F}_n$ is the same as in section 3.2.1. The subfield $\widetilde{\mathcal{F}}_n$ deduced from the noise discretization $(\boldsymbol{x}_n, \boldsymbol{w}_n)$ is such that

$$\widetilde{\mathcal{F}}_n \subset \sigma\big(C_n^{(1)}, \ldots, C_n^{(n)}\big).$$

Assuming that the first coordinates of the centroids $(x_n^{(k)}, w_n^{(k)})$ are distinct one from each other,[5] the previous inclusion is in fact an equality. Then $\widetilde{\mathcal{F}}_n \not\subseteq \mathcal{F}$. Moreover we have $\widetilde{\mathcal{F}}_n \xrightarrow{\text{strong}} \mathcal{A}$ rather than $\mathcal{F}$. The subfield $\widetilde{\mathcal{F}}_n$ deduced from the noise discretization is thus not a good candidate for approximating problem (3.2), unless additional conditions are specified as in Pennanen's approach (see section 4 for further details).

4. Discussion about related works. We have proposed an approximation scheme in which the discretization of the noise and the discretization of the information are done separately. It is interesting to compare this approach with others also taking into account the whole discretization process (noise and information) for stochastic optimal control problems.

Barty's approach. In his Ph.D. thesis (in French), K. Barty proves the convergence of a discretization scheme for problem (1.1). The result he gives (see [3, Theorem IV.28]) makes use of the same notions of convergence as those used in section 3 for the counterexample, both for the σ-field and the random variable. Its approach involves two consecutive steps.

(i) The σ-field $\mathcal{F}$ is approximated by a σ-field $\mathcal{F}_k \subset \mathcal{F}$ which is generated by a finite partition $\mathcal{P}_k = \{\Omega_1, \ldots, \Omega_k\}$ of Ω, and problem (1.1) is replaced by

$$V\big(\boldsymbol{\xi}, \mathcal{F}_k\big) = \min_{\boldsymbol{u} \text{ is } \mathcal{F}_k\text{-measurable}} \mathbb{E}\big[j(\boldsymbol{u}, \boldsymbol{\xi})\big]. \tag{4.1}$$

The optimal value $V\big(\boldsymbol{\xi}, \mathcal{F}_k\big)$ of (4.1) converges with k towards the optimal value $V\big(\boldsymbol{\xi}, \mathcal{F}\big)$ of (1.1) as $\mathcal{F}_k$ strongly converges to $\mathcal{F}$ (see [3, Theorem IV.21]). Note that an $\mathcal{F}_k$-measurable random variable $\boldsymbol{u}$ is constant over each subset Ω_l constituting $\mathcal{P}_k$: Such a random variable $\boldsymbol{u}$ is characterized by a vector $\big(u_1, \ldots, u_k\big) \in U^k$, and the minimization in (see 4.1) is thus performed over a finite dimensional space.

(ii) For a given index k, the random variable $\boldsymbol{\xi}$ is approximated by a finitely valued random variable $\boldsymbol{\xi}_n$ and problem (4.1) is replaced by

$$V\big(\boldsymbol{\xi}_n, \mathcal{F}_k\big) = \min_{\boldsymbol{u} \text{ is } \mathcal{F}_k\text{-measurable}} \mathbb{E}\big[j(\boldsymbol{u}, \boldsymbol{\xi}_n)\big]. \tag{4.2}$$

[5] A property which does not conflict with the L^2 convergence.

The optimal value $V(\boldsymbol{\xi}_n, \mathcal{F}_k)$ of (4.2) converges with n toward the optimal value $V(\boldsymbol{\xi}, \mathcal{F}_k)$ of (4.1) as $\boldsymbol{\xi}_n$ converges in distribution toward $\boldsymbol{\xi}$ (see [3, Theorem IV.26]). Note that this step involves only open-loop problems, which are approximated using the traditional Monte Carlo approach.

In this approach, the global discretization error $\left|V(\boldsymbol{\xi}, \mathcal{F}) - V(\boldsymbol{\xi}_n, \mathcal{F}_k)\right|$ is bounded from above by the sum of two terms:

- $\left|V(\boldsymbol{\xi}, \mathcal{F}) - V(\boldsymbol{\xi}, \mathcal{F}_k)\right|$. *Information structure discretization error*,
- $\left|V(\boldsymbol{\xi}, \mathcal{F}_k) - V(\boldsymbol{\xi}_n, \mathcal{F}_k)\right|$. *Mean computation discretization error*.

Apart from the convergence result itself, this approach enlightens the fact that it is not sufficient to properly deal with the last term (Monte Carlo) in order to obtain a "good" approximation of problem (1.1).

The main difference is thus that problem (1.1) is approximated in a nested manner in Barty's approach, whereas we *simultaneously* approximate the σ-field and the random variable.

Pennanen's approach. In [12], Pennanen adresses a stochastic optimization problem very similar to problem (1.1). He assumes that the observation $\boldsymbol{y}$ is a function[6] of the noise $\boldsymbol{\xi}$:

$$\boldsymbol{y} = h(\boldsymbol{\xi}).$$

Then the problem can be formulated on the probability space $(\Xi, \mathcal{B}_\Xi, \mu)$, μ being the probability distribution of ξ, rather than on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Pennanen chooses a quantification operator q_n on Ξ, leading to an approximated random variable $\boldsymbol{\xi}_n$,

$$\boldsymbol{\xi}_n = q_n(\boldsymbol{\xi}),$$

and then deduces the information quantization from the noise quantization by setting

$$\boldsymbol{y}_n = h(\boldsymbol{\xi}_n) = h \circ q_n(\boldsymbol{\xi}).$$

Now there is no reason for the quantified observation $\boldsymbol{y}_n$ to be measurable with respect to the initial observation $\boldsymbol{y}$. In order to overcome the difficulty, Pennanen assumes that, in terms of subfields of $\mathcal{B}_\Xi$, the following inclusion holds:

$$\sigma(h \circ q_n) \subset \sigma(h). \tag{4.3}$$

Note that condition (4.3) means that if two samples ξ and ξ' of $\boldsymbol{\xi}$ lead to identical observations, so do the quantified noises $q_n(\xi)$ and $q_n(\xi')$. In the dynamic framework of problem (1.2), this assumption implies that the sampled trajectories of the noise are organized in a scenario tree.

The main difference is thus that the approximation of the σ-field is intimately related to the approximation of the random variable in Pennanen's approach, requiring the additional assumption (4.3), whereas these two approximations are designed *separately* in our approach.[7]

[6] An assumption which can be made without loss of generality.

[7] Note, however, that Pennanen's approach is also designed to handle extended-real-valued functions.

Heitsch's approach. In [10], Heitsch, Römisch and Strugarek address a linear multistage stochastic optimization problem subject to nonanticipative constraints. Denoting, respectively, by $\boldsymbol{\xi} = \left(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_T\right) \in L^q(\Omega, \mathcal{A}, \mathbb{P}; \Xi)$ and by $\boldsymbol{u} = \left(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T\right) \in L^r(\Omega, \mathcal{A}, \mathbb{P}; U)$ the noise and the decision stochastic processes, the nonanticipativity constraints "$\boldsymbol{u}_t$ is $\sigma(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_t)$-measurable" are summarized in the compact form $\boldsymbol{u} \in \mathcal{U}(\boldsymbol{\xi})$ so that the optimization problem writes

$$v(\boldsymbol{\xi}) := \min_{\boldsymbol{u} \in \mathcal{U}(\boldsymbol{\xi})} J(\boldsymbol{u}, \boldsymbol{\xi}).$$

The main result [10, Theorem 2.1] states that, under technical assumptions, there exist positive constants L, α, and δ such that the following majoration

$$\left|v(\boldsymbol{\xi}) - v(\boldsymbol{\xi}^{'})\right| \leq L\Big(\left\|\boldsymbol{\xi} - \boldsymbol{\xi}^{'}\right\|_{L^q} + D_\alpha\big(\boldsymbol{\xi}, \boldsymbol{\xi}^{'}\big)\Big) \tag{4.4}$$

holds for all noise processes $\boldsymbol{\xi}^{'}$ such that $\|\boldsymbol{\xi} - \boldsymbol{\xi}^{'}\|_{L^q} \leq \delta$. Here D_α is a given function measuring the distance between the filtrations of $\boldsymbol{\xi}$ and its approximation $\boldsymbol{\xi}^{'}$.

The last result seems to be very similar to the one given in section 2. It, however, differs significantly on the two following points.

- Only one discretization is considered in [10], namely, the noise discretization. The consequence of this discretization is then measured (using function D_α) in terms of information. General conditions ensuring that the term $D_\alpha(\boldsymbol{\xi}, \boldsymbol{\xi}^{'})$ is close to zero are, however, not given in the main theorem of [10].
- According to the authors, the filtration distance D_α is related to the uniform convergence topology on σ-fields. Such a distance allows for a Lipschitz-like result (4.4) which makes sense[8] but presents the drawback that finite partitions are not dense in the space $\mathcal{A}^\star$ equipped with the uniform convergence topology. Approximating information does not appear to be the main purpose in [10], as illustrated by the example in [10, Example 2.7], where the distance D_α is used to reduce an existing scenario tree rather than to approximate a σ-field.

The main difference is thus that [10] essentially presents a stability result, whereas we are concerned with a convergence result.

5. Conclusion. Approximating a stochastic optimization problem somewhat naturally leads to discretizing the underlying noise or its probability distribution. This procedure has been examined by various authors who have given appropriate conditions for convergence (see [8], [16], and [1]).

In our present work, we provide a framework to deal with the approximation of stochastic optimization problems *subject to measurability constraints*. By means of a simple example, we show how the standard Monte Carlo approximation may fail, although it is generally adapted to the situation without measurability constraints. We also point out that deducing an information discretization uncautiously from the noise discretization may also lead to wrong results. Our main contribution consists in treating separately noise and information discretization, providing appropriate separate sufficient conditions to obtain convergence.

[8] Whereas the distance associated with the strong convergence topology is in fact *arbitrary* and thus not well-suited for a Lipschitz property.

Acknowledgments. This paper is based on earlier work by the SOWG (École Nationale des Ponts et Chaussées); we particularly thank Kengy Barty, Guy Cohen, Anes Dallagi, and Cyrille Strugarek. We also want to thank Professor Roger J.-B. Wets for his kind advice on this work.

REFERENCES

[1] Z. Artstein and R. J.-B. Wets, *Consistency of minimizers and the SLLN for stochastic programs*, J. Convex Anal., 2 (1995), pp. 1–17.
[2] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser, Boston, 1988.
[3] K. Barty, *Contributions à la Discrétisation des Contraintes de Mesurabilité, Pour les Problèmes d'Optimisation Stochastique*, Ph.D. thesis, Ecole des Ponts ParisTech, 2004. http://cermics.enpc.fr/theses.
[4] P. Billingsley, *Convergence of Probability Measures*, Wiley, New York, 1999.
[5] E. Boylan, *Equi-convergence of martingales*, Ann. Math. Statist., 42 (1971), pp. 552–559.
[6] J.-P. Chancelier and SOWG, *Epi-convergence of stochastic optimization problems involving both random variables and measurability constraints approximations*, Prépublication du CERMICS, 2006, http://cermics.enpc.fr/reports.
[7] K. Cotter, *Similarity of information and behavior with a pointwise convergence topology*, J. Math. Econom., 15 (1986), pp. 25–38.
[8] J. Dupacova and R. J.-B. Wets, *Asymptotic behavior of statistical estimators and of solutions of stochastic optimization problems*, Ann. Statist., 16 (1988), pp. 1517–1549.
[9] A. Gersho and R. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, Norwell, MA, 1991.
[10] H. Heitsch, W. Römisch, and C. Strugarek, *Stability of multistage stochastic programs*, SIAM J. Optim., 17 (2006), pp. 511–525.
[11] H. Kudo, *A note on the strong convergence of σ-algebras*, Ann. Probab., 2 (1974), pp. 76–83.
[12] T. Pennanen, *Epi-convergent discretizations of multistage stochastic programs*, Math. Oper. Res., 30 (2005), pp. 245–256.
[13] L. Piccinini, *Convergence of nonmonotone sub-σ-fields and convergence of associated subspaces $L^p(\mathcal{B}_n)$ ($p \in [1, +\infty]$)*, J. Math. Anal. Appl., 225 (1998), pp. 73–90.
[14] B. Polyak, *Convergence and convergence rate of iterative stochastic algorithms*, Automat. Remote Control, 37 (1976), pp. 1858–1868.
[15] R. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Ser. Compr. Stud. Math. 317, Springer-Verlag, Berlin, Heidelberg, 1998.
[16] A. Shapiro and T. Homem-de-Mello, *On the rate of convergence of optimal solutions of Monte Carlo approximations of stochastic programs*, SIAM J. Optim., 11 (2000), pp. 70–86.
[17] A. Shapiro, *Inference of statistical bounds for multistage stochastic programming problems*, Math. Methods Oper. Res., 58 (2003), pp. 57–68.
[18] C. Strugarek and SOWG, *On the Fortet-Mourier Metric for the Stability of Stochastic Optimization Problems, an Example*, Stoch. Programming E-Print Ser., 25 (2004), http://speps.org.
[19] C. Strugarek, *Approches Variationnelles et Autres Contributions en Optimisation Stochastique*, Ph.D. thesis, Ecole des Ponts ParisTech, 2006. http://cermics.enpc.fr/theses.

SIAM J. OPTIM.
Vol. 19, No. 4, pp. 1735–1756

STRONG DUALITY FOR THE CDT SUBPROBLEM: A NECESSARY AND SUFFICIENT CONDITION*

WENBAO AI† AND SHUZHONG ZHANG‡

Abstract. In this paper, we consider the problem of minimizing a nonconvex quadratic function, subject to two quadratic inequality constraints. As an application, such a quadratic program plays an important role in the trust region method for nonlinear optimization; such a problem is known as the Celis, Dennis, and Tapia (CDT) subproblem in the literature. The Lagrangian dual of the CDT subproblem is a semidefinite program (SDP), hence convex and solvable. However, a positive duality gap may exist between the CDT subproblem and its Lagrangian dual because the CDT subproblem itself is nonconvex. In this paper, we present a necessary and sufficient condition to characterize when the CDT subproblem and its Lagrangian dual admits no duality gap (i.e., the strong duality holds). This necessary and sufficient condition is easy verifiable and involves only one (any) optimal solution of the SDP relaxation for the CDT subproblem. Moreover, the condition reveals that it is actually rare to render a positive duality gap for the CDT subproblems in general. Moreover, if the strong duality holds, then an optimal solution for the CDT problem can be retrieved from an optimal solution of the SDP relaxation, by means of a matrix rank-one decomposition procedure. The same analysis is extended to the framework where the necessary and sufficient condition is presented in terms of the Lagrangian multipliers at a KKT point. Furthermore, we show that the condition is numerically easy to work with approximatively.

Key words. quadratically constrained quadratic programming, strong Lagrangian duality, Celis, Dennis, and Tapia subproblem, semidefinite program relaxation

AMS subject classifications. 90C33, 90C51, 90C05

DOI. 10.1137/07070601X

1. Introduction.

In this paper, we consider the following nonconvex quadratic optimization problem:

$$(QP) \quad \begin{array}{ll} \text{minimize} & q_0(x) = x^T Q_0 x - 2b_0^T x \\ \text{subject to} & q_i(x) = x^T Q_i x - 2b_i^T x + c_i \le 0, \quad i = 1, \ldots, m. \end{array}$$

In case $m = 1$ and $Q_1 \succ 0$, the problem is known as the *trust region subproblem*, since in the trust region approach to unconstrained optimization, such problems need to be solved repeatedly. In this context, the problem has been thoroughly studied. (For general information on the trust region method, see [6]). It is known that the trust region subproblem can be easily solved. A connection between the solution methods for the trust region subproblem and semidefinite programming (SDP) was established by Sturm and Zhang in [11]. By using a matrix rank-one decomposition procedure, Sturm and Zhang [11] showed that if $m = 1$, then the SDP relaxation of (QP) is tight, and an optimal solution for (QP) can be obtained from an optimal solution of its SDP relaxation. Furthermore, Ye and Zhang [13] showed that if $m = 2$

*Received by the editors October 22, 2007; accepted for publication (in revised form) August 26, 2008; published electronically February 11, 2009.
http://www.siam.org/journals/siopt/19-4/70601.html

†School of Science, Beijing University of Posts and Telecommunications, Beijing 100876, People's Republic of China (wenbaoai@vip.sina.com). This author's work was partially supported by Chinese NSFC Earmarked Grant Project 10701016.

‡Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, Hong Kong (zhang@se.cuhk.edu.hk). This author's work was supported by Hong Kong RGC Earmarked Grant CUHK418505.

and certain additional conditions are satisfied, then the SDP relaxation for (QP) can still be tight in many cases. In fact, the quadratic program (QP) with $m = 2$ has its own history as an extended trust region subproblem. In 1985, Celis, Dennis and Tapia (CDT) [3] proposed a trust region method for constrained optimization, in which (QP) with $m = 2$ plays the role as a model for validating a trust region step. In this particular context, $Q_1 \succ 0$ and $Q_2 \succeq 0$, and the extended trust region subproblem is also referred to as the CDT subproblem. A number of papers have been devoted to studying the structure and the solution algorithms for the CDT subproblem; see, e.g., [4, 5, 9, 10, 11, 13, 14, 15, 16].

A remarkable property which makes the CDT subproblem interesting and intriguing is that at a global optimal solution, the Hessian matrix of the Lagrangian function may not necessarily be positive semidefinite; however, it can have at most one negative eigenvalue (see Yuan [14]). In fact, it is quite rare to encounter examples where the Hessian of the Lagrangian function indeed has a negative eigenvalue at optimum. In 1991, Yuan [15] suggested an algorithm for the CDT subproblem under the assumption that the objective function is convex, and, in 1992, Zhang [16] proposed an algorithm for the CDT subproblem under the assumption that the optimal Lagranian Hessian matrix is positive semidefinite. Chen and Yuan [5] presented a sufficient condition (termed as *Property* $\mathcal{J}$ in [5]) under which the Lagrangian function of the CDT subproblem will have a positive semidefinite Hessian at optimal point. Recently, Beck and Eldar [1] used the complex valued SDP (thus relaxed) approach to come up with a similar sufficient condition to guarantee the nonnegativity of the Hessian matrix of the Lagrangian function at optimum. Beck and Eldar [1] reported that in their experiments on randomly generated instances, their sufficient condition was satisfied for an overwhelming majority of the random instances.

The current paper is concerned with the CDT-type quadratic programs. In particular, we shall present a verifiable condition which indicates whether or not the SDP relaxation for the quadratic program is tight. Since the Lagrangian dual of a general quadratically constrained quadratic program is the dual of its SDP relaxation (see Chapter 13 of [12]), our result is equivalent to a necessary and sufficient condition for the strong duality to hold for this class of nonconvex quadratic programs. Our condition involves only the information of an optimal SDP solution, or alternatively, the information of a given KKT point. The paper is organized as follows. In section 2, we shall formally establish the equivalence between the nonnegativity of the Hessian matrix of the Lagrangian function (of (QP)) at an arbitrary optimal solution and the fact that the SDP relaxation is tight. Section 3 is devoted to a specific problem related to the rank-one decomposition of a positive semidefinite matrix. This technical result is interesting in its own right, and it is used in section 4 to derive a *necessary and sufficient* condition to check whether or not the SDP relaxation is indeed tight. Because the dual of the SDP relaxation coincides with the Lagrangian dual of (QP), a tight SDP relaxation manifests that the strong duality holds for (QP). Our necessary and sufficient condition is different from the other two sufficient conditions previously studied in [5] and [1]. In nonlinear programming, it is customary to use terminologies such as the Lagrangian multipliers or the KKT conditions. For this reason, we shall present our results in section 5 both as an easy verifiable condition based on an optimal solution of the SDP relaxation or, alternatively, as an easy verifiable condition based on a KKT point in terms of the Lagrangian function and multipliers. An example is given in section 6 to show that the information carried by the KKT solutions may not be useful for the optimal solution of the CDT problem when the strong duality fails. In section 7, we propose a numerical implementation of the necessary and sufficient

condition. Our simulation results show that the condition is indeed numerically stable and easy to work with.

Throughout the paper, $\mathcal{S}^{n\times n}$ denotes the set of real $n\times n$ symmetric matrices; $\mathcal{S}^{n\times n}_{+}$ denotes the set of real $n\times n$ positive semidefinite matrices; $\mathcal{S}^{n\times n}_{++}$ denotes the set of real $n\times n$ positive definite matrices; for $A, B\in\mathcal{S}^{n\times n}$, $A\bullet B := \operatorname{tr} AB$ denotes the matrix inner-product between A and B.

2. Convex Lagrangian function and the strong duality. In the literature, there are mainly two ways to solve a general quadratically constrained quadratic program (QP): Either use the Lagrangian function with some appropriately chosen multipliers or base the solution method on the SDP relaxation. In the latter case, the method works well if the SDP relaxation is tight, while, in the former case, the method works well if the Hessian of the Lagrangian function is positive semidefinite. It is therefore natural to believe that these two properties must be essentially identical. In this section, we shall formally prove this point. The result is useful for our subsequent analysis.

First of all, following [11], we use the notation

$$M(q_0) := \begin{bmatrix} 0 & -b_0^T \\ -b_0 & Q_0 \end{bmatrix},\ M(q_i) := \begin{bmatrix} c_i & -b_i^T \\ -b_i & Q_i \end{bmatrix},\ \text{for } i=1,\dots,m.$$

Then, (QP) is equivalently written as

$$\begin{aligned} (QP)\quad \text{minimize}\quad & M(q_0)\bullet \begin{bmatrix} t \\ x \end{bmatrix}\begin{bmatrix} t \\ x \end{bmatrix}^T = x^TQ_0x - 2b_0^Txt \\ \text{subject to}\quad & M(q_i)\bullet \begin{bmatrix} t \\ x \end{bmatrix}\begin{bmatrix} t \\ x \end{bmatrix}^T = x^TQ_ix - 2b_i^Txt + c_it^2 \le 0, \quad i=1,\dots,m \\ & t^2 = 1. \end{aligned}$$

The so-called SDP relaxation of (QP) is

$$\begin{aligned} (SP)\quad \text{minimize}\quad & M(q_0)\bullet X \\ \text{subject to}\quad & M(q_i)\bullet X \le 0, \quad i=1,\dots,m \\ & I_{00}\bullet X = 1 \\ & X \succeq 0, \end{aligned}$$

where $I_{00} = \begin{bmatrix} 1 & 0 \\ 0 & \mathbf{0} \end{bmatrix} \in \mathcal{S}^{(n+1)\times(n+1)}$. The dual problem of (SP) is

$$\begin{aligned} (SD)\quad \text{maximize}\quad & y_0 \\ \text{subject to}\quad & Z = M(q_0) - y_0I_{00} + \textstyle\sum_{i=1}^m y_iM(q_i) \succeq 0 \\ & y_i \ge 0, i=1,\dots,m. \end{aligned}$$

Note that (SD) is also the Lagrangian dual problem for (QP) (see [12]). The following well-known facts regarding the relationship between (SP) and (SD) are either straightforward or well known:

(1) (SP) satisfies the Slater condition if the original problem (QP) satisfies the Slater condition.
(2) (SD) satisfies the Slater condition if at least one of the matrices Q_i's, $i = 0,1,\dots,m$, is positive definite.
(3) If both (SP) and (SD) satisfy the Slater condition, then (SP) and (SD) have attainable optimal solutions. Moreover, a primal-dual feasible pair X and

$(Z, y_0, y_1, \ldots, y_m)$ are optimal if and only if they satisfy the complementary conditions:

$$XZ = 0, \quad y_i M(q_i) \bullet X = 0, \quad i = 1, \ldots, m.$$

Throughout this paper, we assume that $Q_1 \succ 0$ and that (QP) satisfied the Slater condition. Hence, (QP), (SP), and (SD) all have optimal solutions, which we shall denote, respectively, by x^*, $\hat{X}$, and $(\hat{Z}, \hat{y_0}, \hat{y_1}, \ldots, \hat{y_m})$, and their optimal values, respectively, by $v(QP)$, $v(SP)$, and $v(SD)$.

Clearly, $v(SP) \leq v(QP)$ since (SP) is a relaxation of (QP), and $v(SP) = v(SD)$ since both (SP) and (SD) satisfy the Slater condition. Therefore, the strong duality holds for (QP) if and only if the SDP relaxation for (QP) is tight; i.e., $v(SP) = v(QP)$. It is helpful to keep in mind that $\hat{Z}$ can also be rewritten as

$$\hat{Z} = \begin{bmatrix} -\hat{y}_0 + \sum_{i=1}^m \hat{y}_i c_i & -b_0^T - \sum_{i=1}^m \hat{y}_i b_i^T \\ -b_0 - \sum_{i=1}^m \hat{y}_i b_i & Q_0 + \sum_{i=1}^m \hat{y}_i Q_i \end{bmatrix}.$$

On the other hand, the Lagrangian function for (QP), with y_i being the multiplier for the constraint $q_i(x) \leq 0$, $i = 1, \ldots, m$, is given as

$$L(x; y) := q_0(x) + \sum_{i=1}^m y_i q_i(x).$$

Clearly, since the function is quadratic in x for any fixed multiplier y, its Hessian matrix is $\nabla^2_{xx} L(x; y) = Q_0 + \sum_{i=1}^m y_i Q_i$.

THEOREM 2.1. *$v(SP) = v(QP) \iff \nabla^2_{xx} L(x; y) = Q_0 + \sum_{i=1}^m y_i Q_i \succeq 0$, where y is the Lagrangian multiplier for an optimal solution of (QP).*

Proof. "$\Longrightarrow$": For any minimizer x^* of the original problem (QP), the matrix $X^* := \begin{bmatrix} 1 \\ x^* \end{bmatrix} \begin{bmatrix} 1 \\ x^* \end{bmatrix}^T$ is also an optimal solution for (SP). So the primal-dual optimal pair X^* and $(\hat{Z}, \hat{y})$ satisfy complementary conditions, where $(\hat{Z}, \hat{y})$ is optimal to (SD), i.e.,

$$\hat{Z} X^* = 0, \quad \hat{y}_i M(q_i) \bullet X^* = 0, \quad i = 1, \ldots, m. \tag{2.1}$$

Since $\hat{Z} \succeq 0$, the relation $\hat{Z} X^* = 0$ is equivalent to $\hat{Z} \begin{bmatrix} 1 \\ x^* \end{bmatrix} = 0$, which implies that

$$\left(Q_0 + \sum_{i=1}^m \hat{y}_i Q_i \right) x^* = b_0 + \sum_{i=1}^m \hat{y}_i b_i.$$

Also, since $q_i(x^*) = M(q_i) \bullet X^*$, it follows from (2.1) that $\hat{y}_i q_i(x^*) = 0$, $i = 1, \ldots, m$. Therefore, x^* and $\hat{y}$ satisfy the KKT condition, and $\hat{y}$ is the corresponding Lagrangian multiplier with the Hessian matrix being

$$\nabla^2_{xx} \left(q_0(x) + \sum_{i=1}^m \hat{y}_i q_i(x) \right) \Bigg|_{x = x^*} = Q_0 + \sum_{i=1}^m \hat{y}_i Q_i = \hat{Z} \succeq 0.$$

"$\Longleftarrow$": Suppose that the original problem (QP) has an optimal solution x^*, with a positive semidefinite Lagrangian Hessian matrix $Q_0 + \sum_{i=1}^m y_i^* Q_i$. Then x^* and

$y_1^*, \ldots, y_m^*$ satisfy the following KKT condition:

$$\left(Q_0 + \sum_{i=1}^m y_i^* Q_i\right) x^* = b_0 + \sum_{i=1}^m y_i^* b_i, \quad y_i^* q_i(x^*) = 0, \quad y_i^* \geq 0, i = 1, \ldots, m.$$

Let

$$\left\{ \begin{array}{l} X^* := \left[\begin{array}{c} 1 \\ x^* \end{array} \right] \left[\begin{array}{c} 1 \\ x^* \end{array} \right]^T, \quad y_0^* := q_0(x^*), \\ \text{and } Z^* := M(q_0) - y_0^* I_{00} + \sum_{i=1}^m y_i^* M(q_i). \end{array} \right. \tag{2.2}$$

Next, we aim to show that X^* and (y^*, Z^*) are, in fact, optimal to (SP) and (SD). To this end, we need only to verify that $Z^* \succeq 0$ and $Z^* X^* = 0$.

Let the Lagrangian function be

$$L(x; y^*) = q_0(x) + \sum_{i=1}^m y_i^* q_i(x). \tag{2.3}$$

By the Taylor expansion at x^* and the KKT optimality condition, we have

$$\begin{aligned} L(x; y^*) &= L(x^*; y^*) + (x - x^*)^T \left(Q_0 + \sum_{i=1}^m y_i^* Q_i\right)(x - x^*) \\ &\geq L(x^*; y^*) = q_0(x^*) = y_0^* \end{aligned} \tag{2.4}$$

for any x, which implies that x^* is a global minimizer of $L(x; y^*)$. Consider any $(n+1)$-dimensional vector $\left[\begin{smallmatrix} t \\ x \end{smallmatrix}\right]$. If $t \neq 0$, then it follows from (2.2), (2.3), and (2.4) that

$$\begin{aligned} \left[\begin{array}{c} t \\ x \end{array} \right]^T Z^* \left[\begin{array}{c} t \\ x \end{array} \right] &= \left(M(q_0) - y_0^* I_{00} + \sum_{i=1}^m y_i^* M(q_i) \right) \bullet \left[\begin{array}{c} t \\ x \end{array} \right] \left[\begin{array}{c} t \\ x \end{array} \right]^T \\ &= t^2 \left(q_0(x/t) + \sum_{i=1}^m y_i^* q_i(x/t) - y_0^* \right) \\ &= t^2 \left(L(x/t; y^*) - y_0^* \right) \geq 0. \end{aligned}$$

If $t = 0$, then

$$\left[\begin{array}{c} t \\ x \end{array} \right]^T Z^* \left[\begin{array}{c} t \\ x \end{array} \right] = x^T \left(Q_0 + \sum_{i=1}^m y_i^* Q_i \right) x \geq 0.$$

Therefore, $Z^* \succeq 0$. Moreover,

$$Z^* \bullet X^* = \left[\begin{array}{c} 1 \\ x^* \end{array} \right]^T Z^* \left[\begin{array}{c} 1 \\ x^* \end{array} \right] = L(x^*; y^*) - y_0^* = 0,$$

which, together with $Z^* \succeq 0$ and $X^* \succeq 0$, implies that $Z^* X^* = 0$. □

Theorem 2.1 implies that once an optimal solution for (QP) admits a Lagrangian multiplier (vector) with nonnegative Hessian matrix, then $v(QP) = v(SP)$, which in

turn implies that *every* optimal solution admits a Lagrangian multiplier with nonnegative Hessian matrix. We formalize the statement as follows.

COROLLARY 2.2. *If at one optimal solution of* (QP) *there is a Lagrangian multiplier with a nonnegative Hessian matrix, then it follows that at any optimal solution of* (QP)*, there is a Lagrangian multiplier with a nonnegative Hessian matrix.*

We note that Theorem 2.1 can actually be used to bridge an easy provable fact to a less obvious one. For instance, it is relatively easy to show that if $m = 1$, then the Hessian of the Lagrangian function is nonnegative (e.g., Theorem 7.2.1 in [6]). Hence we can conclude $v(QP) = v(SP) = v(SD)$ in this case, simply using Theorem 2.1. On the other hand, if $m = 2$ and (QP) is homogeneous (i.e., $b_0 = b_1 = b_2 = 0$), then Ye and Zhang (section 2.2 of [13]) showed that $v(QP) = v(SP)$. A less obvious fact is that the Lagrangian function always has a nonnegative Hessian matrix in this case.

3. A new matrix rank-one decomposition procedure. Sturm and Zhang [11] proposed a simple (polynomial-time) procedure to compute the following matrix rank-one decomposition problem: Given $X \in \mathcal{S}_+^{n\times n}$ and $A \in \mathcal{S}^{n\times n}$, find $x_j \in \Re^n$, $j = 1, \ldots, r$, where $r = \text{rank}(A)$ such that $X = \sum_{j=1}^r x_j x_j^T$ and $x_j^T A x_j = A \bullet X/r$, $j = 1, \ldots, r$. Huang and Zhang [8] extended the result to the case where the matrices in questions are all Hermitian.

The aim of this section is to study a further extension of such rank-one decomposition in the real symmetric case. Our result will then be applied in the next section to enable a method for (QP) when $m = 2$. Let $x_1 \in \Re^n$ and $X \in \mathcal{S}_+^{n\times n}$. As a convention, we shall call matrix X to be *rank-one decomposable* at x_1 if there exist other $r - 1$ vectors $x_2, \ldots, x_r$ such that $X = x_1x_1^T + x_2x_2^T + \cdots + x_rx_r^T$, where $r = \text{rank}(X)$. To find out when X is a matrix rank-one decomposable at a given vector, we first note the following lemma.

LEMMA 3.1. *Suppose that* $X \in \mathcal{S}_+^{n\times n}$*, with* $\text{rank}(X) = r$ *and* $X = x_1x_1^T + x_2x_2^T + \cdots + x_rx_r^T$*. Let* $X_r = [x_1, \ldots, x_r]$*. Then,* $X = y_1y_1^T + y_2y_2^T + \cdots + y_ry_r^T$ *holds if and only if there exists an orthonormal matrix* $P \in \Re^{r\times r}$ *such that* $Y_r = X_rP$ *where* $Y_r = [y_1, \ldots, y_r]$.

Proof. The sufficiency is obvious. To show the necessity of the condition, let us suppose $X = X_rX_r^T = Y_rY_r^T$ and consider $P = X_r^TY_r(Y_r^TY_r)^{-1}$. Clearly,

$$P^TP = (Y_r^TY_r)^{-1}Y_r^TX_rX_r^TY_r(Y_r^TY_r)^{-1} = (Y_r^TY_r)^{-1}Y_r^TY_rY_r^TY_r(Y_r^TY_r)^{-1} = I_r.$$

Hence P is an orthonormal matrix. At the same time,

$$X_rP = X_rX_r^TY_r(Y_r^TY_r)^{-1} = Y_rY_r^TY_r(Y_r^TY_r)^{-1} = Y_r. \qquad \square$$

Since for any given unit vector one can always construct an orthonormal matrix with this unit vector as the first column, this leads to the following characterization of the rank-one decomposability at a given vector.

PROPOSITION 3.2. *Suppose that* $X \in \mathcal{S}_+^{n\times n}$*, with* $\text{rank}(X) = r$ *and* $X = x_1x_1^T + x_2x_2^T + \cdots + x_rx_r^T$*. Let* $X_r = [x_1, \ldots, x_r]$*. Then,* X *is rank-one decomposable at* $y \in \Re^n$ *if and only if there is* $u \in \Re^r$*, with* $\|u\| = 1$ *and* $y = X_ru$*.*

The next result plays an important role in this paper.

LEMMA 3.3. *Let* $A_1, A_2 \in \mathcal{S}^{n\times n}$*. Suppose that* $X = x_1x_1^T + x_2x_2^T + \cdots + x_rx_r^T$*, where* $r \geq 3$*. If*

$$\begin{aligned} &A_1 \bullet x_1x_1^T = A_1 \bullet x_2x_2^T = \delta_1, \\ &(A_2 \bullet x_1x_1^T - \delta_2)(A_2 \bullet x_2x_2^T - \delta_2) < 0, \end{aligned} \tag{3.1}$$

then in the real-number computation sense (viz. the computational model of Blum, Shub, and Smale [2]*, which we shall abbreviate as the BBS model hereafter), one can find in polynomial-time a vector* $y \in \Re^n$ *such that* X *is rank-one decomposable at* y *and*

$$\begin{aligned} A_1 \bullet yy^T &= \delta_1, \\ A_2 \bullet yy^T &= \delta_2. \end{aligned} \tag{3.2}$$

Proof. Without loss of generality, we assume that

$$A_2 \bullet x_1x_1^T - \delta_2 > 0 \text{ and } A_2 \bullet x_2x_2^T - \delta_2 < 0. \tag{3.3}$$

For given real values α_i, $i = 1, 2, 3$, with $(\alpha_1, \alpha_2, \alpha_3) \neq 0$, define

$$y = \frac{\alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3}{\sqrt{\alpha_1^2 + \alpha_2^2 + \alpha_3^2}}. \tag{3.4}$$

By Proposition 3.2, X is rank-one decomposable at y. Let us substitute (3.4) into (3.2) and consider the following system of equations with respect to the unknown real variables α_1, α_2, and α_3:

$$\begin{aligned} 0 = {} & \alpha_3^2(A_1 \bullet x_3x_3^T - \delta_1) + 2\alpha_1\alpha_2 A_1 \bullet x_1x_2^T \\ & + 2\alpha_1\alpha_3 A_1 \bullet x_1x_3^T + 2\alpha_2\alpha_3 A_1 \bullet x_2x_3^T, \end{aligned} \tag{3.5}$$

$$\begin{aligned} 0 = {} & \alpha_1^2(A_2 \bullet x_1x_1^T - \delta_2) + \alpha_2^2(A_2 \bullet x_2x_2^T - \delta_2) + \alpha_3^2(A_2 \bullet x_3x_3^T - \delta_2) \\ & + 2\alpha_1\alpha_2 A_2 \bullet x_1x_2^T + 2\alpha_1\alpha_3 A_2 \bullet x_1x_3^T + 2\alpha_2\alpha_3 A_2 \bullet x_2x_3^T. \end{aligned} \tag{3.6}$$

In fact, it follows from Finsler's lemma [7] that (3.5) and (3.6) admit a real-valued solution $(\alpha_1, \alpha_2, \alpha_3)$. However, Finsler's lemma is a pure existence result. Below we shall construct such solutions. We proceed by considering two cases.

Case 1. $A_1 \bullet x_1x_2^T = 0$.

We choose $\alpha_1 = 1$ and $\alpha_3 = 0$. Then (3.5) is trivially satisfied for any values of α_2, and (3.6) can be rewritten as follows:

$$(A_2 \bullet x_1x_1^T - \delta_2) + \alpha_2^2(A_2 \bullet x_2x_2^T - \delta_2) + 2\alpha_2 A_2 \bullet x_1x_2^T = 0,$$

which is a quadratic equation in α_2 and must have two distinct real roots because of (3.3); one is positive, and another is negative. Let $\bar{\alpha}_2$ be one of the roots. Then $(\alpha_1, \alpha_2, \alpha_3) = (1, \bar{\alpha}_2, 0)$ is a solution for (3.5) and (3.6).

Case 2. $A_1 \bullet x_1x_2^T \neq 0$.

We choose $\alpha_3 = 1$. Then (3.5) and (3.6) become

$$0 = 2\alpha_2\left(\alpha_1 A_1 \bullet x_1x_2^T + A_1 \bullet x_2x_3^T\right) + 2\alpha_1 A_1 \bullet x_1x_3^T + (A_1 \bullet x_3x_3^T - \delta_1), \tag{3.7}$$

$$\begin{aligned} 0 = {} & \alpha_1^2(A_2 \bullet x_1x_1^T - \delta_2) + \alpha_2^2(A_2 \bullet x_2x_2^T - \delta_2) + 2\alpha_1\alpha_2 A_2 \bullet x_1x_2^T \\ & + 2\alpha_1 A_2 \bullet x_1x_3^T + 2\alpha_2 A_2 \bullet x_2x_3^T + (A_2 \bullet x_3x_3^T - \delta_2). \end{aligned} \tag{3.8}$$

Solving (3.7) yields

$$\alpha_2 = -\frac{2\alpha_1 A_1 \bullet x_1x_3^T + (A_1 \bullet x_3x_3^T - \delta_1)}{2(\alpha_1 A_1 \bullet x_1x_2^T + A_1 \bullet x_2x_3^T)} \triangleq p(\alpha_1). \tag{3.9}$$

Moreover, let us denote

$$\begin{aligned} g(\alpha_1, \alpha_2) := {} & \alpha_1^2(A_2 \bullet x_1x_1^T - \delta_2) + \alpha_2^2(A_2 \bullet x_2x_2^T - \delta_2) + 2\alpha_1\alpha_2 A_2 \bullet x_1x_2^T \\ & + 2\alpha_1 A_2 \bullet x_1x_3^T + 2\alpha_2 A_2 \bullet x_2x_3^T + (A_2 \bullet x_3x_3^T - \delta_2) \end{aligned} \tag{3.10}$$

and define

$$t_1 := -\frac{A_1 \bullet x_2 x_3^T}{A_1 \bullet x_1 x_2^T}.$$

We consider the following two possible subcases.

Case 2.1. $\det \begin{bmatrix} 2A_1 \bullet x_1 x_3^T & A_1 \bullet x_3 x_3^T - \delta_1 \\ A_1 \bullet x_1 x_2^T & A_1 \bullet x_2 x_3^T \end{bmatrix} \neq 0.$

Since

$$\begin{aligned} &\left(2\alpha_1 A_1 \bullet x_1 x_3^T + (A_1 \bullet x_3 x_3^T - \delta_1)\right)\big|_{\alpha_1 = t_1} \\ &= -\det \begin{bmatrix} 2A_1 \bullet x_1 x_3^T & A_1 \bullet x_3 x_3^T - \delta_1 \\ A_1 \bullet x_1 x_2^T & A_1 \bullet x_2 x_3^T \end{bmatrix} \Big/ A_1 \bullet x_1 x_2^T \neq 0, \end{aligned}$$

the function $p(\alpha_1)$ has the properties that

$$\lim_{\alpha_1 \to t_1} p(\alpha_1) = \infty \tag{3.11}$$

and

$$\lim_{\alpha_1 \to \infty} p(\alpha_1) = -\frac{A_1 \bullet x_1 x_3^T}{A_1 \bullet x_1 x_2^T}. \tag{3.12}$$

Substituting (3.9) into (3.10), we obtain an equation in α_1:

$$\begin{aligned} g(\alpha_1, p(\alpha_1)) := {} & \alpha_1^2 (A_2 \bullet x_1 x_1^T - \delta_2) + p(\alpha_1)^2 (A_2 \bullet x_2 x_2^T - \delta_2) + 2\alpha_1 p(\alpha_1) A_2 \bullet x_1 x_2^T \\ & + 2\alpha_1 A_2 \bullet x_1 x_3^T + 2p(\alpha_1) A_2 \bullet x_2 x_3^T + (A_2 \bullet x_3 x_3^T - \delta_2) \\ = {} & 0, \end{aligned}$$

which is essentially a quartic polynomial equation in α_1. Since

$$\lim_{\alpha_1 \to t_1} g(\alpha_1, p(\alpha_1)) = -\infty$$

and

$$\lim_{\alpha_1 \to \infty} g(\alpha_1, p(\alpha_1)) = +\infty$$

due to (3.3), (3.11), and (3.12), it follows that $g(\alpha_1, p(\alpha_1))$ has at least one real root $\bar{\alpha}_1$ in the interval $(t_1, +\infty)$. Moreover, such root can be found by solving a quartic polynomial equation with the standard root-finding formula, which can be regarded as a constant operation in the BSS computational model. Substituting back, we derive $(\bar{\alpha}_1, p_1(\bar{\alpha}_1), 1)$ as a solution for (3.5) and (3.6).

Case 2.2. $\det \begin{bmatrix} 2A_1 \bullet x_1 x_3^T & A_1 \bullet x_3 x_3^T - \delta_1 \\ A_1 \bullet x_1 x_2^T & A_1 \bullet x_2 x_3^T \end{bmatrix} = 0.$

The above implies that there exists k such that

$$(2A_1 \bullet x_1 x_3^T, A_1 \bullet x_3 x_3^T - \delta_1) = k(A_1 \bullet x_1 x_2^T, A_1 \bullet x_2 x_3^T).$$

Thus (3.7) becomes

$$(\alpha_1 A_1 \bullet x_1 x_2^T + A_1 \bullet x_2 x_3^T)(2\alpha_2 + k) = 0,$$

for which the roots are

$$\alpha_1 = -\frac{A_1 \bullet x_2 x_3^T}{A_1 \bullet x_1 x_2^T} =: t_1, \quad \alpha_2 \text{ arbitrary}$$

and

$$\alpha_2 = -k/2 =: t_2, \quad \alpha_1 \text{ arbitrary.}$$

Substituting them back into (3.8), it suffices to solve either $g(t_1, \alpha_2) = 0$ or $g(\alpha_1, t_2) = 0$, which are quadratic equations in α_2 and α_1, respectively. If $g(t_1, \alpha_2)$ has a real root $\bar{\alpha}_2$, then $(t_1, \bar{\alpha}_2, 1)$ is a solution to (3.5) and (3.6); otherwise, we have

$$g(t_1, \alpha_2) < 0 \text{ for all } \alpha_2$$

as $\lim_{\alpha_2 \to +\infty} g(t_1, \alpha_2) = -\infty$ due to (3.3). In particular,

$$g(t_1, t_2) < 0.$$

Thus $g(\alpha_1, t_2)$ has a real root $\bar{\alpha}_1$ for α_1 on the interval $(t_1, +\infty)$ as $\lim_{\alpha_1 \to +\infty} g(\alpha_1, t_2) = +\infty$ due to (3.3). Then $(\bar{\alpha}_1, t_2, 1)$ is a solution to (3.5) and (3.6). □

Remark that in Lemma 3.3, we require that $r = 3$. This condition cannot be removed. Consider the following example:

$$A_1 = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, A_2 = \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}, X = x_1 x_1^T + x_2 x_2^T = \begin{bmatrix} 1 \\ -1 \end{bmatrix} [1, -1] + \begin{bmatrix} 1 \\ 1 \end{bmatrix} [1, 1].$$

Clearly,

$$A_1 \bullet x_1 x_1^T = A_1 \bullet x_2 x_2^T = 0,$$
$$A_2 \bullet x_1 x_1^T = -2 < 0, A_2 \bullet x_2 x_2^T = 2 > 0.$$

However, for any nonzero $x \in \Re^2$, $A_1 \bullet x x^T = 0$ if and only if x is either parallel to x_1 or to x_2, which implies that there is no nontrivial x satisfying both $A_1 \bullet x x^T = 0$ and $A_2 \bullet x x^T = 0$ simultaneously.

Using the above lemma, we now show the following theorem.

THEOREM 3.4. *Let $A_1, A_2 \in \mathcal{S}^{n \times n}$ and $X \in \mathcal{S}_+^{n \times n}$, with*

$$A_1 \bullet X = \delta_1, \quad A_2 \bullet X = \delta_2.$$

If $r := \text{rank}(X) \geq 3$, then in polynomial-time (real-number computation), one finds a rank-one decomposition for X:

$$X = x_1 x_1^T + x_2 x_2^T + \cdots + x_r x_r^T$$

such that

$$A_1 \bullet x_i x_i^T = \delta_1 / r \text{ for } i = 1, \ldots, r,$$
$$A_2 \bullet x_i x_i^T = \delta_2 / r \text{ for } i = 1, \ldots, r-2.$$

Proof. We shall achieve the desired decomposition by the following steps. Initially, we set $X_0 := \emptyset$ and $X_1 := X$. By Lemma 2.2 of [13], one finds a rank-one decomposition for X_1:

$$X_1 = x_1 x_1^T + x_2 x_2^T + \cdots + x_r x_r^T$$

such that $A_1 \bullet x_i x_i^T = \delta_1/r$ for $i = 1, \ldots, r$. Introduce an index set

$$I_0 := \{i \mid A_2 \bullet x_i x_i^T = \delta_2/r, i = 1, \ldots, r\}$$

and then update X_0 and X_1 by setting

$$X_0 := X_0 + \sum_{i \in I_0} x_i x_i^T, \quad X_1 := X_1 - \sum_{i \in I_0} x_i x_i^T.$$

If $\text{rank}(X_1) < 3$, then the procedure is completed; otherwise, i.e., $\text{rank}(X_1) \geq 3$, using Lemma 3.3, we find y for which X_1 is rank-one decomposable at y such that

$$A_1 \bullet yy^T = \delta_1/r, \quad A_2 \bullet yy^T = \delta_2/r.$$

Update X_0 and X_1 by letting

$$X_0 := X_0 + yy^T, \quad X_1 := X_1 - yy^T.$$

In this case, $\text{rank}(X_1)$ is reduced by 1. Repeat the above procedure until $\text{rank}(X_1) < 3$. □

4. Strong duality: A necessary and sufficient condition. In this section, we consider (QP) with $m = 2$, which shall be denoted $(QP)_2$ hereafter. Without loss of generality, we assume $q_1(x) = x^T x - 1$; i.e.,

$$\begin{array}{lll} (QP)_2 & \text{minimize} & q_0(x) = x^T Q_0 x - 2b_0^T x \\ & \text{subject to} & q_1(x) = x^T x - 1 \leq 0 \\ & & q_2(x) = x^T Q_2 x - 2b_2^T x + c_2 \leq 0. \end{array}$$

The above problem is slightly more general than the CDT subproblem in that Q_2 above can be indefinite. The central issue to be considered here is when the corresponding SDP relaxation for $(QP)_2$ is tight, which is shown in section 2 to be equivalent to a strong Lagrangian duality (alternatively, it is also equivalent to the fact that the Lagrangian function has a positive semidefinite Hessian matrix at optimum due to Theorem 2.1). As before, we assume throughout the discussion that the Slater condition is satisfied by $(QP)_2$.

Let $(SP)_2$ be the SDP relxation for $(QP)_2$ and $(SD)_2$ be the dual of $(SP)_2$; that is,

$$\begin{array}{lll} (SP)_2 & \text{minimize} & M(q_0) \bullet X \\ & \text{subject to} & M(q_1) \bullet X \leq 0 \\ & & M(q_2) \bullet X \leq 0 \\ & & I_{00} \bullet X = 1 \\ & & X \succeq 0, \end{array}$$

where

$$M(q_0) := \begin{bmatrix} 0 & -b_0^T \\ -b_0 & Q_0 \end{bmatrix}, M(q_1) := \begin{bmatrix} -1 & 0 \\ 0 & I_n \end{bmatrix},$$
$$M(q_2) := \begin{bmatrix} c_2 & -b_2^T \\ -b_2 & Q_2 \end{bmatrix}, I_{00} := \begin{bmatrix} 1 & 0 \\ 0 & O_n \end{bmatrix}.$$

As we observed earlier, $(SD)_2$ is also the Lagrangian dual of $(QP)_2$. Let $\hat{X}$ and $(\hat{Z}, \hat{y}_0, \hat{y}_1, \hat{y}_2)$ be a pair of optimal solutions to $(SP)_2$ and to $(SD)_2$, respectively. It turns out that the following property of $\hat{X}$ and $(\hat{Z}, \hat{y}_0, \hat{y}_1, \hat{y}_2)$ is important, which we shall call *Property* $\mathcal{I}$ for ease of reference.

DEFINITION 4.1. *For* $\hat{X}$ *and* $(\hat{Z}, \hat{y}_0, \hat{y}_1, \hat{y}_2)$*, a given pair of optimal solutions for* $(SP)_2$ *and* $(SD)_2$*, respectively, we say that this pair has Property* $\mathcal{I}$ *if:*

(1) $\hat{y}_1\hat{y}_2 \neq 0$*;*
(2) $\operatorname{rank}(\hat{Z}) = n - 1$*;*
(3) $\operatorname{rank}(\hat{X}) = 2$*, and there is a rank-one decomposition of* $\hat{X}$*,* $\hat{X} = \hat{x}_1\hat{x}_1^T + \hat{x}_2\hat{x}_2^T$*, such that* $M(q_1) \bullet \hat{x}_i\hat{x}_i^T = 0$*,* $i = 1, 2$*, and* $(M(q_2) \bullet \hat{x}_1\hat{x}_1^T)(M(q_2) \bullet \hat{x}_2\hat{x}_2^T) < 0$*.*

We remark here that it is easy to verify Property $\mathcal{I}$, once $(SP)_2$ and $(SD)_2$ are solved. The first two conditions being straightforward, the last one, due to Proposition 3.2, can be reduced to verifying the condition on a single parameter satisfying a quadratic equation (any 2-by-2 orthonormal matrix can be completely characterized by polar coordinates in a single parameter).

THEOREM 4.2. *Consider* $(QP)_2$ *where the Slater condition is satisfied. Suppose that* $\hat{X}$ *and* $(\hat{Z}, \hat{y}_0, \hat{y}_1, \hat{y}_2)$ *are a pair of optimal solutions for its SDP relaxation problem* $(SP)_2$ *and the dual* $(SD)_2$*, respectively. Then,* $v((SP)_2) < v((QP)_2)$ *holds if and only if the pair* $\hat{X}$ *and* $(\hat{Z}, \hat{y}_0, \hat{y}_1, \hat{y}_2)$ *has Property* $\mathcal{I}$*.*

Proof. We shall complete the proof in two parts. They are, *Part* 1: If Property $\mathcal{I}$ does not hold, then the SDP relaxation is tight; and *Part* 2: If Property $\mathcal{I}$ holds, then the relaxation must not be tight.

In *Part* 1, we enumerate four exhaustive (but not mutually exclusive) possibilities, to be denoted by *Part* 1.i, with $i = 1, 2, 3, 4$.

Part 1.1. $\hat{y}_1\hat{y}_2 = 0$.

The proof that the SDP relaxation is tight in this case can be found in Ye and Zhang [13].

Part 1.2. $\hat{y}_1\hat{y}_2 \neq 0$ and $\operatorname{rank}(\hat{X}) \neq 2$.

$\hat{y}_1\hat{y}_2 \neq 0$ implies by the complementary conditions that

$$\hat{Z}\hat{X} = 0, \quad M(q_1) \bullet \hat{X} = 0, \quad M(q_2) \bullet \hat{X} = 0.$$

Let $r := \operatorname{rank}(\hat{X})$. Obviously, $r > 0$ since $I_{00} \bullet \hat{X} = 1$, and if $r = 1$, then the theorem is already true. Therefore, we need only to consider the nontrivial case $r \geq 3$. By Theorem 3.4, there is a rank-one decomposition of $\hat{X}$ satisfying

$$\begin{aligned} &X = x_1x_1^T + x_2x_2^T + \cdots + x_rx_r^T \\ &M(q_1) \bullet x_ix_i^T = 0, \text{ for } i = 1, \ldots, r \\ &M(q_2) \bullet x_ix_i^T = 0, \text{ for } i = 1, \ldots, r-2. \end{aligned}$$

Thus $x_1x_1^T/t_1^2$ satisfies the complementary conditions hence optimal to $(SP)_2$. This implies that x_1/t_1 is a homogenized optimal solution to (QP), where t_1 denotes the first element of x_1, which must be nonzero because $M(q_1) \bullet x_1x_1^T = 0$, $x_1 \neq 0$, and $Q_1 \succ 0$.

Part 1.3. $\hat{y}_1\hat{y}_2 \neq 0$ and $\operatorname{rank}(\hat{X}) = 2$, and $M(q_2) \bullet \hat{x}_1\hat{x}_1^T = M(q_2) \bullet \hat{x}_2\hat{x}_2^T = 0$.

In this case, both $\hat{x}_1\hat{x}_1^T/\hat{t}_1^2$ and $\hat{x}_2\hat{x}_2^T/\hat{t}_2^2$ are optimal to $(SP)_2$. Thus both $\hat{x}_1/\hat{t}_1$ and $\hat{x}_2/\hat{t}_2$ are optimal solutions for $(QP)_2$, where $\hat{t}_1$ and $\hat{t}_2$ are the first elements of $\hat{x}_1$ and $\hat{x}_2$, respectively, which are both nonzero as argued before.

Part 1.4. $\hat{y}_1\hat{y}_2 \neq 0$ and $\operatorname{rank}(\hat{X}) = 2$, $\left(M(q_2) \bullet \hat{x}_1\hat{x}_1^T\right)\left(M(q_2) \bullet \hat{x}_2\hat{x}_2^T\right) < 0$, and $\operatorname{rank}(\hat{Z}) \neq n - 1$.

Since $\operatorname{rank}(\hat{Z})+\operatorname{rank}(\hat{X}) \le n+1$ and $\operatorname{rank}(\hat{X}) = 2$, it follows that $\operatorname{rank}(\hat{Z}) \le n-1$, and therefore, in this particular case, $\operatorname{rank}(\hat{Z}) < n-1$. Now $\hat{X}+\hat{Z}$ is singular and both $\hat{X}$ and $\hat{Z}$ are positive semidefinite, so there must be a nontrivial y in the intersection of the null spaces of $\hat{X}$ and $\hat{Z}$. Let

$$X := \hat{X} + yy^T = \hat{x}_1\hat{x}_1^T + \hat{x}_2\hat{x}_2^T + yy^T.$$

Obviously, $\operatorname{rank}(X) = 3$ and $\hat{Z}X = 0$. Since

$$M(q_1) \bullet \hat{x}_1\hat{x}_1^T = M(q_1) \bullet \hat{x}_1\hat{x}_1^T = 0, \tag{4.1}$$

$$(M(q_2) \bullet \hat{x}_1\hat{x}_1^T)(M(q_2) \bullet \hat{x}_1\hat{x}_1^T) < 0. \tag{4.2}$$

By applying Lemma 3.3, we obtain x such that X is rank-one decomposable at x and that

$$M(q_1) \bullet xx^T = 0, \quad M(q_2) \bullet xx^T = 0.$$

Since x is in the range space of X, it must be in the null space of $\hat{Z}$. That is, $\hat{Z} \bullet xx^T = 0$, implying that xx^T/t^2 is an optimal solution to $(SP)_2$ and x/t is an optimal solution to $(QP)_2$, where t is the first component of x (which must be nonzero as argued before).

This concludes *Part* 1.

Next we proceed to *Part* 2, in which we shall prove that if Property $\mathcal{I}$ holds, then there is definitely a gap between $(QP)_2$ and $(SP)_2$, i.e., $v((SP)_2) < v((QP)_2)$. To see why this is true, we use a contradiction argument. Suppose that Property $\mathcal{I}$ holds, while $v((SP)_2) = v((QP)_2)$. Let x^* be an optimal solution of $(QP)_2$ (we extend the dimension of x^* to be $(n+1)$-dimensional by putting 1 in the first component). Then, since $v((SP)_2) = v((QP)_2)$, $x^*(x^*)^T$ must be an optimal solution to $(SP)_2$. Consequently, $x^*(x^*)^T$ and $(\hat{Z}, \hat{y}_0, \hat{y}_1, \hat{y}_2)$ must satisfy the complementarity condition; i.e.,

$$\hat{Z}x^*(x^*)^T = 0, \quad M(q_1) \bullet x^*(x^*)^T = 0, \quad M(q_2) \bullet x^*(x^*)^T = 0. \tag{4.3}$$

This implies that x^* must be in the null space of $\hat{Z}$, which is two-dimensional in this case. In other words, it must be a linear combination of $\hat{x}_1$ and $\hat{x}_2$. Let us assume that there are two numbers α and β such that

$$x^* = \alpha\hat{x}_1 + \beta\hat{x}_2. \tag{4.4}$$

Substituting (4.4) into the equations $M(q_1) \bullet x^*(x^*)^T = 0$ and $M(q_2) \bullet x^*(x^*)^T = 0$ and noting (4.1) and (4.2), we obtain

$$\alpha\beta\hat{x}_1^T M(q_1)\hat{x}_2 = 0, \tag{4.5}$$

$$\alpha^2 M(q_2) \bullet \hat{x}_1\hat{x}_1^T + 2\alpha\beta M(q_2) \bullet \hat{x}_1\hat{x}_2^T + \beta^2 M(q_2) \bullet \hat{x}_2\hat{x}_2^T = 0. \tag{4.6}$$

Due to (4.2), neither α nor β can be zero. (For example, if $\alpha = 0$, then by (4.6) and (4.2), it necessarily follows that $\beta = 0$ and vice versa). Thus, from (4.5), it follows that $\hat{x}_1^T M(q_1)\hat{x}_2 = 0$. Let $\hat{x}_1 = (t_1, u^T)^T$ and $\hat{x}_2 = (t_2, v^T)^T$, where $u, v \in \Re^n$. We have $0 = M(q_1) \bullet \hat{x}_1\hat{x}_1^T = t_1^2 - \|u\|^2$ and $0 = M(q_1) \bullet \hat{x}_2\hat{x}_2^T = t_2^2 - \|v\|^2$. Now, $\hat{x}_1^T M(q_1)\hat{x}_2 = 0$ leads to $0 = t_1t_2 - u^Tv$, and so $(u^Tv)^2 = \|u\|^2\|v\|^2$. By the Cauchy–Schwartz inequality, this is only possible when u is a multiple of v. Consequently, $\hat{x}_1$ and $\hat{x}_2$ must be linearly dependent, a contradiction to the fact that $2 = \operatorname{rank}(\hat{X}) = \hat{x}_1\hat{x}_1^T + \hat{x}_2\hat{x}_2^T$. ∎

5. The Lagrangian function and the KKT condition. It is intuitively clear that the Lagrangian function must be related to the SDP relaxation, as Theorem 2.1 has already indicated, primarily due to the fact that the Lagrangian dual of quadratically constrained quadratic program (QP) is identical to the dual of its SDP relaxation. It is, however, useful to translate Property $\mathcal{I}$ using the terms of the Lagrangian function and the KKT conditions explicitly due to its relevance in the nonlinear programming community.

First, let us formally introduce an analog of Property $\mathcal{I}$ in the context of Lagrangian multipliers.

DEFINITION 5.1. *For given Lagrangian multipliers λ and μ for the quadratic program $(QP)_2$, we say that they have Property $\mathcal{I}'$ if*

(1) $\lambda > 0$ *and* $\mu > 0$*;*
(2) $H(\lambda,\mu) = Q_0 + \lambda I + \mu Q_2 \succeq 0$ *and* $\operatorname{rank}(H(\lambda,\mu)) = n-1$*;*
(3) *The system of linear equations $H(\lambda,\mu)x = b_0 + \mu b_2$ has two solutions x_1 and x_2 satisfying $x_i^T x_i = 1$, $i = 1, 2$, and $q_2(x_1)q_2(x_2) < 0$.*

THEOREM 5.2. *Suppose that $(QP)_2$ satisfies the Slater condition. Then, $(QP)_2$ has no strong duality if and only if there exist multipliers λ and μ such that Property $\mathcal{I}'$ holds.*

Proof. The Slater condition for $(QP)_2$ implies $v((SP)_2) = v((SD)_2) = v((QD)_2)$, where $(QD)_2$ denotes the dual problem of $(QP)_2$. Therefore, "$(QP)_2$ has no strong duality" is equivalent to "$v((SP)_2) < v((QP)_2)$." By Theorem 4.2, it is again equivalent to "Property $\mathcal{I}$ holds." What remains to show is that Property $\mathcal{I}$ holds if and only if the above Property $\mathcal{I}'$ holds. To put things in perspective, we restate Property $\mathcal{I}$ as follows: There exist three numbers y_0, λ, μ and two linearly independent $(n+1)$-dimensional vectors $\hat{x}_1 = [t_1, x_1^T]^T$ and $\hat{x}_2 = [t_2, x_2^T]^T$ such that

$$
\begin{cases}
x_1^T x_1 - t_1^2 = x_2^T x_2 - t_2^2 = 0, \\
M(q_2) \bullet \hat{x}_1\hat{x}_1^T + M(q_2) \bullet \hat{x}_2\hat{x}_2^T = 0, \\
(M(q_2) \bullet \hat{x}_1\hat{x}_1^T)(M(q_2) \bullet \hat{x}_2\hat{x}_2^T) < 0, \\
t_1^2 + t_2^2 = 1, \\
(\lambda,\mu) > 0, \\
Z := M(q_0) - y_0 I_{00} + \lambda M(q_1) + \mu M(q_2) \succeq 0, \\
\operatorname{rank}(Z) = n-1, \\
Z\hat{x}_1 = Z\hat{x}_2 = 0.
\end{cases} \tag{5.1}
$$

"Property $\mathcal{I} \Longrightarrow$ Property $\mathcal{I}'$".

First we note that from the sixth equation in (5.1), we may write Z as

$$
Z = \begin{bmatrix} -y_0 + \lambda + \mu c_2 & -b_0^T - \mu b_2^T \\ -b_0 - \mu b_2 & Q_0 + \lambda I + \mu Q_2 \end{bmatrix}. \tag{5.2}
$$

By $x_1^T x_1 - t_1^2 = x_2^T x_2 - t_2^2 = 0$ and the linear independence of $\hat{x}_1$ and $\hat{x}_2$, we have $t_1 t_2 \neq 0$. Let

$$
\bar{x}_1 := x_1/t_1,\ \bar{x}_2 := x_2/t_2.
$$

By (5.1), it immediately follows that $\bar{x}_1$ and $\bar{x}_2$ must satisfy the following:

$$\begin{array}{l}
\|\bar{x}_1\| = \|\bar{x}_2\| = 1, \\
q_2(\bar{x}_1)q_2(\bar{x}_2) < 0, \\
(\lambda, \mu) > 0, \\
Q_0 + \lambda I + \mu Q_2 \succeq 0, \\
(Q_0 + \lambda I + \mu Q_2)\bar{x}_i = b_0 + \mu b_2,\ i = 1, 2, \\
\bar{x}_1, \bar{x}_2 \text{ are linearly indepedent.}
\end{array}$$

It now remains only to check if $\mathrm{rank}(Q_0 + \lambda I + \mu Q_2) = n - 1$. By using $Z\hat{x}_1$=0 and (5.2), we have

$$\left[\begin{array}{c} -y_0 + \lambda + \mu c_2 \\ -b_0 - \mu b_2 \end{array} \right] = - \left[\begin{array}{c} -b_0^T - \mu b_2^T \\ Q_0 + \lambda I + \mu Q_2 \end{array} \right] \bar{x}_1,$$

which implies that

$$\begin{aligned}
n - 1 = \mathrm{rank}(Z) &= \mathrm{rank} \left(\left[\begin{array}{cc} -y_0 + \lambda + \mu c_2 & -b_0^T - \mu b_2^T \\ -b_0 - \mu b_2 & Q_0 + \lambda I + \mu Q_2 \end{array} \right] \right) \\
&= \mathrm{rank} \left(\left[\begin{array}{c} -b_0^T - \mu b_2^T \\ Q_0 + \lambda I + \mu Q_2 \end{array} \right] \right) = \mathrm{rank}(Q_0 + \lambda I + \mu Q_2).
\end{aligned}$$

"Property $\mathcal{I}'$ $\Longrightarrow$ Property $\mathcal{I}$".

Let us assume, without loss of generality, that $q_2(x_1) < 0, q_2(x_2) > 0$, and let us define

$$\begin{array}{l}
y_0 := q_0(x_1) + \lambda q_1(x_1) + \mu q_2(x_1), \\
t_1 := \sqrt{\dfrac{-q_2(x_2)}{q_2(x_1) - q_2(x_2)}}, \\
t_2 := \sqrt{\dfrac{q_2(x_1)}{q_2(x_1) - q_2(x_2)}}, \\
\hat{x}_1 := t_1 \left[\begin{array}{c} 1 \\ x_1 \end{array} \right], \\
\hat{x}_2 := t_2 \left[\begin{array}{c} 1 \\ x_2 \end{array} \right], \\
Z := M(q_0) - y_0 I_{00} + \lambda M(q_1) + \mu M(q_2).
\end{array}$$

Then, it can be straightforwardly checked that

$$\begin{array}{l}
M(q_1) \bullet \hat{x}_1\hat{x}_1^T = M(q_1) \bullet \hat{x}_2\hat{x}_2^T = 0, \\
M(q_2) \bullet \hat{x}_1\hat{x}_1^T + M(q_2) \bullet \hat{x}_2\hat{x}_2^T = 0, \\
M(q_2) \bullet \hat{x}_1\hat{x}_1^T < 0,\ M(q_2) \bullet \hat{x}_2\hat{x}_2^T > 0, \\
t_1^2 + t_2^2 = 1, \\
(\lambda, \mu) > 0.
\end{array}$$

To complete the proof, one needs only to show that $Z \succeq 0$, $Z\hat{x}_1 = Z\hat{x}_2 = 0$, and $\mathrm{rank}(Z) = n - 1$.

Consider the Lagrangian function

$$L(x;\lambda,\mu) := q_0(x) + \lambda q_1(x) + \mu q_2(x),$$

whose Hessian matrix $H(\lambda,\mu) = Q_0 + \lambda I + \mu Q_2$ is semidefinite, due to (2) of Property $\mathcal{I}'$. This implies that $L(x;\lambda,\mu)$ is a convex quadratic function in x. Furthermore, (2) and (3) of Property $\mathcal{I}'$ imply that the minimizers of $L(x;\lambda,\mu)$ consist of all the points on the straight line connecting x_1 and x_2. Consequently,

$$y_0 = L(x_1;\lambda,\mu) = L(x_2;\lambda,\mu) = \min_{x\in\Re^n} L(x;\lambda,\mu).$$

Consider any $(n+1)$-dimensional vector $(t, x^T)^T$, where $t \in \Re^1$ and $x \in \Re^n$. If $t = 0$, then

$$\begin{aligned}\begin{bmatrix} t \\ x \end{bmatrix}^T Z \begin{bmatrix} t \\ x \end{bmatrix} &= \begin{bmatrix} 0 \\ x \end{bmatrix}^T \begin{bmatrix} -y_0 + \lambda + \mu c_2 & -b_0^T - \mu b_2^T \\ -b_0 - \mu b_2 & Q_0 + \lambda I + \mu Q_2 \end{bmatrix} \begin{bmatrix} 0 \\ x \end{bmatrix} \\ &= x^T(Q_0 + \lambda I + \mu Q_2)x \geq 0.\end{aligned}$$

Otherwise, if $t \neq 0$, then

$$\begin{aligned}\begin{bmatrix} t \\ x \end{bmatrix}^T Z \begin{bmatrix} t \\ x \end{bmatrix} &= \begin{bmatrix} t \\ x \end{bmatrix}^T (M(q_0) - y_0 I_{00} + \lambda M(q_1) + \mu M(q_2)) \begin{bmatrix} t \\ x \end{bmatrix} \\ &= t^2 q_0(x/t) - t^2 y_0 + \lambda t^2 q_1(x/t) + \mu t^2 q_2(x/t) = t^2(L(x/t;\lambda,\mu) - y_0) \\ &\geq t^2(L(x_1;\lambda,\mu) - L(x_1;\lambda,\mu)) = 0.\end{aligned}$$

Moreover, $\hat{x}_1^T Z \hat{x}_1 = t_1^2(L(x_1;\lambda,\mu) - y_0) = 0$ and $\hat{x}_2^T Z \hat{x}_2 = t_2^2(L(x_2;\lambda,\mu) - y_0) = t_2^2(L(x_2;\lambda,\mu) - L(x_1;\lambda,\mu)) = 0$. Therefore, $Z\hat{x}_1 = 0$ and $Z\hat{x}_2 = 0$ because $Z \succeq 0$. Since $\hat{x}_1$ and $\hat{x}_2$ are linearly independent, it follows that $\operatorname{rank}(Z) \leq n-1$. On the other hand, $\operatorname{rank}(Z) \geq \operatorname{rank}(H(\lambda,\mu)) = n-1$, leading to $\operatorname{rank}(Z) = n-1$. □

Property $\mathcal{I}'$ is closely related to Property $\mathcal{J}$ studied in Chen and Yuan [5] for the CDT subproblem. Since Chen and Yuan [5] considered the CDT subproblem, they considered problem $(QP)_2$ with an additional condition that $Q_2 \succeq 0$. To put things in perspective, their Property $\mathcal{J}$ can be stated as follows.

DEFINITION 5.3. *For given Lagrangian multipliers λ and μ for the quadratic program $(QP)_2$, we say that they have Property $\mathcal{J}$ if*

(1) *$\lambda > 0$ and $\mu > 0$;*
(2) *$H(\lambda,\mu) = Q_0 + \lambda I + \mu Q_2 \succeq 0$ and $\operatorname{rank}(H(\lambda,\mu)) = n-1$;*
(3) *The following "surrogate" problem*

$$(P)_{\frac{\lambda}{\lambda+\mu}} \quad \begin{aligned} &\textit{minimize} && q_0(x) \\ &\textit{subject to} && \frac{\lambda}{\lambda+\mu} q_1(x) + \frac{\mu}{\lambda+\mu} q_2(x) \leq 0 \end{aligned}$$

has two solutions x_1 and x_2 satisfying

$$H(\lambda,\mu)x = b_0 + \mu b_2,$$

and $x_1^T x_1 < 1$ and $x_2^T x_2 > 1$.

The above Property $\mathcal{J}$ (see [5]) is based on the idea of surrogate representation of the constraints, hence different from ours. Moreover, Chen and Yuan in (see [5]) proved just only that if $(QP)_2$ with $Q_2 \succeq 0$ has no strong duality, then Property $\mathcal{J}$

holds, in other words, the converse proposition cannot be proved so far. However, the appearances of Property $\mathcal{J}$ and Property $\mathcal{I}'$ are quite similar indeed. Despite the similar appearances, below we shall show that they are not identical in all circumstances. Before our discussion, we shall first remark that the existence of multipliers satisfying Property $\mathcal{J}$ cannot be directly verified, while Property $\mathcal{I}$ can be checked in polynomial-time by solving a pair of SDP problems.

PROPOSITION 5.4. *If $Q_2 \succeq 0$, then Property $\mathcal{J}$ is equivalent to Property $\mathcal{I}'$. If $Q_2 \not\succeq 0$, then Property $\mathcal{J}$ is not identical to Property $\mathcal{I}'$, the latter being a necessary and sufficient condition for $(QP)_2$ to admit a gap with its SDP relaxation.*

Proof. First consider the situation when $Q_2 \succeq 0$. We shall prove in this case that Property $\mathcal{J}$ leads to Property $\mathcal{I}'$.

Restricting the quadratic function $\frac{\lambda}{\lambda+\mu}q_1(x) + \frac{\mu}{\lambda+\mu}q_2(x)$ on the line connecting x_1 and x_2, we obtain a univariate function

$$g(t) := \frac{\lambda}{\lambda+\mu}q_1((1-t)x_1 + tx_2) + \frac{\mu}{\lambda+\mu}q_2((1-t)x_1 + tx_2), \quad t \in \Re.$$

Since $\frac{\lambda}{\lambda+\mu}q_1(x) + \frac{\mu}{\lambda+\mu}q_2(x)$ is strictly convex and quadratic, we have

$$g(t) \begin{cases} = 0, & \text{if } t = 0 \text{ and } 1; \\ < 0, & \text{if } 0 < t < 1; \\ > 0, & \text{else.} \end{cases} \tag{5.3}$$

Similarly, $h(t) := q_1((1-t)x_1 + tx_2) = \|(1-t)x_1 + tx_2\|^2$ is also a strictly convex quadratic function of t. Therefore, $h(0) < 0$ and $h(1) > 0$ lead to the existence of two numbers $t_1 \in (-\infty, 0)$ and $t_2 \in (0, 1)$ such that $h(t_1) = h(t_2) = 0$. Denote $x_3 = (1-t_1)x_1 + t_1x_2$ and $x_4 = (1-t_2)x_1 + t_2x_2$. Based on (5.3), we have

$$q_1(x_3) = q_1(x_4) = 0, q_2(x_3) > 0, q_2(x_4) < 0,$$

which means that Property $\mathcal{I}'$ holds.

Now consider the case where $Q_2 \not\succeq 0$. We shall prove our assertion by the following example:

$$\begin{aligned} \text{minimize} \quad & q_0(x) = x_1^2 - 3x_1 \\ \text{subject to} \quad & q_1(x) = x_1^2 + x_2^2 - 1 \le 0, \\ & q_2(x) = -x_1^2 - x_2^2 + 2x_1 \le 0, \end{aligned}$$

where $x = (x_1, x_2)^T$. It is easy to see that the two circles $q_1(x) = 0$ and $q_2(x) = 0$ intersect at two points: P_1 with coordinates $(\frac{1}{2}, \frac{\sqrt{3}}{2})$ and P_2 with coordinates $(\frac{1}{2}, -\frac{\sqrt{3}}{2})$. It is easy to see that P_1 and P_2 are two unique optimal solutions for this problem, for which the corresponding multipliers are $\lambda = \mu = 1$, with the corresponding Hessian matrix of the Lagrangian function being

$$H(\lambda, \mu) = Q_0 + \lambda Q_1 + \mu Q_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

which is positive semidefinite with rank $n-1$ ($n = 2$). So, this problem has optimal solutions with positive semidefinite Lagrangian Hessian matrices. The KKT points satisfy

$$\begin{cases} 1x_1 = \frac{1}{2}, \\ 0x_2 = 0, \end{cases}$$

which lie on the line connecting P_1 and P_2. In this case, however, by Theorem 5.2 we know that Property $\mathcal{I}'$ is violated. We shall see below that Property $\mathcal{J}$ still holds nevertheless. Choose, for instance, $x^{(1)} = (\frac{1}{2}, 0)^T$, $x^{(2)} = (\frac{1}{2}, 1)^T$, and $\lambda = \mu = 1$. We have $\frac{\lambda}{\lambda+\mu} q_1(x) + \frac{\mu}{\lambda+\mu} q_2(x) = 0$ for $x = x^{(1)}$ and $x = x^{(2)}$, and $\|x^{(1)}\| < 1$ and $\|x^{(2)}\| > 1$. After checking the conditions, we see that Property $\mathcal{J}$ is indeed satisfied in this case; however, Property $\mathcal{I}'$ is violated as we have observed. ∎

Another related result is due to Beck and Eldar [1]. Their approach is based on a comparison between the real and the complex valued SDP relaxations. They showed that if the dimension of the null space of $H(\lambda, \mu)$ is not equal to 1, or equivalently, $\operatorname{rank}(H(\lambda, \mu)) \neq n-1$, then the SDP relaxation is tight. In the context of Theorem 5.2, this is clear, since this sufficient condition guarantees that Property $\mathcal{I}'$ does not hold, and hence the SDP relaxation must be tight.

Since, in Property $\mathcal{I}'$ of Theorem 5.2, the constraint $q_2(x) \leq 0$ plays a role only in the last part, the following corollary is immediate.

Consider

$$\begin{array}{lll} (Q(\rho))_2 & \text{minimize} & q_0(x) = x^T Q_0 x - 2b_0^T x \\ & \text{subject to} & q_1(x) = x^T x - 1 \leq 0 \\ & & q_2(x) - \rho \leq 0, \end{array}$$

where ρ is a parameter.

COROLLARY 5.5. *Suppose that Property $\mathcal{I}'$ holds for $(QP)_2$ with x_1 and x_2 being the two solutions in* (3) *of Property $\mathcal{I}'$ satisfying $q_2(x_1) < 0 < q_2(x_2)$. Then for any $\rho \in (q_2(x_1), q_2(x_2))$, problem $(Q(\rho))_2$ will not have a positive semidefinite Hessian for its Lagrangian function at any optimal solution.*

6. The optimal line of the dual problem. As shown in the previous sections, if Property $\mathcal{I}'$ holds for a CDT subproblem, then there exists a gap between the optimal values of the primal and the dual problems. In the case of Property $\mathcal{I}'$, we obtain two dual optimal solutions x_1 and x_2, one of which is feasible for the primal problem, say, x_1. It can be easily proved that each point of the entire line connecting x_1 and x_2 is also an optimal solution to the dual problem. Let us call this line *the optimal line of the dual problem*. Naturally, we may wish to minimize the original quadratic function along this line to obtain a better approximate solution than x_1 for the primal problem. It is tempting to conjecture that this will always lead to an improvement. However, below we shall give an example to show that this approach may not yield a solution with any quality assurance.

Example 6.1.

$$\begin{array}{ll} \text{minimize} & q_0(x_1, x_2) = x_1(p - x_1) \\ \text{subject to} & q_1(x_1, x_2) = x_1^2 + x_2^2 \leq \dfrac{17}{16} p^2, \\ & q_2(x_1, x_2) = (x_1 - 2p)^2 + (x_2 - p)^2 \leq \dfrac{73}{16} p^2, \end{array}$$

where p is a positive parameter. The global optimal solution for this problem is $x^* \approx \begin{bmatrix} -0.1359p \\ 1.0218p \end{bmatrix}$, which is one of two intersection points of the circles $q_1(x_1, x_2) = 0$ and $q_2(x_1, x_2) = 0$, and the corresponding optimal value is $v^* \approx -0.1544p^2$. The system $(Q_0 + \lambda Q_1 + \mu Q_2)x = b_0 + \mu b_2$ is in this case

$$\begin{aligned} (-2 + 2\lambda + 2\mu)x_1 &= (4\mu - 1)p \\ (2\lambda + 2\mu)x_2 &= 2\mu p. \end{aligned}$$

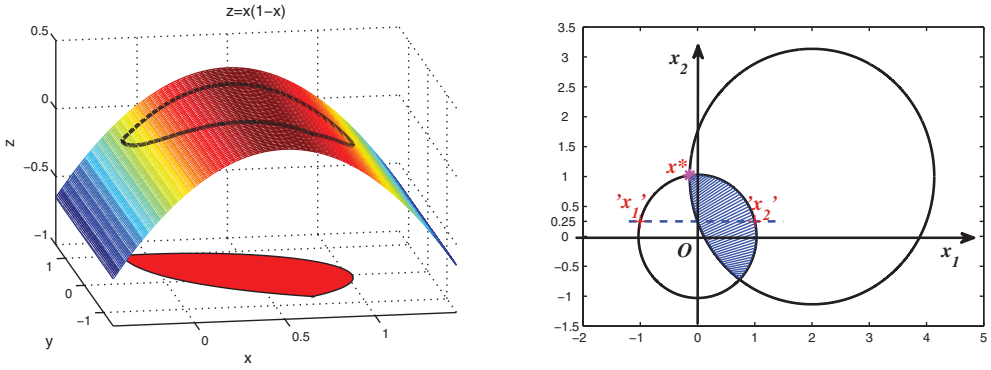


FIG. 6.1. *The graph of* $z = x_1(1 - x_1)$ *(on the left) and the feasible domain (on the right) at* $p = 1$.

One easily verifies that Property $\mathcal{I}'$ holds at $(\lambda, \mu) = (0.75, 0.25)$, and the solutions "$x_1$" and "$x_2$" in (3) of Property $\mathcal{I}'$ are $\begin{bmatrix} p \\ 0.25p \end{bmatrix}$ and $\begin{bmatrix} -p \\ 0.25p \end{bmatrix}$, respectively (see Figure 6.1). The optimal value of the SDP relaxation $(SP)_2$ is $y_0 = -0.75p^2$, and the gap between y_0 and v^* is $v^* - y_0 \approx 0.5926p^2$. The line segment that connects "x_1" and "x_2" and is contained in the feasible domain can be expressed by

$$\left\{ \begin{bmatrix} tp \\ 0.25p \end{bmatrix} \,\middle|\, 0 \le t \le 1 \right\}.$$

On this line segment, the optimal value of $q_0(x_1, x_2)$ is identically 0 for any p, which can be attained at the point "x_2." This shows that there cannot be any bound, in neither absolute nor relative sense of error measurements, regarding the quality of the solution obtained by the heuristic method of searching along the line segment. It remains to be a challenge to solve $(QP)_2$ efficiently, if, after solving its SDP relaxation, it turns out that Property $\mathcal{I}$ indeed holds, although numerical experiments in [1] suggest that this is highly unlikely for randomly generated instances.

7. Testing Property $\mathcal{I}$ numerically. In its direct form, Property $\mathcal{I}$ requires the knowledge of an exact solution for the SDP relaxation. As is well known, in general, it is impossible to solve an SDP problem exactly. It is therefore natural to test its predictive power if one uses the necessary and sufficient condition involving Property $\mathcal{I}$ in an approximative sense. In other words, if we use an ε_1-approximation solution of the SDP relaxation, then a similarly relaxed Property $\mathcal{I}$ can be verified, leading to the conclusion whether or not the original CDT subproblem satisfies the strong duality within an ε_2 error tolerance. The question is, How does the approximation work in practice?

First, we need to relax the requirement on the optimal solution. Applying an SDP solver (such as SeDuMi) to solve the SDP relaxation will return with a solution $\bar{X} \succeq 0$ and a dual solution $(\bar{Z}, \bar{y}_0, \bar{y}_1, \bar{y}_2)$, with $\bar{Z} \succeq 0$. Of course, these solutions might, however, violate the equality and inequality constraints of the primal-dual feasibility requirements, say, by an amount no more than ε_1. Then, to purify the ranks of $\bar{X}$ and $\bar{Z}$, we may operate a spectral decomposition on $\bar{X}$ and $\bar{Z}$: $\bar{X} = Q_1^T \Lambda_1 Q_1$ and $\bar{Z} = Q_2^T \Lambda_2 Q_2$, where Q_i is orthonormal and $\Lambda_i = \mathrm{diag}(\lambda_{i1}, \ldots, \lambda_{in})$, with $\lambda_{ij} \ge 0$,

$j = 1, \ldots, n$, $i = 1, 2$. Introduce

$$\hat{\lambda}_{ij} := \begin{cases} \lambda_{ij}, & \text{if } \lambda_{ij} \geq \varepsilon_2, \\ 0, & \text{if } \lambda_{ij} < \varepsilon_2, \end{cases}$$

for $j = 1, \ldots, n$, $i = 1, 2$, and let us purify the solutions by using

$$\hat{X} := Q_1^T \mathrm{diag}(\hat{\lambda}_{11}, \ldots, \hat{\lambda}_{1n}) Q_1 \text{ and } \hat{Z} := Q_2^T \mathrm{diag}(\hat{\lambda}_{21}, \ldots, \hat{\lambda}_{2n}) Q_2$$

instead of $\bar{X}$ and $\bar{Z}$, while keeping $\hat{y}_i := \bar{y}_i$, $i = 0, 1, 2$. We call $\hat{X}$ and $(\hat{Z}, \hat{y}_0, \hat{y}_1, \hat{y}_2)$ to be *a pair of purified* $(\varepsilon_1, \varepsilon_2)$*-approximate optimal solutions.*

DEFINITION 7.1. *Suppose that* $\hat{X}$ *and* $(\hat{Z}, \hat{y}_0, \hat{y}_1, \hat{y}_2)$ *are a pair of purified* $(\varepsilon_1, \varepsilon_2)$*-approximate optimal solutions for* $(SP)_2$ *and* $(SD)_2$*, respectively. We say this pair has Property* $\mathcal{I}(\varepsilon_2)$ *if*

(1) $\hat{y}_1 > \varepsilon_2$ *and* $\hat{y}_2 > \varepsilon_2$*;*
(2) $\mathrm{rank}(\hat{Z}) = n - 1$*;*
(3) $\mathrm{rank}(\hat{X}) = 2$*, and there is a rank-one decomposition of* $\hat{X}$*,* $\hat{X} = \hat{x}_1 \hat{x}_1^T + \hat{x}_2 \hat{x}_2^T$*, such that* $M(q_1) \bullet \hat{x}_i \hat{x}_i^T = M(q_1) \bullet \hat{X}/2$*,* $i = 1, 2$*, and* $M(q_2) \bullet \hat{x}_1 \hat{x}_1^T < -\varepsilon_2$ *and* $M(q_2) \bullet \hat{x}_2 \hat{x}_2^T > \varepsilon_2$*.*

Below we shall introduce a polynomial-time procedure to test the strong duality for the CDT problem, based on the ε_1-optimal SDP relaxation solution Property $\mathcal{I}(\varepsilon_2)$ and the matrix decomposition technique.

ALGORITHM 7.2. *Input* ε_2*,* $M(q_0)$*,* $M(q_1)$*, and* $M(q_2)$*.*

Step 1. *Let* $\hat{X}$ *and* $(\hat{Z}, \hat{y}_0, \hat{y}_1, \hat{y}_2)$ *be the purified* $(\varepsilon_1, \varepsilon_2)$*-approximate solutions for* $(SP)_2$ *and its dual.*

Step 2. *Test whether or not Property* $\mathcal{I}(\varepsilon_2)$ *is satisfied by checking Definition* 7.1*, which runs in polynomial-time.*

Step 3. *If Property* $\mathcal{I}(\varepsilon_2)$ *is violated, then use the matrix decomposition technique presented in the previous sections to obtain an approximate solution to the original CDT problem; otherwise, get an approximate solution by searching along the optimal line of the dual problem.*

We now use SeDuMi to test this procedure by numerical simulations. Throughout our tests, we let $\varepsilon_2 = 10^{-4}$ and ε_1 be set as the default precision of SeDuMi. For a given positive integer n, our MATLAB code would generate two $(n+1) \times (n+1)$ matrices $M(q_0)$ and $M(q_2)$, of which the upper triangular part (including diagonal) of the entries are uniformly generated random numbers on the interval $[-50, 50]$ (the lower part takes the values by symmetry). In order to guarantee that $(SP)_2$ have an interior feasible solution, we first solve

$$\begin{array}{ll} \text{minimize} & M(q_2) \bullet X \\ \text{subject to} & M(q_1) \bullet X \leq 0, \\ & I_{00} \bullet X = 1, \\ & X \succeq 0. \end{array}$$

Let f^* denote its optimal value. If $f^* > -10^{-4}$, we decrease the first entry (the (1,1)th position) of $M(q_2)$ by the amount $f^* + 10^{-4}$. This ensures that the Slater condition is satisfied. We apply Algorithm 7.2 on 90 randomly generated instances. The numerical results are summarized in Tables 1, 2, and 3, where "n" denotes the dimension of the CDT problem, "Value 1" is equal to $M(q_0) \bullet \bar{X}$, i.e., the ε_1-optimal value of the SDP relaxation solution returned by SeDuMi, "Value 2" denotes the objective value of the feasible solution for the CDT problem generated by Algorithm 7.2, and "Gap"

TABLE 1
Numerical results.

n	Value 1	Value 2	Gap	Rank	$I(\varepsilon_2)$
1	31.8310	31.8310	$-1.9315e-008$	1	V
2	−61.6350	−61.6350	$1.0267e-007$	1	V
3	−92.6195	−92.6195	$1.5046e-008$	1	V
4	−64.3479	−64.3479	$6.3392e-009$	1	V
5	−76.0429	−76.0429	$1.2039e-007$	1	V
6	−148.3942	−148.3942	$8.4647e-008$	1	V
7	−149.2147	−149.2147	$1.3788e-007$	1	V
8	−165.2366	−165.2366	$2.2856e-007$	1	V
9	−146.7020	−146.7020	$6.5012e-010$	1	V
10	−193.3607	−193.3607	$1.0247e-007$	1	V
11	−194.9409	−194.9409	$4.3410e-006$	1	V
12	−131.2606	−131.2606	$3.4186e-009$	1	V
13	−174.0891	−174.0891	$4.6756e-008$	1	V
14	−215.5152	−215.5152	$2.8498e-008$	1	V
15	−232.2548	−232.2548	$1.1953e-007$	1	V
16	−288.2241	−288.2241	$8.9968e-008$	1	V
17	−180.2632	−180.2632	$1.5133e-007$	1	V
18	−257.0321	−257.0321	$1.1875e-007$	1	V
19	−307.8921	−307.8921	$7.1101e-008$	1	V
20	−250.2240	−250.2240	$2.4064e-008$	1	V
21	−216.6837	−216.6837	$1.5005e-007$	1	V
22	−285.2257	−285.2257	$1.1723e-006$	1	V
23	−305.7068	−305.7068	$1.1012e-007$	1	V
24	−273.7716	−273.7716	$2.2697e-008$	1	V
25	−305.1200	−305.1200	$2.6449e-010$	1	V
26	−311.0972	−311.0972	$9.3392e-008$	1	V
27	−269.2598	−269.2598	$1.5854e-008$	1	V
28	−349.2378	−349.2378	$1.3295e-009$	1	V
29	−280.3103	−280.3103	$7.1443e-007$	1	V
30	−322.0861	−322.0861	$1.9794e-008$	1	V

TABLE 2
Numerical results for $n = 5$.

Instance	Value 1	Value 2	Gap	Rank	$I(\varepsilon_2)$
1	−72.2487	−72.2487	$-9.0962e-010$	1	V
2	−78.8733	−78.8733	$3.1875e-007$	1	V
3	−129.3945	−129.3945	$2.4719e-009$	1	V
4	−78.6061	−78.6061	$3.1858e-007$	1	V
5	−87.7781	−87.7781	$4.0048e-009$	1	V
6	−162.4757	−162.4757	$3.2261e-009$	1	V
7	−181.4192	−181.4192	$1.2105e-006$	1	V
8	−148.9920	−131.6450	17.3470	2	H
9	−84.6160	−84.6160	$1.2004e-007$	1	V
10	−106.1400	−106.1400	$2.6063e-007$	1	V
11	−80.2952	−80.2952	$8.1327e-010$	1	V
12	−93.9455	−37.5482	56.3973	2	H
13	−182.7852	−182.7852	$7.6042e-008$	1	V
14	−47.4945	−47.4945	$3.5781e-008$	1	V
15	−107.2132	−107.2132	$7.4877e-008$	1	V
16	−195.9235	−195.9235	$4.2069e-008$	1	V
17	−91.5627	−91.5627	$1.7774e-009$	1	V
18	−149.4562	−149.4562	$2.0514e-007$	1	V
19	−199.7809	−199.7809	$4.9602e-010$	1	V
20	−96.7141	−96.7141	$2.0592e-006$	1	V
21	−193.2582	−193.2582	$1.0298e-006$	1	V
22	−121.9034	−121.9034	$1.3054e-009$	1	V
23	−132.7388	−132.7388	$5.9610e-008$	1	V
24	−221.9654	−221.9654	$-7.9771e-010$	1	V
25	−69.0899	−69.0899	$9.3646e-006$	1	V
26	−48.9339	−38.6602	10.2737	2	H
27	−204.1014	−204.1014	$1.4712e-007$	1	V
28	−50.4021	−50.4021	$1.8484e-006$	1	V
29	−95.7052	−95.7052	$4.4413e-008$	1	V
30	−162.5680	−162.5680	$6.1157e-009$	1	V

TABLE 3
Numerical results for $n = 50$.

Instance	Value 1	Value 2	Gap	Rank	$I(\varepsilon_2)$
1	−329.1350	−329.1350	$1.3564e-008$	1	V
2	−418.0411	−418.0411	$1.3500e-010$	1	V
3	−334.9108	−334.9108	$8.4879e-010$	1	V
4	−314.3538	−314.3538	$4.0116e-007$	1	V
5	−406.6970	−406.6970	$1.8738e-008$	1	V
6	−376.4849	−376.4849	$6.9003e-009$	1	V
7	−436.8686	−436.8686	$1.1316e-009$	1	V
8	−456.1419	−456.1419	$1.0745e-009$	1	V
9	−420.0406	−420.0406	$2.3637e-009$	1	V
10	−443.0921	−443.0921	$2.8577e-009$	1	V
11	−398.1299	−398.1299	$1.2683e-008$	1	V
12	−381.3000	−381.3000	$2.2239e-009$	1	V
13	−400.2680	−400.2680	$1.5546e-007$	1	V
14	−337.3982	−337.3982	$4.3128e-008$	1	V
15	−433.0168	−433.0168	$1.5800e-007$	1	V
16	−353.2036	−353.2036	$1.1395e-007$	1	V
17	−422.6912	−422.6912	$1.7024e-007$	1	V
18	−373.7865	−373.7865	$5.1733e-010$	1	V
19	−356.4418	−356.4418	$4.0084e-007$	1	V
20	−449.4164	−449.4164	$1.8588e-010$	1	V
21	−363.3087	−363.3087	$6.4648e-010$	1	V
22	−422.4459	−422.4459	$3.0531e-009$	1	V
23	−376.0524	−376.0524	$1.4611e-007$	1	V
24	−399.0962	−399.0962	$3.5397e-007$	1	V
25	−428.4575	−428.4575	$1.6672e-010$	1	V
26	−422.2624	−422.2624	$2.8901e-009$	1	V
27	−422.8571	−422.8571	$5.3685e-009$	1	V
28	−344.5267	−344.5267	$7.2918e-007$	1	V
29	−448.3855	−448.3855	$1.4571e-008$	1	V
30	−403.9283	−403.9283	$4.7542e-009$	1	V

indicates the difference between "Value 1" and "Value 2" (Gap = Value 2 − Value 1), which reflects the eventual performance of Algorithm 7.2. Finally, "Rank" indicates the rank of $\hat{X}$, and at the column "$\mathcal{I}(\varepsilon_2)$," the symbol "V" denotes that Property $\mathcal{I}(\varepsilon_2)$ is *violated* and "H" signifies that Property $\mathcal{I}(\epsilon_2)$ *holds*.

Among 90 runs summarized in Tables 1 through 3, there are 87 instances violating Property $\mathcal{I}(\varepsilon_2)$ and only 3 cases holding Property $\mathcal{I}(\varepsilon_2)$. For all these 87 instances, the gaps between "Value 1" and "Value 2" are far less than the tolerance ε_2, which show that Algorithm 7.2 is indeed effective. Furthermore, the rank of the purified solution $\hat{X}$ for the 87 instances are all actually one, meaning that the eigenvector of $\hat{X}$ is the approximate optimal solution for the original CDT problem. We also made a test for two different values of the dimension: $n = 5$ and $n = 50$. Tables 2 and 3 show that it is less likely for Property $\mathcal{I}(\varepsilon_2)$ to hold for the larger n.

REFERENCES

[1] A. BECK AND Y. ELDAR, *Strong duality in nonconvex quadratic optimization with two quadratic constraints*, SIAM J. Optim., 17 (2006), pp. 844–860.

[2] L. BLUM, M. SHUB, AND S. SMALE, *On a theory of computation and complexity over the real numbers: NP-completeness, recursive functions and universal machines*, Bull. Amer. Math. Soc., 21 (1989), pp. 1–46.

[3] M.R. CELIS, J.E. DENNIS, AND R.A. TAPIA, *A trust region algorithm for nonlinear equality constrained optimization*, in Numerical Optimization, R.T. Boggs, R.H. Byrd, and R.B. Schnabel, eds., SIAM, Philadelphia, 1984, pp. 71–82.

[4] X. CHEN AND Y. YUAN, *On local solutions of the Celis–Dennis–Tapia subproblem*, SIAM J. Optim., 10 (2000), pp. 359–383.

[5] X. Chen and Y. Yuan, *On maxima of dual function of the CDT subproblem*, J. Comput. Math., 19 (2001), pp. 113–124.
[6] A.R. Conn, N.L.M. Gould, and P.L. Toint, *Trust-Region Methods*, MPS/SIAM Ser. Optim. 1, SIAM, Philadelphia, 2000.
[7] P. Finsler, *Über das Vorkommen definiter und semidefiniteter Formen in Scharen quadratischer Formen*, Comment. Math. Helv., 9 (1937), pp 188–192.
[8] Y.W. Huang and S. Zhang, *Complex matrix decomposition and quadratic programming*, Math. Oper. Res., 32 (2007), pp. 758–768.
[9] J.M. Martinez, *Local minimizers of quadratic functions on Euclidean balls and spheres*, SIAM J. Optim., 4 (1994), pp. 159–176.
[10] J. Peng and Y. Yuan, *Optimality conditions for the minimization of a quadratic with two quadratic constraints*, SIAM J. Optim., 7 (1997), pp. 579–594.
[11] J.F. Sturm and S. Zhang, *On cones of nonnegative quadratic functions*, Math. Oper. Res., 28 (2003), pp. 246–267.
[12] H. Wolkowicz, R. Saigal, and L. Vandenberghe, eds., *Handbook on Semidefinite Programming: Theory, Algorithms, and Applications*, Kluwer Academic Publishers, Dordrecht, 2000.
[13] Y. Ye and S. Zhang, *New results on quadratic minimization*, SIAM J. Optim., 14 (2003), pp. 245–267.
[14] Y. Yuan, *On a subproblem of trust region algorithms for constrained optimization*, Math. Program., 47 (1990), pp. 53–63.
[15] Y. Yuan, *A dual algorithm for minimizing a quadratic function with two quadratic constraints*, J. Comput. Math., 9 (1991), pp. 348–359.
[16] Y. Zhang, *Computing a Celis-Dennis-Tapia trust-region step for equality constrained optimization*, Math. Program., 55 (1992), pp. 109–124.

SIAM J. OPTIM.
Vol. 19, No. 4, pp. 1757–1780

APPROXIMATE PRIMAL SOLUTIONS AND RATE ANALYSIS FOR DUAL SUBGRADIENT METHODS*

ANGELIA NEDIĆ† AND ASUMAN OZDAGLAR‡

Abstract. In this paper, we study methods for generating approximate primal solutions as a byproduct of subgradient methods applied to the Lagrangian dual of a primal convex (possibly nondifferentiable) constrained optimization problem. Our work is motivated by constrained primal problems with a favorable dual problem structure that leads to efficient implementation of dual subgradient methods, such as the recent resource allocation problems in large-scale networks. For such problems, we propose and analyze dual subgradient methods that use averaging schemes to generate approximate primal optimal solutions. These algorithms use a constant stepsize in view of its simplicity and practical significance. We provide estimates on the primal infeasibility and primal suboptimality of the generated approximate primal solutions. These estimates are given per iteration, thus providing a basis for analyzing the trade-offs between the desired level of error and the selection of the stepsize value. Our analysis relies on the Slater condition and the inherited boundedness properties of the dual problem under this condition. It also relies on the boundedness of subgradients, which is ensured by assuming the compactness of the constraint set.

Key words. subgradient methods, averaging, approximate primal solutions, convergence rate estimates

AMS subject classifications. 90C25, 90C30

DOI. 10.1137/070708111

1. Introduction.

Lagrangian relaxation and duality have been effective tools for solving large-scale convex optimization problems and for systematically providing lower bounds on the optimal value of nonconvex (continuous and discrete) optimization problems. Subgradient methods have played a key role in this framework providing computationally efficient means to obtain near-optimal dual solutions and bounds on the optimal value of the original optimization problem. Most remarkably, in networking applications, over the last few years, subgradient methods have been used with great success in developing decentralized cross-layer resource allocation mechanisms (see Low and Lapsley [18], Shakkottai and Srikant [30], and Srikant [33] for more on this subject).

Subgradient methods for solving nondifferentiable problems have been studied extensively starting with Polyak [26] and Ermoliev [9]. Their convergence properties under various stepsize rules have been established, for example, in Shor [32], Demyanov and Vasilyev [8], Polyak [27], Hiriart-Urruty and Lemaréchal [10], Bertsekas [4], and Bertsekas, Nedić, and Ozdaglar [5]. Nonasymptotic convergence rate estimates have been provided in the seminal work of Nemirovskii and Yudin [23], [24], and more recently in Ben-Tal, Margalit, and Nemirovski [2], Ben-Tal and Nemirovski [3], and Nesterov [25]. Numerous extensions and implementations including parallel and incremental versions have been proposed and analyzed (for example, see Ben-Tal, Margalit, and Nemirovski [2], Ben-Tal and Nemirovski [3], Kiwiel and Lindberg [13],

*Received by the editors November 13, 2007; accepted for publication (in revised form) October 9, 2008; published electronically February 11, 2009.
http://www.siam.org/journals/siopt/19-4/70811.html
†Department of Industrial and Enterprise Systems Engineering, University of Illinois, Urbana, IL 61801 (angelia@uiuc.edu).
‡Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 (asuman@mit.edu).

Zhao, Luh, and Wang [35], Nedić and Bertsekas [20], [21], and Nedić, Bertsekas, and Borkar [22]).

Our development in this paper is motivated by problems with a favorable dual structure, so that solving the problem using dual subgradient methods leads to efficient implementations.[1] For such problems, we develop methods that exploit the subgradient information generated in the dual space directly to construct approximate primal solutions with explicit error estimates. Our methods use an *averaging scheme* that constructs primal solutions by forming running averages of the primal iterates generated when evaluating the subgradient of the dual function. We provide convergence rate estimates for the infeasibility and error estimates on suboptimality of the generated approximate primal solutions.

Averaging schemes for generating primal solutions have been studied in a number of earlier works. Primal-averaging was first proposed and analyzed within a primal-dual subgradient method by Nemirovskii and Yudin [23]. Subsequently, a related primal-averaging scheme based on subgradient information generated by a dual subgradient method has been proposed for linear (primal) optimization problems by Shor [32] and applied to a scheduling problem by Zhurbenko et al. [36]. Shor's ideas have been further developed and computationally tested by Larsson and Liu [14]. Sherali and Choi [31] have extended these results to allow for more general averaging schemes (i.e., more general choices of the weights for convex combinations) and a wider class of stepsize choices. More recently, Larsson, Patriksson, and Strömberg generalized these results in a series of papers (see [15], [16], [17]) to convex constrained optimization problems and demonstrated promising applications of these schemes in the context of traffic equilibrium and road pricing. Sen and Sherali [29] have studied a more complex scheme combining a subgradient method and an auxiliary penalty problem to recover primal solutions. A dual subgradient method producing primal solutions, the volume algorithm, for linear problems has been proposed by Barahona and Anbil [1]. They have reported experimental results for several problems including set partitioning, set covering, and max-cut, but have not analyzed the convergence properties of the algorithm. Kiwiel, Larsson, and Lindberg [12] have studied the convergence of primal-averaging in dual subgradient methods using a target-level based stepsize. Recently, Nesterov [25] has proposed a subgradient algorithm using averaging and provided convergence rate estimates assuming the availability of a bound on the Euclidean norm of an optimal solution. Nesterov's algorithm generates a solution to a convex minimization problem, and it is not a primal-recovery scheme. More recently, Ruszczyński [28] has proposed a new subgradient method that uses averaging to identify both an optimal solution of a convex minimization problem and a subgradient that appears in the optimality condition.

Among the papers cited above, our work is most closely related to the primal-recovery algorithms of Shor [32], Sherali and Choi [31], Larsson, Patriksson, and Strömberg [15], [16], [17], and Kiwiel, Larsson, and Lindberg [12]. The focus of these works is on exact recovery methods for primal solutions and the convergence properties for diminishing stepsize rules (with divergent sum).[2] In contrast, our focus in this paper is on methods generating approximate primal solutions for general

[1] Section 2.2 illustrates a motivating example of a network resource allocation problem, where the use of dual subgradient methods leads to *decentralized resource allocation policies* for communication networks.

[2] The exception is the paper [12], where a target-level based stepsize (i.e., a modification of Polyak's stepsize [26]) has been considered.

(possibly nondifferentiable) convex constrained optimization problems and providing convergence rate guarantees. We consider subgradient methods that use a constant stepsize, mainly because of its practical importance and simplicity for implementations. We provide convergence rate estimates for the approximate solutions under the Slater constraint qualification, including estimates for the amount of feasibility violation, and upper and lower bounds for the primal objective function. Moreover, our estimates are per iteration and illustrate the trade-offs between the approximate solution error and the stepsize value.

This paper is organized as follows: In section 2, we define the primal and dual problems, and provide an explicit bound on the level sets of the dual function under Slater condition. In section 3, we consider a subgradient method with a constant stepsize and study its properties under Slater. In section 4, we introduce approximate primal solutions generated through averaging and provide bounds on their feasibility violation and primal cost values. In section 5, we consider an alternative to the basic subgradient method based on the boundedness properties of the dual optimal solution set under the Slater condition, and we provide error estimates for the generated approximate primal solutions. We conclude in section 6 by summarizing our work and providing some comments.

2. Primal and dual problems. In this section, we formulate the primal and dual problems of interest. We provide a motivating example for the use of dual subgradient methods and give some preliminary results that we use in the subsequent development. We start by introducing the notation and the basic terminology that we use throughout this paper.

2.1. Notation and terminology. We consider the n-dimensional vector space $\mathbb{R}^n$ and the m-dimensional vector space $\mathbb{R}^m$. We view a vector as a column vector, and we denote by $x'y$ the inner product of two vectors x and y. We use $\|y\|$ to denote the standard Euclidean norm, $\|y\| = \sqrt{y'y}$. Occasionally, we also use the standard 1-norm and ∞-norm denoted, respectively, by $\|y\|_1$ and $\|y\|_\infty$, i.e., $\|y\|_1 = \sum_i |y_i|$ and $\|y\|_\infty = \max_i |y_i|$. We write $dist(\bar{y}, Y)$ to denote the standard Euclidean distance of a vector $\bar{y}$ from a set Y, i.e.,

$$dist(\bar{y}, Y) = \inf_{y \in Y} \|\bar{y} - y\|.$$

For a vector $u \in \mathbb{R}^m$, we write u^+ to denote the projection of u on the nonnegative orthant in $\mathbb{R}^m$; i.e., u^+ is the componentwise maximum of the vector u and the zero vector:

$$u^+ = (\max\{0, u_1\}, \ldots, \max\{0, u_m\})' \quad \text{for } u = (u_1, \ldots, u_m)'.$$

For a concave function $q : \mathbb{R}^m \to [-\infty, \infty]$, we denote the domain of q by $\mathrm{dom}(q)$, where

$$\mathrm{dom}(q) = \{\mu \in \mathbb{R}^m \mid q(\mu) > -\infty\}.$$

We use the notion of a subgradient of a concave function $q(\mu)$. In particular, a subgradient $s_{\bar{\mu}}$ of a concave function $q(\mu)$ at a given vector $\bar{\mu} \in \mathrm{dom}(q)$ provides a linear overestimate of the function $q(\mu)$ for all $\mu \in \mathrm{dom}(q)$. We use this as the subgradient defining property: *$s_{\bar{\mu}} \in \mathbb{R}^m$ is a subgradient of a concave function $q(\mu)$ at a given vector $\bar{\mu} \in \mathrm{dom}(q)$* if the following relation holds:

$$q(\bar{\mu}) + s'_{\bar{\mu}}(\mu - \bar{\mu}) \geq q(\mu) \qquad \text{for all } \mu \in \mathrm{dom}(q). \tag{1}$$

The set of all subgradients of q at $\bar{\mu}$ is denoted by $\partial q(\bar{\mu})$.

In this paper, we focus on the following constrained optimization problem:

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && g(x) \leq 0, \\ &&& x \in X, \end{aligned} \tag{2}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function, $g = (g_1, \ldots, g_m)'$ and each $g_j : \mathbb{R}^n \to \mathbb{R}$ is a convex function, and $X \subset \mathbb{R}^n$ is a nonempty closed convex set. We refer to this as the *primal problem*. We denote the primal optimal value by f^*, and throughout this paper, we assume that the value f^* is finite.

To generate approximate solutions to the primal problem in (2), we consider solving its dual using subgradient methods. Here, the *dual problem* is the one arising from the Lagrangian relaxation of the inequality constraints $g(x) \leq 0$, and it is given by

$$\begin{aligned} &\text{maximize} && q(\mu) \\ &\text{subject to} && \mu \geq 0, \\ &&& \mu \in \mathbb{R}^m, \end{aligned} \tag{3}$$

where q is the dual function defined by

$$q(\mu) = \inf_{x \in X} \{f(x) + \mu' g(x)\}. \tag{4}$$

We often refer to a vector $\mu \in \mathbb{R}^m$ with $\mu \geq 0$ as a *multiplier*. We denote the dual optimal value by q^* and the dual optimal set by M^*. We say that there is *zero duality gap* if the optimal values of the primal and the dual problems are equal, i.e., $f^* = q^*$.

We assume that the minimization problem associated with the evaluation of the dual function $q(\mu)$ has a solution for every $\mu \geq 0$. This is the case, for instance, when the set X is compact (since f and $g_j's$ are continuous due to being convex over $\mathbb{R}^n$). Furthermore, we assume that the minimization problem in (4) is simple enough so that it either has a closed form solution or can be solved efficiently. For example, this is the case when the functions f and g_j's are affine or affine plus norm-square term (i.e., $c\|x\|^2 + a'x + b$), and the set X is a nonnegative orthant or a box in $\mathbb{R}^n$. Many practical problems of interest, such as those arising in network optimization (see section 2.2), often have this structure.

In our subsequent development, we consider subgradient methods as applied to the dual problem given by (3) and (4). Due to the form of the dual function q, the subgradients of q at a vector μ are related to the primal vectors x_μ attaining the minimum in (4). Specifically, the set $\partial q(\mu)$ of subgradients of q at a given $\mu \geq 0$ is given by

$$\partial q(\mu) = \operatorname{conv}\left(\{g(x_\mu) \mid x_\mu \in X_\mu\}\right), \qquad X_\mu = \{x_\mu \in X \mid q(\mu) = f(x_\mu) + \mu' g(x_\mu)\}, \tag{5}$$

where $\operatorname{conv}(Y)$ denotes the convex hull of a set Y.

Before proceeding with our analysis, we discuss an example that motivates the subsequent development of dual subgradient methods.

2.2. Motivating example. Here, we describe the canonical utility-based network resource allocation problem and briefly discuss how dual subgradient methods lead to decentralized policies that can be used over a network. This approach was

proposed in the seminal work of Kelly, Maulloo, and Tan [11] and further developed by Low and Lapsley [18], Shakkottai and Srikant [30], and Srikant [33].

Consider a network that consists of a set $\mathcal{S} = \{1, \ldots, S\}$ of sources and a set $\mathcal{L} = \{1, \ldots, L\}$ of undirected links, where a link l has capacity c_l. Let $\mathcal{L}(i) \subset \mathcal{L}$ denote the set of links used by source i. The application requirements of source i are represented by a concave increasing utility function $u_i : [0, \infty) \to [0, \infty)$; i.e., each source i gains a utility $u_i(x_i)$ when it sends data at a rate x_i. We further assume that rate x_i is constrained to lie in the interval $I_i = [0, M_i]$ for all $i \in \mathcal{S}$, where the scalar M_i denotes the maximum allowed rate for source i. Let $\mathcal{S}(l) = \{i \in \mathcal{S} \mid l \in \mathcal{L}(i)\}$ denote the set of sources that use link l. The goal of the *network utility maximization problem* is to allocate the source rates as the optimal solution of the problem

$$\begin{aligned} &\text{maximize} \sum_{i \in \mathcal{S}} u_i(x_i) && (6)\\ &\text{subject to} \sum_{i \in \mathcal{S}(l)} x_i \le c_l \quad \text{for all } l \in \mathcal{L}, \\ &x_i \in I_i \quad \text{for all } i \in \mathcal{S}. && (7) \end{aligned}$$

Solving problem (6) directly by applying existing subgradient methods requires coordination among sources and therefore may be impractical for real networks. This is because in real networks, such as the Internet, there is no central entity that has access to both the source utility functions and the capacity of all the links in the network. The utility function $u_i(x_i)$ is known only by source i, while the link capacities may be known only by "network providers." At the same time, in view of the separable structure of the objective and constraint functions, the dual problem can be evaluated exactly while using decentralized information. In particular, the dual problem of (6) is given by (3), where the dual function takes the form

$$\begin{aligned} q(\mu) &= \max_{x_i \in I_i,\ i \in \mathcal{S}} \sum_{i \in \mathcal{S}} u_i(x_i) - \sum_{l \in \mathcal{L}} \mu_l \left(\sum_{i \in \mathcal{S}(l)} x_i - c_l \right) \\ &= \max_{x_i \in I_i,\ i \in \mathcal{S}} \sum_{i \in \mathcal{S}} \left(u_i(x_i) - x_i \sum_{l \in \mathcal{L}(i)} \mu_l \right) + \sum_{l \in \mathcal{L}} \mu_l c_l. \end{aligned}$$

Since the optimization problem on the right-hand side of the preceding relation is separable in the variables x_i, the problem decomposes into subproblems for each source i. Letting $\mu_i = \sum_{l \in \mathcal{L}(i)} \mu_l$ for each i (i.e., μ_i is the sum of the multipliers corresponding to the links used by source i), we can write the dual function as

$$q(\mu) = \sum_{i \in \mathcal{S}} \max_{x_i \in I_i} \{u_i(x_i) - x_i \mu_i\} + \sum_{l \in \mathcal{L}} \mu_l c_l.$$

Hence, to evaluate the dual function, each source i needs to solve the one-dimensional optimization problem $\max_{x_i \in I_i} \{u_i(x_i) - x_i \mu_i\}$. This involves only its own utility function u_i and the value μ_i, which is available to source i in practical networks (through a direct feedback mechanism from its destination).

This favorable structure of the dual problem has motivated much interest in using dual subgradient methods to solve the network utility maximization problem in an iterative decentralized manner (see Chiang et al. [6]). Other problems where

the dual problem has a structure that allows exact evaluation of the dual function using local information include the problem of processor speed control considered by Mutapcic et al. [19], and the traffic equilibrium and road pricing problems considered by Larsson, Patriksson, and Strömberg [15], [16], [17].

2.3. Slater condition and boundedness of the multiplier sets. In this section, we consider sets of the form $\{\mu \geq 0 \mid q(\mu) \geq q(\bar{\mu})\}$ for a fixed $\bar{\mu} \geq 0$, which are obtained by intersecting the nonnegative orthant in $\mathbb{R}^m$ and (upper) level sets of the concave dual function q. We show that these sets are bounded when the primal problem satisfies the standard Slater constraint qualification, formally given in the following.

Assumption 1 (Slater condition). There exists a vector $\bar{x} \in X$ such that

$$g_j(\bar{x}) < 0 \qquad \text{for all } j = 1, \ldots, m.$$

We refer to a vector $\bar{x}$ satisfying the Slater condition as *a Slater vector*.

Under the assumption that f^* is finite, it is well known that the Slater condition is sufficient for a zero duality gap as well as for the existence of a dual optimal solution (see, for example, Bertsekas [4] or Bertsekas, Nedić, and Ozdaglar [5]). Furthermore, the dual optimal set is bounded (see Hiriart-Urruty and Lemaréchal [10]). This property of the dual optimal set under the Slater condition has been observed and used as early as in Uzawa's analysis of the Arrow–Hurwicz gradient method in [34]. This property will be key in our subsequent development and analysis.

The following proposition extends the result on the optimal dual set boundedness under the Slater condition. In particular, it shows that the Slater condition also guarantees the boundedness of the (level) sets $\{\mu \geq 0 \mid q(\mu) \geq q(\bar{\mu})\}$.

LEMMA 1. *Let the Slater condition hold (cf. Assumption* 1*). Then, the set $Q_{\bar{\mu}}$ is bounded and, in particular, we have*

$$\max_{\mu \in Q_{\bar{\mu}}} \|\mu\| \leq \frac{1}{\gamma} \left(f(\bar{x}) - q(\bar{\mu})\right),$$

where $\gamma = \min_{1 \leq j \leq m}\{-g_j(\bar{x})\}$ and $\bar{x}$ is a Slater vector.

Proof. We have, for any $\mu \in Q_{\bar{\mu}}$,

$$q(\bar{\mu}) \leq q(\mu) = \inf_{x \in X} \{f(x) + \mu' g(x)\} \leq f(\bar{x}) + \mu' g(\bar{x}) = f(\bar{x}) + \sum_{j=1}^{m} \mu_j g_j(\bar{x}),$$

implying that

$$-\sum_{j=1}^{m} \mu_j g_j(\bar{x}) \leq f(\bar{x}) - q(\bar{\mu}).$$

Because $g_j(\bar{x}) < 0$ and $\mu_j \geq 0$ for all j, it follows that

$$\min_{1 \leq j \leq m}\{-g_j(\bar{x})\} \sum_{j=1}^{m} \mu_j \leq -\sum_{j=1}^{m} \mu_j g_j(\bar{x}) \leq f(\bar{x}) - q(\bar{\mu}).$$

Therefore,

$$\sum_{j=1}^{m} \mu_j \leq \frac{f(\bar{x}) - q(\bar{\mu})}{\min_{1 \leq j \leq m}\{-g_j(\bar{x})\}}.$$

Since $\mu \geq 0$, we have $\|\mu\| \leq \sum_{j=1}^{m} \mu_j$ and the estimate follows. ∎

It follows from the preceding lemma that under the Slater condition, the dual optimal set M^* is nonempty. In particular, by noting that $M^* = \{\mu \geq 0 \mid q(\mu) \geq q^*\}$ and by using Lemma 1, we see that

$$\max_{\mu^* \in M^*} \|\mu^*\| \leq \frac{1}{\gamma} \left(f(\bar{x}) - q^*\right), \tag{8}$$

with $\gamma = \min_{1 \leq j \leq m}\{-g_j(\bar{x})\}$.

In practice, the dual optimal value q^* is not readily available. However, having a dual function value $q(\tilde{\mu})$ for some $\tilde{\mu} \geq 0$, we can still provide a bound on the norm of the dual optimal solutions. In particular, since $q^* \geq q(\tilde{\mu})$, from relation (8) we obtain the following bound:

$$\max_{\mu^* \in M^*} \|\mu^*\| \leq \frac{1}{\gamma} \left(f(\bar{x}) - q(\tilde{\mu})\right).$$

Furthermore, having any multiplier sequence $\{\mu_k\}$, we can use the dual function values $q(\mu_k)$ to generate a sequence of (possibly improving) upper bounds on the dual optimal solution norms $\|\mu^*\|$. Formally, since $q^* \geq \max_{0 \leq i \leq k} q(\mu_i)$, from relation (8) we have

$$\max_{\mu^* \in M^*} \|\mu^*\| \leq \frac{1}{\gamma} \left(f(\bar{x}) - \max_{0 \leq i \leq k} q(\mu_i)\right) \qquad \text{for all } k \geq 0.$$

Note that these bounds are nonincreasing in k. These bounds have important implications in the development and analysis of subgradient methods since they allow us to "locate dual optimal solutions" by using only a Slater vector $\bar{x}$ and a multiplier sequence $\{\mu_k\}$ generated by a subgradient method.

Such bounds play a key role in our subsequent development. In particular, we use these bounds to provide error estimates of our approximate solutions as well as to design a dual algorithm that projects on a set containing the dual optimal solution.

3. Subgradient method. To solve the dual problem, we consider the classical subgradient algorithm with a constant stepsize:

$$\mu_{k+1} = [\mu_k + \alpha g_k]^+ \qquad \text{for } k = 0, 1, \ldots, \tag{9}$$

where the vector $\mu_0 \geq 0$ is an initial iterate and the scalar $\alpha > 0$ is a stepsize. The vector g_k is a subgradient of q at μ_k given by

$$g_k = g(x_k), \qquad x_k \in \operatorname*{argmin}_{x \in X}\{f(x) + \mu_k' g(x)\} \qquad \text{for all } k \geq 0; \tag{10}$$

see (5).

Our choice of the constant stepsize is primarily motivated by its practical importance and, in particular, because in practice the stepsize typically stays bounded away from zero. Furthermore, the error estimates for this stepsize can be explicitly written in terms of the problem parameters that are often available. Also, when implementing a subgradient method with a constant stepsize rule, the stepsize length α is the only parameter that a user has to select, which is often preferred to more complex stepsize choices involving several stepsize parameters without a good guidance on their selection.

3.1. Basic relations. In this section, we establish some basic relations that hold for a sequence $\{\mu_k\}$ obtained by the subgradient algorithm of (9). These properties are important in our construction of approximate primal solutions and, in particular, in our analysis of the error estimates of these solutions.

We start with a lemma providing some basic relations that hold under minimal assumptions. The relations given in this lemma have been known and used in various ways to analyze subgradient approaches (for example, see Shor [32], Polyak [27], Demyanov and Vasilyev [8], Correa and Lemaréchal [7], Nedić and Bertsekas [20], [21]). The proofs are provided here for completeness.

LEMMA 2 (basic iterate relation). *Let the sequence $\{\mu_k\}$ be generated by the subgradient algorithm* (9). *We then have the following:*

(a) *For any $\mu \geq 0$,*

$$\|\mu_{k+1} - \mu\|^2 \leq \|\mu_k - \mu\|^2 - 2\alpha\,(q(\mu) - q(\mu_k)) + \alpha^2\|g_k\|^2 \qquad \text{for all } k \geq 0.$$

(b) *When the optimal solution set M^* is nonempty, there holds*

$$dist^2(\mu_{k+1}, M^*) \leq dist^2(\mu_k, M^*) - 2\alpha\,(q^* - q(\mu_k)) + \alpha^2\|g_k\|^2 \qquad \text{for all } k \geq 0,$$

where $dist(y, Y)$ denotes the Euclidean distance from a vector y to a set Y.

Proof. (a) By using the nonexpansive property of the projection operation, from relation (9) we obtain, for any $\mu \geq 0$ and all k,

$$\|\mu_{k+1} - \mu\|^2 = \left\|[\mu_k + \alpha g_k]^+ - \mu\right\|^2 \leq \|\mu_k + \alpha g_k - \mu\|^2.$$

Therefore,

$$\|\mu_{k+1} - \mu\|^2 \leq \|\mu_k - \mu\|^2 + 2\alpha g_k'(\mu_k - \mu) + \alpha^2\|g_k\|^2 \qquad \text{for all } k.$$

Since g_k is a subgradient of q at μ_k (cf. (1)), we have

$$g_k'(\mu - \mu_k) \geq q(\mu) - q(\mu_k),$$

implying that

$$g_k'(\mu_k - \mu) \leq -\,(q(\mu) - q(\mu_k)).$$

Hence, for any $\mu \geq 0$,

$$\|\mu_{k+1} - \mu\|^2 \leq \|\mu_k - \mu\|^2 - 2\alpha\,(q(\mu) - q(\mu_k)) + \alpha^2\|g_k\|^2 \qquad \text{for all } k.$$

(b) By using the preceding relation with $\mu = \mu^*$ for any optimal solution μ^*, we obtain

$$\|\mu_{k+1} - \mu^*\|^2 \leq \|\mu_k - \mu^*\|^2 - 2\alpha\,(q^* - q(\mu_k)) + \alpha^2\|g_k\|^2 \qquad \text{for all } k \geq 0.$$

The desired relation follows by taking the infimum over all $\mu^* \in M^*$ in both sides of the preceding relation. ∎

3.2. Bounded multipliers. Here, we show that the multiplier sequence $\{\mu_k\}$ produced by the subgradient algorithm is bounded under the Slater condition and the bounded subgradient assumption. We formally state the latter requirement in the following assumption.

Assumption 2 (bounded set X). The constraint set X in problem (2) is bounded.

Under this assumption, due to the convexity of the constraint functions g_j over $\mathbb{R}^n$, each g_j is continuous over $\mathbb{R}^n$. Thus, $\max_{x\in X}\|g(x)\|$ is finite and provides an upper bound on the norms of the subgradients g_k, i.e.,

$$\|g_k\| \le L \qquad \text{for all } k \ge 0, \qquad \text{with } \ L = \max_{x\in X}\|g(x)\|.$$

In the following lemma, we establish the boundedness of the multiplier sequence. In this, we use the boundedness of the dual sets $\{\mu \ge 0 \mid q(\mu) \ge q(\bar{\mu})\}$ (cf. Lemma 1) and the basic relation for the sequence $\{\mu_k\}$ of Lemma 2(a).

LEMMA 3 (bounded multipliers). *Let the multiplier sequence $\{\mu_k\}$ be generated by the subgradient algorithm of* (9)*. Also, let the Slater condition and the bounded set assumption hold (cf. Assumptions* 1 *and* 2*). Then, the sequence $\{\mu_k\}$ is bounded and, in particular, we have*

$$\|\mu_k\| \le \frac{2}{\gamma}\,(f(\bar{x}) - q^*) + \max\left\{\|\mu_0\|, \frac{1}{\gamma}\,(f(\bar{x}) - q^*) + \frac{\alpha L^2}{2\gamma} + \alpha L\right\},$$

where $\gamma = \min_{1\le j\le m}\{-g_j(\bar{x})\}$, $\bar{x}$ is a Slater vector, L is the subgradient norm bound, and $\alpha > 0$ is the stepsize.

Proof. Under the Slater condition the optimal dual set M^* is nonempty. Consider the set Q_α defined by

$$Q_\alpha = \left\{\mu \ge 0 \mid q(\mu) \ge q^* - \frac{\alpha L^2}{2}\right\},$$

which is nonempty in view of $M^* \subset Q_\alpha$. We fix an arbitrary $\mu^* \in M^*$, and we first prove that, for all $k \ge 0$,

$$\|\mu_k - \mu^*\| \le \max\left\{\|\mu_0 - \mu^*\|, \frac{1}{\gamma}\,(f(\bar{x}) - q^*) + \frac{\alpha L^2}{2\gamma} + \|\mu^*\| + \alpha L\right\}, \tag{11}$$

where $\gamma = \min_{1\le j\le m}\{-g_j(\bar{x})\}$ and L is the bound on the subgradient norms $\|g_k\|$. Then, we use Lemma 1 to prove the desired estimate.

We show that relation (11) holds by induction on k. Note that the relation holds for $k = 0$. Assume now that it holds for some $k > 0$, i.e.,

$$\|\mu_k - \mu^*\| \le \max\left\{\|\mu_0 - \mu^*\|, \frac{1}{\gamma}\,(f(\bar{x}) - q^*) + \frac{\alpha L^2}{2\gamma} + \|\mu^*\| + \alpha L\right\} \text{ for some } k > 0. \tag{12}$$

We now consider two cases: $q(\mu_k) \ge q^* - \alpha L^2/2$ and $q(\mu_k) < q^* - \alpha L^2/2$.

Case 1. $q(\mu_k) \ge q^* - \alpha L^2/2$. By using the definition of the iterate μ_{k+1} in (9) and the subgradient boundedness, we obtain

$$\|\mu_{k+1} - \mu^*\| \le \|\mu_k + \alpha g_k - \mu^*\| \le \|\mu_k\| + \|\mu^*\| + \alpha L.$$

Since $q(\mu_k) \geq q^* - \alpha L^2/2$, it follows that $\mu_k \in Q_\alpha$. According to Lemma 1, the set Q_α is bounded and, in particular, $\|\mu\| \leq \frac{1}{\gamma}\left(f(\bar{x}) - q^* + \alpha L^2/2\right)$ for all $\mu \in Q_\alpha$. Therefore

$$\|\mu_k\| \leq \frac{1}{\gamma}\left(f(\bar{x}) - q^*\right) + \frac{\alpha L^2}{2\gamma}.$$

By combining the preceding two relations, we obtain

$$\|\mu_{k+1} - \mu^*\| \leq \frac{1}{\gamma}\left(f(\bar{x}) - q^*\right) + \frac{\alpha L^2}{2\gamma} + \|\mu^*\| + \alpha L,$$

thus showing that the estimate in (11) holds for $k+1$.

Case 2. $q(\mu_k) < q^* - \alpha L^2/2$. By using Lemma 2(a) with $\mu = \mu^*$, we obtain

$$\|\mu_{k+1} - \mu^*\|^2 \leq \|\mu_k - \mu^*\|^2 - 2\alpha\left(q^* - q(\mu_k)\right) + \alpha^2\|g_k\|^2.$$

By using the subgradient boundedness, we further obtain

$$\|\mu_{k+1} - \mu^*\|^2 \leq \|\mu_k - \mu^*\|^2 - 2\alpha\left(q^* - q(\mu_k) - \frac{\alpha L^2}{2}\right).$$

Since $q(\mu_k) < q^* - \alpha L^2/2$, it follows that $q^* - q(\mu_k) - \alpha L^2/2 > 0$, which when combined with the preceding relation yields

$$\|\mu_{k+1} - \mu^*\| < \|\mu_k - \mu^*\|.$$

By the induction hypothesis (cf. (12)), it follows that the estimate in (11) holds for $k+1$ as well. Hence, the estimate in (11) holds for all $k \geq 0$.

From (11) we obtain, for all $k \geq 0$,

$$\|\mu_k\| \leq \|\mu_k - \mu^*\| + \|\mu^*\| \leq \max\left\{\|\mu_0 - \mu^*\|, \frac{1}{\gamma}\left(f(\bar{x}) - q^*\right) + \frac{\alpha L^2}{2\gamma} + \|\mu^*\| + \alpha L\right\} + \|\mu^*\|.$$

By using $\|\mu_0 - \mu^*\| \leq \|\mu_0\| + \|\mu^*\|$, we further have, for all $k \geq 0$,

$$\begin{aligned}\|\mu_k\| &\leq \max\left\{\|\mu_0\| + \|\mu^*\|, \frac{1}{\gamma}\left(f(\bar{x}) - q^*\right) + \frac{\alpha L^2}{2\gamma} + \|\mu^*\| + \alpha L\right\} + \|\mu^*\| \\ &= 2\|\mu^*\| + \max\left\{\|\mu_0\|, \frac{1}{\gamma}\left(f(\bar{x}) - q^*\right) + \frac{\alpha L^2}{2\gamma} + \alpha L\right\}.\end{aligned}$$

Since $M^* = \{\mu \geq 0 \mid q(\mu) \geq q^*\}$, according to Lemma 1 we have the following bound on the dual optimal solutions:

$$\max_{\mu^* \in M^*} \|\mu^*\| \leq \frac{1}{\gamma}\left(f(\bar{x}) - q^*\right),$$

implying that, for all $k \geq 0$,

$$\|\mu_k\| \leq \frac{2}{\gamma}\left(f(\bar{x}) - q^*\right) + \max\left\{\|\mu_0\|, \frac{1}{\gamma}\left(f(\bar{x}) - q^*\right) + \frac{\alpha L^2}{2\gamma} + \alpha L\right\}. \qquad \square$$

The error estimate of Lemma 3 depends explicitly on the dual optimal value q^*. In practice, the value q^* is not readily available. However, since $q^* \geq q(\mu_0)$, by replacing q^* with $q(\mu_0)$, we have obtain the following norm bound for the multiplier sequence:

$$\|\mu_k\| \leq \frac{2}{\gamma}\left(f(\bar{x}) - q(\mu_0)\right) + \max\left\{\|\mu_0\|, \frac{1}{\gamma}\left(f(\bar{x}) - q(\mu_0)\right) + \frac{\alpha L^2}{2\gamma} + \alpha L\right\},$$

where $\gamma = \min_{1\leq j\leq m}\{-g_j(\bar{x})\}$. Note that this bound depends on the algorithm parameters and problem data only. Specifically, it involves the initial iterate μ_0 of the subgradient method, the stepsize α, the vector $\bar{x}$ satisfying the Slater condition, and the subgradient norm bound L. In some practical applications, such as those in network optimization, such data is readily available. One may think of optimizing this bound with respect to the Slater vector $\bar{x}$. This might be an interesting and challenging problem on its own. However, this is outside the scope of our paper.

4. Approximate primal solutions. In this section, we provide approximate primal solutions by considering the running averages of the primal sequence $\{x_k\}$ generated as a byproduct of the subgradient method (cf. (10)). Intuitively, one would expect that, by averaging, the primal cost and the amount of constraint violation of primal infeasible vectors can be reduced due to the convexity of the cost and the constraint functions. It turns out that the benefits of averaging are far more reaching than merely cost and infeasibility reduction. We show here that under the Slater condition, we can also provide upper bounds for the number of subgradient iterations needed to generate a primal solution within a given level of constraint violation. We also derive upper and lower bounds on the gap from the optimal primal value. These bounds depend on some assumptions and prior information such as a Slater vector and a bound on subgradient norms.

We now introduce the notation that we use in our averaging scheme throughout the rest of this paper. We consider the multiplier sequence $\{\mu_k\}$ generated by the subgradient algorithm of (9), and the corresponding sequence of primal vectors $\{x_k\} \subset X$ that provide the subgradients g_k in the algorithm, i.e.,

$$g_k = g(x_k), \qquad x_k \in \operatorname*{argmin}_{x\in X}\{f(x) + \mu_k' g(x)\} \qquad \text{for all } k \geq 0$$

(cf. (10)). We define $\hat{x}_k$ as the average of the vectors $x_0, \ldots, x_{k-1}$, i.e.,

$$\hat{x}_k = \frac{1}{k}\sum_{i=0}^{k-1} x_i \qquad \text{for all } k \geq 1. \tag{13}$$

The average vectors $\hat{x}_k$ lie in the set X because X is convex and $x_i \in X$ for all i. However, these vectors need not satisfy the primal inequality constraints $g_j(x) \leq 0$, $j = 1, \ldots, m$, and therefore, they can be primal infeasible.

In the rest of this section, we study some basic properties of the average vectors $\hat{x}_k$. Using these properties and the Slater condition, we provide estimates for the primal optimal value and the feasibility violation at each iteration of the subgradient method.

4.1. Basic properties of the averaged primal sequence. In this section, we provide upper and lower bounds on the primal cost of the running averages $\hat{x}_k$. We also provide an upper and a lower bound on the amount of feasibility violation of these vectors. These bounds are given per iteration, as seen in the following.

PROPOSITION 1. *Let the multiplier sequence $\{\mu_k\}$ be generated by the subgradient method of* (9)*. Let the vectors $\hat{x}_k$ for $k \geq 1$ be the averages given by* (13)*. Then, for all $k \geq 1$, the following hold:*

(a) *An upper bound on the amount of constraint violation of the vector $\hat{x}_k$ is given by*

$$\left\|g(\hat{x}_k)^+\right\| \leq \frac{\|\mu_k\|}{k\alpha}.$$

(b) *An upper bound on the primal cost of the vector $\hat{x}_k$ is given by*

$$f(\hat{x}_k) \leq q^* + \frac{\|\mu_0\|^2}{2k\alpha} + \frac{\alpha}{2k}\sum_{i=0}^{k-1}\|g(x_i)\|^2.$$

(c) *A lower bound on the primal cost of the vector $\hat{x}_k$ is given by*

$$f(\hat{x}_k) \geq q^* - \|\mu^*\|\|g(\hat{x}_k)^+\|,$$

where μ^* is a dual optimal solution.

Proof. (a) By using the definition of the iterate μ_{k+1} in (9), we obtain

$$\mu_k + \alpha g_k \leq [\mu_k + \alpha g_k]^+ = \mu_{k+1} \qquad \text{for all } k \geq 0.$$

Since $g_k = g(x_k)$ with $x_k \in X$, it follows that

$$\alpha g(x_k) \leq \mu_{k+1} - \mu_k \qquad \text{for all } k \geq 0.$$

Therefore,

$$\sum_{i=0}^{k-1} \alpha g(x_i) \leq \mu_k - \mu_0 \leq \mu_k \qquad \text{for all } k \geq 1,$$

where the last inequality in the preceding relation follows from $\mu_0 \geq 0$. Since $x_k \in X$ for all k, by the convexity of X, we have $\hat{x}_k \in X$ for all k. Hence, by the convexity of each of the functions g_j, it follows that

$$g(\hat{x}_k) \leq \frac{1}{k}\sum_{i=0}^{k-1} g(x_i) = \frac{1}{k\alpha}\sum_{i=0}^{k-1}\alpha g(x_i) \leq \frac{\mu_k}{k\alpha} \qquad \text{for all } k \geq 1.$$

Because $\mu_k \geq 0$ for all k, we have $g(\hat{x}_k)^+ \leq \mu_k/(k\alpha)$ for all $k \geq 1$, and, therefore,

$$\|g(\hat{x}_k)^+\| \leq \frac{\|\mu_k\|}{k\alpha} \qquad \text{for all } k \geq 1.$$

(b) By the convexity of the primal cost $f(x)$ and the definition of x_k as a minimizer of the Lagrangian function $f(x) + \mu_k' g(x)$ over $x \in X$ (cf. (10)), we have

$$f(\hat{x}_k) \leq \frac{1}{k}\sum_{i=0}^{k-1} f(x_i) = \frac{1}{k}\sum_{i=0}^{k-1}\{f(x_i) + \mu_i' g(x_i)\} - \frac{1}{k}\sum_{i=0}^{k-1}\mu_i' g(x_i).$$

Since $q(\mu_i) = f(x_i) + \mu_i' g(x_i)$ and $q(\mu_i) \leq q^*$ for all i, it follows that, for all $k \geq 1$,

$$f(\hat{x}_k) \leq \frac{1}{k}\sum_{i=0}^{k-1} q(\mu_i) - \frac{1}{k}\sum_{i=0}^{k-1}\mu_i' g(x_i) \leq q^* - \frac{1}{k}\sum_{i=0}^{k-1}\mu_i' g(x_i). \tag{14}$$

From the definition of the algorithm in (9), by using the nonexpansive property of the projection, and the facts $0 \in \{\mu \in \mathbb{R}^m \mid \mu \geq 0\}$ and $g_i = g(x_i)$, we obtain

$$\|\mu_{i+1}\|^2 \leq \|\mu_i\|^2 + 2\alpha\mu_i' g(x_i) + \alpha^2\|g(x_i)\|^2 \qquad \text{for all } i \geq 0,$$

implying that

$$-\mu_i' g(x_i) \leq \frac{\|\mu_i\|^2 - \|\mu_{i+1}\|^2 + \alpha^2 \|g(x_i)\|^2}{2\alpha} \qquad \text{for all } i \geq 0.$$

By summing over $i = 0, \ldots, k-1$ for $k \geq 1$, we have

$$-\frac{1}{k} \sum_{i=0}^{k-1} \mu_i' g(x_i) \leq \frac{\|\mu_0\|^2 - \|\mu_k\|^2}{2k\alpha} + \frac{\alpha}{2k} \sum_{i=0}^{k-1} \|g(x_i)\|^2 \qquad \text{for all } k \geq 1.$$

Combining the preceding relation and (14), we further have

$$f(\hat{x}_k) \leq q^* + \frac{\|\mu_0\|^2 - \|\mu_k\|^2}{2k\alpha} + \frac{\alpha}{2k} \sum_{i=0}^{k-1} \|g(x_i)\|^2 \qquad \text{for all } k \geq 1,$$

implying the desired estimate.

(c) Given a dual optimal solution μ^*, we have

$$f(\hat{x}_k) = f(\hat{x}_k) + (\mu^*)' g(\hat{x}_k) - (\mu^*)' g(\hat{x}_k) \geq q(\mu^*) - (\mu^*)' g(\hat{x}_k).$$

Because $\mu^* \geq 0$ and $g(\hat{x}_k)^+ \geq g(\hat{x}_k)$, we further have

$$-(\mu^*)' g(\hat{x}_k) \geq -(\mu^*)' g(\hat{x}_k)^+ \geq -\|\mu^*\| \|g(\hat{x}_k)^+\|.$$

From the preceding two relations and the fact $q(\mu^*) = q^*$, it follows that

$$f(\hat{x}_k) \geq q^* - \|\mu^*\| \|g(\hat{x}_k)^+\|. \qquad \square$$

An immediate consequence of Proposition 1(a) is that the maximum violation $\|g(\hat{x}_k)^+\|_\infty$ of constraints $g_j(x)$, $j = 1, \ldots, m$, at $x = \hat{x}_k$ is bounded by the same bound. In particular, we have

$$\max_{1 \leq j \leq m} g_j(\hat{x}_k)^+ \leq \frac{\|\mu_k\|}{k\alpha} \qquad \text{for all } k \geq 1,$$

which follows from the proposition in view of the relation $\|y\|_\infty \leq \|y\|$ for any y.

We note that the results of Proposition 1 in parts (a) and (c) show how the amount of feasibility violation $\|g(\hat{x}_k)^+\|$ affects the lower estimate of $f(\hat{x}_k)$. Furthermore, we note that the results of Proposition 1 indicate that the bounds on the feasibility violation and the primal value $f(\hat{x}_k)$ are readily available provided that we have bounds on the multiplier norms $\|\mu_k\|$, optimal solution norms $\|\mu^*\|$, and subgradient norms $\|g(x_k)\|$. This is precisely what we use in the next section to establish our estimates.

Let us also note that the bounds on the primal cost of Proposition 1 in parts (b) and (c) hold for a more general subgradient algorithm than the algorithm of (9). In particular, the result in part (c) is independent of the algorithm that is used to generate the multiplier sequence $\{\mu_k\}$. The proof of the result in part (c) relies on the nonexpansive property of the projection operation and the fact that the zero vector belongs to the projection set $\{\mu \in \mathbb{R}^m \mid \mu \geq 0\}$. Therefore, the results in parts (b) and (c) hold when we use a more general subgradient algorithm of the following form:

$$\mu_{k+1} = \mathcal{P}_D[\mu_k + \alpha g_k] \qquad \text{for } k \geq 1,$$

where $D \subseteq \{\mu \in \mathbb{R}^m \mid \mu \geq 0\}$ is a closed convex set containing the zero vector. We study a subgradient algorithm of this form in section 5 and establish similar error estimates.

Finally, bounds similar to those of Proposition 1 can be established for a multiplier sequence $\{\mu_k\}$ generated by a subgradient algorithm that uses a general (nonequal) stepsize sequence. In particular, given a stepsize sequence $\{\alpha_k\}$ and an initial iterate $\mu_0 \geq 0$, consider the subgradient method

$$\mu_{k+1} = [\mu_k + \alpha_k g_k]^+ \qquad \text{for } k = 0, 1, \ldots, \tag{15}$$

where g_k is given by

$$g_k = g(x_k), \qquad x_k \in \operatorname*{argmin}_{x \in X}\{f(x) + \mu_k' g(x)\} \qquad \text{for all } k \geq 0.$$

Define $\tilde{x}_k$ as the convex combination of the vectors $x_0, \ldots, x_{k-1}$ with weights $\alpha_0, \ldots \alpha_{k-1}$,

$$\tilde{x}_k = \frac{\sum_{i=0}^{k-1} \alpha_i x_i}{\sum_{i=0}^{k-1} \alpha_i}. \tag{16}$$

Following a similar analysis to that of Proposition 1, we establish the following estimates for the weighted-average vectors $\tilde{x}_k$.

PROPOSITION 2. *Let the multiplier sequence $\{\mu_k\}$ be generated by the subgradient method of* (15)*. Let the vectors $\tilde{x}_k$ for $k \geq 1$ be the averages given by* (16)*. Then, for all $k \geq 1$, the following hold:*

(a) *An upper bound on the amount of constraint violation of the vector $\tilde{x}_k$ is given by*

$$\|g(\tilde{x}_k)^+\| \leq \frac{\|\mu_k\|}{\sum_{i=0}^{k-1} \alpha_i}.$$

(b) *An upper bound on the primal cost of the vector $\tilde{x}_k$ is given by*

$$f(\tilde{x}_k) \leq q^* + \frac{\|\mu_0\|^2}{2\sum_{i=0}^{k-1} \alpha_i} + \frac{\sum_{i=0}^{k-1} \alpha_i^2 \|g(x_i)\|^2}{2\sum_{i=0}^{k-1} \alpha_i}.$$

(c) *A lower bound on the primal cost of the vector $\tilde{x}_k$ is given by*

$$f(\tilde{x}_k) \geq q^* - \|\mu^*\| \|g(\tilde{x}_k)^+\|,$$

where μ^* is a dual optimal solution.

4.2. Properties of the averaged primal sequence under Slater. Here, we strengthen the relations of Proposition 1 under the Slater condition and the boundedness of the set X. Our main result is given in the following proposition.

PROPOSITION 3. *Let the sequence $\{\mu_k\}$ be generated by the subgradient algorithm* (9)*. Let the Slater condition and the bounded set assumption hold (cf. Assumptions* 1 *and* 2*). Also, let*

$$B^* = \frac{2}{\gamma}\left(f(\bar{x}) - q^*\right) + \max\left\{\|\mu_0\|, \frac{1}{\gamma}\left(f(\bar{x}) - q^*\right) + \frac{\alpha L^2}{2\gamma} + \alpha L\right\}. \tag{17}$$

Let the vectors $\hat{x}_k$ for $k \geq 1$ be the averages given by (13)*. Then, the following hold for all $k \geq 1$:*

(a) *An upper bound on the amount of constraint violation of the vector $\hat{x}_k$ is given by*

$$\|g(\hat{x}_k)^+\| \leq \frac{B^*}{k\alpha}.$$

(b) *An upper bound on the primal cost of the vector $\hat{x}_k$ is given by*

$$f(\hat{x}_k) \leq f^* + \frac{\|\mu_0\|^2}{2k\alpha} + \frac{\alpha L^2}{2}.$$

(c) *A lower bound on the primal cost of the vector $\hat{x}_k$ is given by*

$$f(\hat{x}_k) \geq f^* - \frac{1}{\gamma}\,[\,f(\bar{x}) - q^*\,]\,\|g(\hat{x}_k)^+\|.$$

Proof. (a) Under the assumptions, by Lemma 3 we have

$$\|\mu_k\| \leq \frac{2}{\gamma}\,(f(\bar{x}) - q^*) + \max\left\{\|\mu_0\|, \frac{1}{\gamma}\,(f(\bar{x}) - q^*) + \frac{\alpha L^2}{2\gamma} + \alpha L\right\} \qquad \text{for all } k \geq 0.$$

By the definition of B^* in (17), the preceding relation is equivalent to

$$\|\mu_k\| \leq B^* \qquad \text{for all } k \geq 0. \tag{18}$$

By using Proposition 1(a), we obtain

$$\|g(\hat{x}_k)^+\| \leq \frac{\|\mu_k\|}{k\alpha} \leq \frac{B^*}{k\alpha} \qquad \text{for all } k \geq 1.$$

(b) From Proposition 1(b), we obtain

$$f(\hat{x}_k) \leq q^* + \frac{\|\mu_0\|^2}{2k\alpha} + \frac{\alpha}{2k}\sum_{i=0}^{k-1}\|g(x_i)\|^2 \qquad \text{for all } k \geq 1.$$

Under the Slater condition, there is zero duality gap, i.e., $q^* = f^*$. Furthermore, the subgradients are bounded by a scalar L (cf. Assumption 2), so that

$$f(\hat{x}_k) \leq f^* + \frac{\|\mu_0\|^2}{2k\alpha} + \frac{\alpha L^2}{2} \qquad \text{for all } k \geq 1.$$

(c) Under the Slater condition, a dual optimal solution exists and there is zero duality gap, i.e., $q^* = f^*$. Thus, by Proposition 1(c), for any dual solution μ^* we have

$$f(\hat{x}_k) \geq f^* - \|\mu^*\|\|g(\hat{x}_k)^+\| \qquad \text{for all } k \geq 1.$$

By using Lemma 1 with $\bar{\mu} = \mu^*$, we see that the dual set is bounded and, in particular, $\|\mu^*\| \leq \frac{1}{\gamma}(f(\bar{x}) - q^*)$ for all dual optimal vectors μ^*. Hence,

$$f(\hat{x}_k) \geq f^* - \frac{1}{\gamma}\,[\,f(\bar{x}) - q^*\,]\,\|g(\hat{x}_k)^+\| \qquad \text{for all } k \geq 1. \qquad \square$$

It seems reasonable to choose the initial iterate as $\mu_0 = 0$, as suggested by the upper bound for $f(\hat{x}_k)$ in part (b) of Proposition 3. In this case, the bound B^* in part (a) of Proposition 3, with $q^* = f^*$, reduces to

$$B^* = \frac{3}{\gamma}\,[f(\bar{x}) - f^*] + \frac{\alpha L^2}{2\gamma} + \alpha L \qquad \text{for } k \geq 1, \tag{19}$$

while the estimate in part (b) reduces to

$$f(\hat{x}_k) \leq f^* + \frac{\alpha L^2}{2} \qquad \text{for all } k \geq 1. \tag{20}$$

Using the preceding two relations, one can estimate the order of the number of iterations needed to achieve an ϵ-feasible and ϵ-optimal solution.[3] In particular, to achieve the ϵ-optimality, from (20) the stepsize should satisfy $\alpha \leq 2\epsilon/L^2$. Assuming for the sake of simplicity that the term $\frac{1}{\gamma}\,[f(\bar{x}) - f^*]$ is dominant in (19), to obtain ϵ-feasibility, the number of iterations k should satisfy $k \geq \frac{[f(\bar{x})-f^*]\,L^2}{c\gamma\epsilon^2}$, where c is some constant independent of ϵ. Thus, to achieve ϵ-feasible and ϵ-optimal solution, the number of iterations is of the order $1/\epsilon^2$, which is typical for subgradient methods.

The results can alternatively be interpreted for a fixed stepsize value α. In this case, by Proposition 3(a), the amount of feasibility violation $\|g(\hat{x}_k)^+\|$ of the vector $\hat{x}_k$ diminishes to zero as the number of subgradient iterations k increases. By combining the results in (a)–(c), we see that the limits of the function values $f(\hat{x}_k)$, as $k \to \infty$, are within the range $\big[f^*, f^* + \alpha L^2/2\big]$.

Finally, we note that a more practical bound than the bound B^* in Proposition 3 can be obtained by using $\max_{0\leq i\leq k} q(\mu_i)$ as an approximation of the optimal value $f^* = q^*$.

5. Modified subgradient method under Slater. In this section, we consider a modified version of the subgradient method under the Slater assumption. The motivation is coming from the fact that under the Slater assumption, the set of dual optimal solutions is bounded (cf. Lemma 1). Therefore, it is of interest to consider a subgradient method in which dual iterates are projected onto a bounded superset of the dual optimal solution set. We consider such algorithms and generate primal solutions using averaging as described in section 4. Also, we provide estimates for the amount of constraint violation and cost of the average primal sequence. Our goal is to compare these estimates with the error estimates obtained for the "ordinary" subgradient method in section 4.

Formally, we consider subgradient methods of the following form:

$$\mu_{k+1} = \mathcal{P}_D\,[\mu_k + \alpha g_k], \tag{21}$$

where the set D is a compact convex set containing the set of dual optimal solutions (to be discussed shortly) and $\mathcal{P}_D$ denotes the projection on the set D. The vector $\mu_0 \in D$ is an arbitrary initial iterate and the scalar $\alpha > 0$ is a constant stepsize. The vector g_k is a subgradient of q at μ_k given by

$$g_k = g(x_k), \qquad x_k \in \operatorname*{argmin}_{x\in X}\{f(x) + \mu_k' g(x)\} \qquad \text{for all } k \geq 0$$

(see (5)).

Under the Slater condition, the dual optimal set M^* is nonempty and bounded, and a bound on the norms of the dual optimal solutions is given by

$$\sum_{j=1}^{m} \mu_j^* \leq \frac{1}{\gamma}(f(\bar{x}) - q^*) \qquad \text{for all } \mu^* \in M^*,$$

[3]We thank an anonymous referee for this insight.

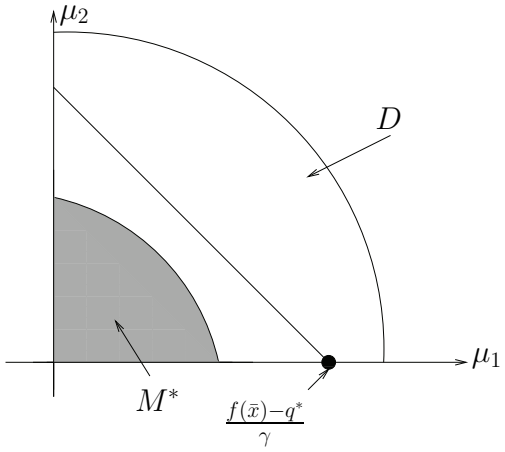


FIG. 1. *The dual optimal set M^* and the set D, which is considered in the modified subgradient method.*

with $\gamma = \min_{1\le j\le m}\{-g_j(\bar{x})\}$ (cf. Lemma 1). Thus, having the dual value $\tilde{q} = q(\tilde{\mu})$ for some $\tilde{\mu} \ge 0$, since $q^* \ge \tilde{q}$, we obtain

$$\sum_{j=1}^{m} \mu_j^* \le \frac{1}{\gamma}(f(\bar{x}) - \tilde{q}) \qquad \text{for all } \mu^* \in M^*. \tag{22}$$

This motivates the following choice for the set D:

$$D = \left\{ \mu \ge 0 \;\middle|\; \|\mu\| \le \frac{f(\bar{x}) - \tilde{q}}{\gamma} + r \right\}, \tag{23}$$

with a scalar $r > 0$. Clearly, the set D is compact and convex, and it contains the set of dual optimal solutions in view of relation (22) and the fact $\|y\| \le \|y\|_1$ for any vector y (the illustration of the set D is provided in Figure 1).

Similar to section 4, we provide near-feasible and near-optimal primal vectors by averaging the vectors from the sequence $\{x_k\}$. In particular, we define $\hat{x}_k$ as the average of the vectors $x_0, \ldots, x_{k-1}$, i.e.,

$$\hat{x}_k = \frac{1}{k}\sum_{i=0}^{k-1} x_i \qquad \text{for all } k \ge 1. \tag{24}$$

In the next proposition, we provide per-iterate bounds for the constraint violation and primal cost values of the average vectors $\hat{x}_k$.

PROPOSITION 4. *Let the Slater condition and the bounded set assumption hold (cf. Assumptions* 1 *and* 2*). Let the dual sequence $\{\mu_k\}$ be generated by the modified subgradient method of* (21)*. Let $\{\hat{x}_k\}$ be the average sequence defined in* (24)*. Then, for all $k \ge 1$, we have the following:*

(a) *An upper bound on the amount of constraint violation of the vector $\hat{x}_k$ is given by*

$$\|g(\hat{x}_k)^+\| \le \frac{2}{k\alpha r}\left(\frac{f(\bar{x}) - \tilde{q}}{\gamma} + r\right)^2 + \frac{\alpha L^2}{2r}.$$

(b) *An upper bound on the primal cost of the vector $\hat{x}_k$ is given by*

$$f(\hat{x}_k) \leq f^* + \frac{\|\mu_0\|^2}{2k\alpha} + \frac{\alpha L^2}{2}.$$

(c) *A lower bound on the primal cost of the vector $\hat{x}_k$ is given by*

$$f(\hat{x}_k) \geq f^* - \left(\frac{f(\bar{x}) - \tilde{q}}{\gamma}\right) \|g(\hat{x}_k)^+\|.$$

Here, the scalars $r > 0$ and $\tilde{q}$ with $\tilde{q} \leq q^*$ are those from the definition of the set D in (23).

Proof. (a) Using the definition of the iterate μ_{k+1} in (21) and the nonexpansive property of projection on a closed convex set, we obtain, for all $\mu \in D$ and all $i \geq 0$,

$$\begin{aligned}\|\mu_{i+1} - \mu\|^2 &= \|\mathcal{P}_D\,[\mu_i + \alpha g_i] - \mu\|^2 \\ &\leq \|\mu_i + \alpha g_i - \mu\|^2 \\ &\leq \|\mu_i - \mu\|^2 + 2\alpha g_i'(\mu_i - \mu) + \alpha^2\|g_i\|^2 \\ &\leq \|\mu_i - \mu\|^2 + 2\alpha g_i'(\mu_i - \mu) + \alpha^2 L^2.\end{aligned}$$

Therefore, for any $\mu \in D$,

$$g_i'(\mu - \mu_i) \leq \frac{\|\mu_i - \mu\|^2 - \|\mu_{i+1} - \mu\|^2}{2\alpha} + \frac{\alpha L^2}{2} \qquad \text{for all } i \geq 0. \tag{25}$$

Since g_i is a subgradient of the dual function q at μ_i, using the subgradient inequality (cf. (1)), we obtain, for any dual optimal solution μ^*,

$$g_i'(\mu_i - \mu^*) \leq q(\mu_i) - q(\mu^*) \leq 0 \qquad \text{for all } i \geq 0,$$

where the last inequality follows from the optimality of μ^* and the feasibility of each $\mu_i \in D$ (i.e., $\mu_i \geq 0$). We then have, for all $\mu \in D$ and all $i \geq 0$,

$$g_i'(\mu - \mu^*) = g_i'(\mu - \mu^* - \mu_i + \mu_i) = g_i'(\mu - \mu_i) + g_i'(\mu_i - \mu^*) \leq g_i'(\mu - \mu_i).$$

From the preceding relation and (25), we obtain, for any $\mu \in D$,

$$g_i'(\mu - \mu^*) \leq \frac{\|\mu_i - \mu\|^2 - \|\mu_{i+1} - \mu\|^2}{2\alpha} + \frac{\alpha L^2}{2} \qquad \text{for all } i \geq 0.$$

Summing over $i = 0, \ldots, k-1$ for $k \geq 1$, we obtain, for any $\mu \in D$ and $k \geq 1$,

$$\sum_{i=0}^{k-1} g_i'(\mu - \mu^*) \leq \frac{\|\mu_0 - \mu\|^2 - \|\mu_k - \mu\|^2}{2\alpha} + \frac{\alpha k L^2}{2} \leq \frac{\|\mu_0 - \mu\|^2}{2\alpha} + \frac{\alpha k L^2}{2}.$$

Therefore, for any $k \geq 1$,

$$\max_{\mu \in D} \left\{\sum_{i=0}^{k-1} g_i'(\mu - \mu^*)\right\} \leq \frac{1}{2\alpha} \max_{\mu \in D} \|\mu_0 - \mu\|^2 + \frac{\alpha k L^2}{2}. \tag{26}$$

We now provide a lower estimate on the left-hand side of the preceding relation. Let $k \geq 1$ be arbitrary and, for simplicity, we suppress the explicit dependence on k by letting

$$s = \sum_{i=0}^{k-1} g_i. \tag{27}$$

Let s^+ be the componentwise maximum of s and the zero vector; i.e., the jth entry of the vector s^+ is given by $s_j^+ = \max\{s_j, 0\}$. If $s^+ = 0$, then the bound in part (a) of this proposition trivially holds. Thus, assume that $s^+ \neq 0$ and define a vector $\bar{\mu}$ as follows:

$$\bar{\mu} = \mu^* + r\,\frac{s^+}{\|s^+\|}.$$

Note that $\bar{\mu} \geq 0$ since $\mu^* \geq 0$, $s^+ \geq 0$, and $r > 0$. By Lemma 1, the dual optimal solution set is bounded and, in particular, $\|\mu^*\| \leq \frac{f(\bar{x})-q^*}{\gamma}$. Furthermore, since $\tilde{q} \leq q^*$, it follows that $\|\mu^*\| \leq \frac{f(\bar{x})-\tilde{q}}{\gamma}$ for any dual solution μ^*. Therefore, by the definition of the vector $\bar{\mu}$, we have

$$\|\bar{\mu}\| \leq \|\mu^*\| + r \leq \frac{f(\bar{x}) - \tilde{q}}{\gamma} + r, \tag{28}$$

implying that $\bar{\mu} \in D$. Using the definition of the vector s in (27) and relation (26), we obtain

$$s'(\bar{\mu} - \mu^*) = \sum_{i=0}^{k-1} g_i'(\bar{\mu} - \mu^*) \leq \max_{\mu \in D} \left\{ \sum_{i=0}^{k-1} g_i'(\mu - \mu^*) \right\} \leq \frac{1}{2\alpha} \max_{\mu \in D} \|\mu_0 - \mu\|^2 + \frac{\alpha k L^2}{2}.$$

Since $\bar{\mu} - \mu^* = r\frac{s^+}{\|s^+\|}$, we have $s'(\bar{\mu} - \mu^*) = r\|s^+\|$. Thus, by the definition of s in (27) and the fact $g_i = g(x_i)$, we have

$$s'(\bar{\mu} - \mu^*) = r \left\| \left[\sum_{i=0}^{k-1} g(x_i) \right]^+ \right\|.$$

Combining the preceding two relations, it follows that

$$\left\| \left[\sum_{i=0}^{k-1} g(x_i) \right]^+ \right\| \leq \frac{1}{2\alpha r} \max_{\mu \in D} \|\mu_0 - \mu\|^2 + \frac{\alpha k L^2}{2r}.$$

Dividing both sides of this relation by k, and using the convexity of the functions g_j in $g = (g_1, \ldots, g_m)$ and the definition of the average primal vector $\hat{x}_k$, we obtain

$$\|g(\hat{x}_k)^+\| \leq \frac{1}{k} \left\| \left[\sum_{i=0}^{k-1} g(x_i) \right]^+ \right\| \leq \frac{1}{2k\alpha r} \max_{\mu \in D} \|\mu_0 - \mu\|^2 + \frac{\alpha L^2}{2r}. \tag{29}$$

Since $\mu_0 \in D$, we have

$$\max_{\mu \in D} \|\mu_0 - \mu\|^2 \leq \max_{\mu \in D} (\|\mu_0\| + \|\mu\|)^2 \leq 4 \max_{\mu \in D} \|\mu\|^2.$$

By using the definition of the set D of (23), we have

$$\max_{\mu\in D}\|\mu\| \leq \frac{f(\bar{x})-\tilde{q}}{\gamma}+r.$$

By substituting the preceding two estimates in relation (29), we obtain

$$\|g(\hat{x}_k)^+\| \leq \frac{2}{k\alpha r}\left(\frac{f(\bar{x})-\tilde{q}}{\gamma}+r\right)^2+\frac{\alpha L^2}{2r}.$$

(b) The proof follows from an identical argument to that used in the proof of Proposition 1(b), and therefore is omitted.

(c) The result follows from an identical argument to that used in the proof of Proposition 1(c) and the bound on the dual optimal solution set that follows in view of the Slater condition (cf. Assumption 1 and Lemma 1 with $\bar{\mu}=\mu^*$). □

We note here that the subgradient method of (21) with the set D given in (23) couples the computation of multipliers. In some applications, it might be desirable to accommodate distributed computation models whereby the multiplier components μ_j^* are processed in a distributed manner among a set of processors or agents. To accommodate such computations, one may modify the subgradient method of (21) by replacing the set D of (23) with the following set:

$$D_\infty=\left\{\mu\geq 0 \;\middle|\; \|\mu\|_\infty \leq \frac{f(\bar{x})-\tilde{q}}{\gamma}+r\right\}.$$

It can be seen that the results of Proposition 4 also hold for this choice of the projection set. In particular, this can be seen by following the same line of argument as in the proof of Proposition 4 and by using the following relation:

$$\|\bar{\mu}\|_\infty \leq \|\mu^*\|+r \leq \frac{f(\bar{x})-\tilde{q}}{\gamma}+r$$

(cf. (28) and the fact $\|y\|_\infty \leq \|y\|$ for any vector y).

We next consider selecting the parameter r, which is used in the definition of the set D, such that the right-hand side of the bound in part (a) of Proposition 4 is minimized at each iteration k. Given some $k\geq 1$, we choose r as the optimal solution of the problem

$$\min_{r>0}\left\{\frac{2}{k\alpha r}\left(\frac{f(\bar{x})-\tilde{q}}{\gamma}+r\right)^2+\frac{\alpha L^2}{2r}\right\}.$$

It can be seen that the optimal solution of the preceding problem, denoted by $r^*(k)$, is given by

$$r^*(k)=\sqrt{\left(\frac{f(\bar{x})-\tilde{q}}{\gamma}\right)^2+\frac{\alpha^2L^2k}{4}} \qquad \text{for } k\geq 1. \tag{30}$$

Consider now an algorithm where the dual iterates are obtained by

$$\mu_{i+1}=\mathcal{P}_{D_k}\left[\mu_i+\alpha g_k\right] \qquad \text{for each } i\geq 0,$$

with $\mu_0 \in D_0$ and the set D_k given by

$$D_k = \left\{ \mu \geq 0 \mid \|\mu\| \leq \frac{f(\bar{x}) - \tilde{q}}{\gamma} + r^*(k) \right\},$$

where $r^*(k)$ is given by (30). Hence, at each iteration i, the algorithm projects onto the set D_k, which contains the set of dual optimal solutions M^*.

Substituting $r^*(k)$ in the bound of Proposition 4(a), we can see that

$$\begin{aligned} \|g(\hat{x}_k)^+\| &\leq \frac{4}{k\alpha} \left(\frac{f(\bar{x}) - \tilde{q}}{\gamma} + \sqrt{\left(\frac{f(\bar{x}) - \tilde{q}}{\gamma} \right)^2 + \frac{\alpha^2 L^2 k}{4}} \right) \\ &\leq \frac{4}{k\alpha} \left(\frac{2(f(\bar{x}) - \tilde{q})}{\gamma} + \frac{\alpha L \sqrt{k}}{2} \right) \\ &= \frac{8}{k\alpha} \left(\frac{f(\bar{x}) - \tilde{q}}{\gamma} \right) + \frac{2L}{\sqrt{k}}. \end{aligned}$$

The preceding discussion combined with Proposition 4(a) immediately yields the following result.

PROPOSITION 5. *Let the Slater condition and the bounded set assumption hold (cf. Assumptions* 1 *and* 2*). Given some $k \geq 1$, define the set D_k as*

$$D_k = \left\{ \mu \geq 0 \;\middle|\; \|\mu\|_2 \leq \frac{f(\bar{x}) - \tilde{q}}{\gamma} + \sqrt{\left(\frac{f(\bar{x}) - \tilde{q}}{\gamma} \right)^2 + \frac{\alpha^2 L^2 k}{4}} \right\}. \tag{31}$$

Let the dual sequence $\{\mu_i\}$ be generated by the following modified subgradient method: let $\mu_0 \in D_k$, and for each $i \geq 0$ the dual iterate μ_i is obtained by

$$\mu_{i+1} = \mathcal{P}_{D_k} \left[\mu_i + \alpha g_i \right].$$

Then, an upper bound on the amount of feasibility violation of the vector $\hat{x}_k$ is given by

$$\|g(\hat{x}_k)^+\| \leq \frac{8}{k\alpha} \left(\frac{f(\bar{x}) - \tilde{q}}{\gamma} \right) + \frac{2L}{\sqrt{k}}. \tag{32}$$

This result shows that at a given k, the error estimate provided in (32) can be achieved if we use a modified subgradient method where each dual iterate is projected on the set D_k defined in (31). Given a prespecified accuracy for the amount of feasibility violation, this bound can be used to select the stepsize value and the set D_k. Furthermore, using the estimate (32) in Proposition 4(c), we can obtain a lower bound on the cost $f(\hat{x}_k)$.

We now compare the feasibility violation bound of Proposition 3(a) for the ordinary subgradient method with the result of Proposition 5 for the modified subgradient method. As we will see, depending on the values of γ, L, and the estimate $\tilde{q}$, each of these bounds can be better or worse than the other one. For the sake of comparison, let us assume that $\tilde{q}$ is the same in both bounds. Then, by Proposition 5(a) and by using $q^* \geq \tilde{q}$ in the definition of the bound B^* in (17), we obtain, for all $k \geq 1$,

$$\|g(\hat{x}_k)^+\| \leq \frac{2}{k\alpha\gamma} \left(f(\bar{x}) - \tilde{q} \right) + \frac{1}{k\alpha} \max \left\{ \|\mu_0\|, \frac{1}{\gamma} \left(f(\bar{x}) - \tilde{q} \right) + \frac{\alpha L^2}{2\gamma} + \alpha L \right\}. \tag{33}$$

This bound, as compared to that of Proposition 5, is better when k is very large, since the feasibility violation decreases in the order of $1/k$, while in Proposition 5, it decreases in the order of $1/\sqrt{k}$. However, initially, the feasibility violation in (33) for the ordinary subgradient method can be worse because it depends on the initial iterate μ_0. Suppose that the initial iterate is $\mu_0 = 0$. Then, the bound in (33) reduces to

$$\|g(\hat{x}_k)^+\| \leq \frac{3}{k\alpha\gamma}\,[\,f(\bar{x}) - \tilde{q}\,] + \frac{1}{k}\left(\frac{L^2}{2\gamma} + L\right).$$

Even in this case, initially, this bound can be worse than that of Proposition 5 because the bound depends inversely on γ which can be very small (recall $\gamma = \min_{1\leq j\leq m}\{-g_j(\bar{x})\}$ with a Slater vector $\bar{x}$). To complement our theoretical analysis, we need to perform some numerical experiments to further study and compare these algorithms.

6. Conclusions. In this paper, we have studied the application of dual subgradient algorithms for generating primal near-feasible and near-optimal optimal solutions. We have proposed and analyzed two such algorithms under Slater condition. Both of the proposed algorithms use projections to generate a dual sequence and an averaging scheme to produce approximate primal vectors. The algorithms employ the same averaging scheme in the primal space. However, they operate on different sets when projecting in the dual space. One algorithm uses the projections on the nonnegative orthant, while the other algorithm uses the projections on nested compact convex sets that change with each iteration but always contain the dual optimal solutions. Nevertheless, both algorithms produce primal vectors whose infeasibility diminishes to zero and function values in the limit stay within the interval $[f^*, f^* + \alpha L/2]$.

Let us note that, in general, one may consider solving the dual problem to generate a good approximate "dual solution" (with a method more efficient than the subgradient method with a constant step), and then proceed with our algorithm with averaging. This can be advantageous since $x(\mu_k)$ for μ_k far from an optimum can be less informative than those closer to an optimum.[4]

Our comparison of the two methods is purely based on our *theoretical analysis*, which need not reflect the real behavior of these algorithms for practical implementations. Our future goal is to numerically test and evaluate these algorithms in order to gain deeper insights into their behavior.

Acknowledgments. We thank Robert Freund and Pablo Parrilo for useful comments and discussions. Also, we thank Vladimir Norkin and the two anonymous referees for valuable comments and suggestions.

REFERENCES

[1] F. Barahona and R. Anbil, *The volume algorithm: Producing primal solutions with a subgradient method*, Math. Program., 87 (2000), pp. 385–399.
[2] A. Ben-Tal, T. Margalit, and A. Nemirovski, *The ordered subsets mirror descent optimization method and its use for the positron emission tomography reconstruction problem*, SIAM J. Optim., 12 (2001), pp. 79–108.
[3] A. Ben-Tal and A. Nemirovski, *Non-Euclidean restricted memory level method for large-scale convex optimization*, Math. Program., 102 (2005), pp. 407–456.
[4] D. Bertsekas, *Nonlinear Programming*, 2nd ed., Athena Scientific, Cambridge, MA, 1999.
[5] D. Bertsekas, A. Nedić, and A. Ozdaglar, *Convex Analysis and Optimization*, Athena Scientific, Cambridge, MA, 2003.

[4]We thank an anonymous referee for bringing this to our attention.

[6] M. Chiang, S. Low, A. Calderbank, and J. Doyle, *Layering as optimization decomposition: A mathematical theory of network architectures*, Proceedings of the IEEE, 95 (2007), pp. 255–312.
[7] R. Correa and C. Lemaréchal, *Convergence of some algorithms for convex minimization*, Math. Programming, 62 (1993), pp. 261–275.
[8] V. Demyanov and L. Vasilyev, *Nondifferentiable Optimization*, Optimization Software, Inc., New York, 1985.
[9] Y. Ermoliev, *Methods for solving nonlinear extremal problems*, Kibernetika, 4 (1966), pp. 1–17.
[10] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms, Vol.* I. *Fundamentals*, Springer-Verlag, Berlin, 1993.
[11] F. Kelly, A. Maulloo, and D. Tan, *Rate control in communication networks: Shadow prices, proportional fairness, and stability*, Journal of the Operational Research Society, 49 (1998), pp. 237–252.
[12] K. Kiwiel, T. Larsson, and P. Lindberg, *Lagrangian relaxation via ballstep subgradient methods*, Math. Oper. Res., 32 (2007), pp. 669–686.
[13] K. C. Kiwiel and P. O. Lindberg, *Parallel subgradient methods for convex optimization*, in Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications, Stud. Comput. Math. 8, D. Butnariu, Y. Censor, and S. Reich, eds., North–Holland, Amsterdam, 2001, pp. 335–344.
[14] T. Larsson and Z. Liu, *A Lagrangian relaxation scheme for structured linear programs with application to multicommodity network flows*, Optimization, 40 (1997), pp. 247–284.
[15] T. Larsson, M. Patriksson, and A. Strömberg, *Ergodic results and bounds on the optimal value in subgradient optimization*, in Operations Research Proceedings, P. Kleinschmidt et al., eds., Springer-Verlag, Berlin, 1995, pp. 30–35.
[16] T. Larsson, M. Patriksson, and A.-B. Strömberg, *Ergodic convergence in subgradient optimization*, Optim. Methods Softw., 9 (1998), pp. 93–120.
[17] T. Larsson, M. Patriksson, and A.-B. Strömberg, *Ergodic primal convergence in dual subgradient schemes for convex programming*, Math. Program., 86 (1999), pp. 283–312.
[18] S. Low and D. Lapsley, *Optimization flow control,* I: *Basic algorithm and convergence*, IEEE/ACM Transactions on Networking, 7 (1999), pp. 861–874.
[19] A. Mutapcic, S. Boyd, S. Murali, D. Atienza, G. D. Micheli, and R. Gupta, *Processor speed control with thermal constraints*, submitted.
[20] A. Nedić and D. Bertsekas, *Convergence rate of incremental subgradient algorithms*, in Stochastic Optimization: Algorithms and Applications, Appl. Optim. 54, S. Uryasev and P. Pardalos, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001, pp. 223–264.
[21] A. Nedić and D. Bertsekas, *Incremental subgradient methods for nondifferentiable optimization*, SIAM J. Optim., 12 (2001), pp. 109–138.
[22] A. Nedić, D. Bertsekas, and V. Borkar, *Distributed asynchronous incremental subgradient methods*, in Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications, Stud. Comput. Math. 8, D. Butnariu, Y. Censor, and S. Reich, eds., North–Holland, Amsterdam, 2001, pp. 381–407.
[23] A. S. Nemirovskii and D. B. Yudin, *Cezare convergence of gradient method approximation of saddle points for convex-concave functions*, Dokl. Akad. Nauk SSSR, 239 (1978), pp. 1056–1059.
[24] A. S. Nemirovskii and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*, John Wiley and Sons, London, 1983.
[25] Y. Nesterov, *Primal-Dual Subgradient Methods for Convex Problems*, Technical Report 67, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), Louvain, Belgium, 2005.
[26] B. Polyak, *A general method for solving extremum problems*, Soviet Math. Dokl., 8 (1967), pp. 593–597.
[27] B. Polyak, *Introduction to Optimization*, Optimization Software, Inc., New York, 1987.
[28] A. Ruszczyński, *A merit function approach to the subgradient method with averaging*, Optim. Methods Softw., 23 (2008), pp. 161–172.
[29] S. Sen and H. D. Sherali, *A class of convergent primal-dual subgradient algorithms for decomposable convex programs*, Math. Programming, 35 (1986), pp. 279–297.
[30] S. Shakkottai and R. Srikant, *Network optimization and control*, Foundations and Trends in Networking, 2 (2007), pp. 271–379.
[31] H. D. Sherali and G. Choi, *Recovery of primal solutions when using subgradient optimization methods to solve Lagrangian duals of linear programs*, Oper. Res. Lett., 19 (1996), pp. 105–113.

[32] N. Shor, *Minimization Methods for Nondifferentiable Functions*, translated from the Russian by K. C. Kiwiel and A. Ruszczynski, Springer-Verlag, Berlin, 1985.
[33] R. Srikant, *Mathematics of Internet Congestion Control*, Birkhäuser Boston, Boston, 2004.
[34] H. Uzawa, *Iterative methods in concave programming*, in Studies in Linear and Nonlinear Programming, K. Arrow, L. Hurwicz, and H. Uzawa, eds., Stanford University Press, Palo Alto, CA, 1958, pp. 154–165.
[35] X. Zhao, P. B. Luh, and J. Wang, *Surrogate gradient algorithm for Lagrangian relaxation*, J. Optim. Theory Appl., 100 (1999), pp. 699–712.
[36] N. G. Zhurbenko, E. G. Pinaev, N. Z. Shor, and G. N. Yun, *Choice of fleet composition and allocation of aircraft to civil airline routes*, Cybernetics and Systems Analysis, 12 (1976), pp. 636–640 (in English); Kibernetika, 4 (1976), pp. 138–141 (in Russian).

SIAM J. OPTIM.
Vol. 19, No. 4, pp. 1781–1806

ON MULTIVARIATE DISCRETE MOMENT PROBLEMS: GENERALIZATION OF THE BIVARIATE MIN ALGORITHM FOR HIGHER DIMENSIONS*

GERGELY MÁDI-NAGY†

Abstract. The objective of the multivariate discrete moment problem (MDMP) is to find the minimum and/or maximum of the expected value of a function of a random vector with a discrete finite support where the probability distribution is unknown, but some of the moments are given. The moments may be binomial, power, or of a more general type. The MDMP can be formulated as a linear programming problem with a very ill-conditioned coefficient matrix. Hence, the LP problem can be solved with difficulty or cannot be solved at all. The central results of the field of the MDMP concern the structure of the dual feasible bases. These bases, on one hand, provide us with bounds without any numerical difficulties. On the other hand, they can be used as an initial basis of the dual simplex method. That results in shorter running time and better numerical stability because the first phase can be skipped. This paper introduces a new type of MDMP, where the bivariate moments up to a certain order m consisting of the first variable and further univariate moments up to the order m_j, $j = 1, \dots, s$, are given. Then we generalize the bivariate Min Algorithm of Mádi-Nagy and Prékopa [*Math. Oper. Res.*, 29 (2004), pp. 229–258] for higher dimensions, which gives numerous dual feasible bases of the MDMP. By the aid of this, on one hand, we can give useful bounds for MDMPs with higher dimensional random vectors even if the usual solvers cannot give acceptable results. On the other hand, applying our algorithm for the binomial MDMP, we can give better bounds for probabilities of Boolean functions of event sequences than the recent bounds in the literature. These results are illustrated by numerical examples.

Key words. discrete moment problem, multivariate Lagrange interpolation, linear programming, expectation bounds, probability bounds

AMS subject classifications. 62H99, 90C05, 65D05

DOI. 10.1137/070705878

1. Introduction.

1.1. Moment problems. The classical moment problem was introduced in 1894–1895 by Stieltjes [38] (see Kjeldsen [18]). The problem is about trying to invert the mapping that takes a measure P on $I \subset \mathbb{R}$ to the sequences of moments μ_k, $k = 1, 2, 3, \dots$, where the kth, so-called power, moments are defined as

$$\mu_k = \int_I z^k dP.$$

In what follows, we will use the notation μ_0 for $k = 0$, but we know $\mu_0 = 1$. More general moments than the power moments can be defined if we consider the sequence of functions $u_k(z)$, $k = 0, 1, 2, \dots$, which are assumed to be measurable and integrable functions on I. Then the generalized moment sequence with respect to $\{u_k(z)\}$ is

$$\int_I u_k(z) dP, \qquad k = 0, 1, \dots.$$

*Received by the editors October 17, 2007; accepted for publication (in revised form) October 12, 2008; published electronically February 11, 2009. This research was partially supported by OTKA grants F-046309 and T-047340 in Hungary.
http://www.siam.org/journals/siopt/19-4/70587.html
†Mathematical Institute, Budapest University of Technology and Economics, Műegyetem rakpart 1-3, Budapest 1111, Hungary (gnagy@math.bme.hu).

Sometimes the moment problem is based on these general moments, but typically power moments are used. The questions of the classical moment problem can be the following. Given a sequence of numbers μ_k, $k = 1, 2, 3, \ldots$,

(a) does there exist a probability measure on I with the moments μ_k, $k = 1, 2, 3, \ldots$?
(b) is this probability measure uniquely determined by the moments μ_k, $k = 1, 2, 3, \ldots$?
(c) how can one describe all probability measures on I with the moments μ_k, $k = 1, 2, 3, \ldots$?

Depending on the set I, classical results can be found in Hamburger [12], [13], Hausdorff [14], [15], and Riesz [36]. Akhiezer [1] gives a comprehensive survey of that field.

The truncated variation of the classical moment problem studies the properties of measures with fixed first k moments (for a finite k) or with finite set of moments. Beside the classical questions (existence, uniqueness of the measure P) the following bounding problem arises. Given the moments μ_k, $k = 1, \ldots, m$, or any finite collection of the moments, what are the possible lower and upper bounds for

$$\int_I f(z)dP,$$

where f is a given real function on I and P is unknown, but P has to have the given moments? Let X be a random variable with support I. Assume that the expected values of X^k, $k = 1, \ldots, m$, are given finite values, but the distribution of X is not known. Then the *bounding moment problem* can be formulated as

$$\begin{aligned} &\inf(\sup)\ E[f(X)] = \int_I f(z)dP \\ &\text{subject to} \\ &E[X^k] = \int_I z^k dP = \mu_k, \quad k = 0, 1, \ldots, m, \end{aligned} \tag{1.1}$$

where the probability measure P is unknown, and I, f, μ_k, $k = 0, 1, \ldots, m$, are given. Bounding problems related to moments were first considered by Bienaymé [2], Chebyshev [6], [7], and Markov [23]. The bounding moment problem frequently appears in the literature as Chebyshev-type inequalities. A good summary of these results can be found in Krein and Nudelman [19]. For historical background, see Kjeldsen [18] and Prékopa and Alexe [33].

At the end of the 1980s, Prékopa [25], [26], [27] and Samuels and Studden [37] independently introduced and studied the *univariate discrete moment problem (DMP)*, where $I = \{z_0, z_1, \ldots, z_n\}$ is a discrete finite set. Samuels and Studden use the classical approach and determine the solutions in closed form whenever possible; however, their method is applicable only to small size problems. Prékopa invented a linear programming methodology, presented briefly below. It turns out that in the special case of the DMP, linear programming techniques provide us with more general and simpler algorithmic solutions than the classical ones. Moreover, it allows for an efficient solution of large size moment problems as well as for finding closed form sharp bounds.

The DMP has (at least) three useful properties. The first one is that it uses the discrete property of the support I beside the moment information. Hence, it can give tighter bounds than the classical moment problems. The second one is that the

DMP can be formulated as a linear programming (LP) problem. Let us designate the (unknown) probability distribution of X in the following way:

$$p_i = P(X = z_i), \qquad i = 0, 1, \ldots, n,$$

and let $f(z_i) = f_i$. Then

$$E[f(X)] = f_0 p_0 + f_1 p_1 + \cdots + f_n p_n$$

and

$$E[X^k] = z_0^k p_0 + z_1^k p_1 + \cdots + z_n^k p_n, \qquad k = 0, 1, \ldots, m.$$

The LP corresponding to the DMP is

$$\begin{array}{lrrrcl} \min(\max) & f_0 p_0 & +f_1 p_1 & +\cdots & +f_n p_n & \\ \text{subject to} & & & & & \\ & p_0 & +p_1 & +\cdots & +p_n & = \mu_0 (= 1), \\ & z_0 p_0 & +z_1 p_1 & +\cdots & +z_n p_n & = \mu_1, \\ & z_0^2 p_0 & +z_1^2 p_1 & +\cdots & +z_n^2 p_n & = \mu_2, \\ & & & & & \vdots \\ & z_0^m p_0 & +z_1^m p_1 & +\cdots & +z_n^m p_n & = \mu_m, \\ & p_0, & p_1, & \ldots & p_n & \geq 0, \end{array} \tag{1.2}$$

where the unknown variables are p_i, $i = 0, 1, \ldots, n$. The support $I = \{z_0, z_1, \ldots, z_n\}$, the values of the function $f(z)$, $z \in I$, and the moments μ_k, $k = 0, 1, \ldots, m$, are given. Since the coefficient matrix of (1.2) is an ill-conditioned Vandermonde matrix, the problem usually cannot be solved by general solution methods. However, under some conditions of the function f, Prékopa [27] developed a numerically stable dual method for the solution. This method is based on theorems which give the subscript structures of columns of all dual feasible bases. By the aid of the known dual feasible bases, closed form bounds in terms of the moments can also be given; see Boros and Prékopa [3].

The third useful property is that the optimal solution of the so-called binomial moment problem can give sharp Bonferroni-type bounds and other probability bounds, as well. The kth binomial moment of a random variable X with the support $I \subset \mathbb{N}$ is defined as

$$E\left[\binom{X}{k}\right].$$

Let $A_1, A_2, \ldots, A_n$ be arbitrary events. Let the random variable X with the support $\{0, 1, \ldots, n\}$ be the number of those events which occur. Then the binomial moments of X equals the following (see, e.g., Prékopa [29, p. 182]):

$$E\left[\binom{X}{k}\right] = S_k = \sum_{0 \leq i_1 < i_2 < \cdots < i_k \leq n} P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}), \tag{1.3}$$

$k = 1, 2, \ldots$, and also $E[\binom{X}{0}] = S_0 = 1$. The binomial moment problem has the following form:

$$
\begin{array}{lrlll}
\min(\max) & f_0p_0 & +f_1p_1 \quad +\cdots & +f_np_n & \\
\text{subject to} & & & & \\
& p_0 & +p_1 \quad +\cdots & +p_n & = S_0(=1), \\
& \binom{0}{1}p_0 & +\binom{1}{1}p_1 \quad +\cdots & +\binom{n}{1}p_n & = S_1, \\
& \binom{0}{2}p_0 & +\binom{1}{2}p_1 \quad +\cdots & +\binom{n}{2}p_n & = S_2, \\
& & & & \vdots \\
& \binom{0}{m}p_0 & +\binom{1}{m}p_1 \quad +\cdots & +\binom{n}{m}p_n & = S_m, \\
& p_0, & p_1, \quad \ldots & p_n & \geq 0.
\end{array} \tag{1.4}
$$

If we would like to give bounds for the probability of the union or intersection of all the n events, then we have to consider

$$
f(z) = \begin{cases} 0 \text{ if } z = 0, \\ 1 \text{ otherwise} \end{cases} \quad \text{or } f(z) = \begin{cases} 1 \text{ if } z = s, \\ 0 \text{ otherwise,} \end{cases} \tag{1.5}
$$

respectively. If in problem (1.2) we assume that $\{z_0, z_1, \ldots, z_n\} = \{0, 1, \ldots n\}$, then problems (1.2) and (1.4) can be transformed into each other by simple nonsingular transformations (see Prékopa [25]). This means that the matrices of the equality constraints can be transformed into each other by a nonsingular square matrix and its inverse, respectively. This fact implies that *a basis in problem* (1.2) *is dual feasible if and only if it is dual feasible in* (1.4). Hence, the dual method of Prékopa [27] can be applied as well as closed form bounds can be given for the binomial DMP, too. If we consider the first function in (1.5), then sharp Bonferroni-type bounds can be given for the union of events in terms of $S_0, S_1, \ldots, S_m$; see Prékopa [26], [28], [29] and Boros and Prékopa [3].

1.2. The multivariate discrete moment problem. The multivariate discrete moment problem (MDMP) has been introduced by Prékopa [28] and also discussed in papers by Prékopa [30], [32], Mádi-Nagy and Prékopa [21], and Mádi-Nagy [20]. The MDMP can be formulated as follows. Let $\boldsymbol{X} = (X_1, \ldots, X_s)$ be a random vector and assume that the support of X_j is a known finite set $Z_j = \{z_{j0}, \ldots, z_{jn_j}\}$, consisting of distinct elements, $j = 1, \ldots, s$. A certain set of the following moments will be considered.

DEFINITION 1.1. *The* $(\alpha_1, \ldots, \alpha_s)$-order power moment *of the random vector* $(X_1, \ldots, X_s)$ *is defined as*

$$
\mu_{\alpha_1 \ldots \alpha_s} = E\left[X_1^{\alpha_1} \cdots X_s^{\alpha_s}\right],
$$

where $\alpha_1, \ldots, \alpha_s$ *are nonnegative integers. The sum* $\alpha_1 + \cdots + \alpha_s$ *will be called the* total order of the moment.

We use the following notation for the (unknown) distribution of $\boldsymbol{X}$:

$$
p_{i_1 \ldots i_s} = P(X_1 = z_{1i_1}, \ldots, X_s = z_{si_s}), \qquad 0 \leq i_j \leq n_j,\ j = 1, \ldots, s. \tag{1.6}
$$

Then the moments can be written in the form

$$
\mu_{\alpha_1 \ldots \alpha_s} = \sum_{i_1=0}^{n_1} \cdots \sum_{i_s=0}^{n_s} z_{1i_1}^{\alpha_1} \cdots z_{si_s}^{\alpha_s} p_{i_1 \ldots i_s}.
$$

Let $Z = Z_1 \times \cdots \times Z_s$ and $f(\boldsymbol{z})$, $\boldsymbol{z} \in Z$, be a function. Let

$$
f_{i_1 \ldots i_s} = f(z_{1i_1}, \ldots, z_{si_s}).
$$

The (power) MDMP is to give bounds for

$$E[f(X_1,\dots,X_s)],$$

where distribution of X (i.e., (1.6)) is unknown, but known are the moments

$$\mu_{\alpha_1\dots\alpha_s} \text{ for } (\alpha_1\dots\alpha_s)\in H.$$

We can formulate the problem by the following LP:

$$\begin{aligned} &\min(\max)\ \sum_{i_1=0}^{n_1}\cdots\sum_{i_s=0}^{n_s} f_{i_1\dots i_s}p_{i_1\dots i_s} \\ &\text{subject to} \\ &\sum_{i_1=0}^{n_1}\cdots\sum_{i_s=0}^{n_s} z_{1i_1}^{\alpha_1}\cdots z_{si_s}^{\alpha_s}p_{i_1\dots i_s}=\mu_{\alpha_1\dots\alpha_s} \\ &\text{for } (\alpha_1\dots\alpha_s)\in H, \\ &p_{i_1\dots i_s}\ge 0 \text{ for all } i_1,\dots,i_s. \end{aligned} \tag{1.7}$$

In problem (1.7) $p_{i_1\dots i_s}$, $0\le i_j\le n_j$, $j=1,\dots,s$, are the unknown variables; all other parameters (i.e., the function f and the moments) are given. Regarding the set H, in the literature the following are considered. In Prékopa [30], [32]

$$H=\{(\alpha_1,\dots,\alpha_s)|\ 0\le\alpha_j,\ \alpha_j \text{ integer}, \alpha_1+\cdots+\alpha_s\le m,\ j=1,\dots,s\}, \tag{1.8}$$

where m is a given nonnegative integer, and

$$H=\{(\alpha_1\dots\alpha_s)|\ 0\le\alpha_j\le m_j, \alpha_j \text{ integer},\ j=1,\dots,s\}, \tag{1.9}$$

where m_j, $j=1,\dots,s$, are given nonnegative integers. In Mádi-Nagy and Prékopa [21] and Mádi-Nagy [20]

$$\begin{aligned} &H=\{(\alpha_1,\dots,\alpha_s)|\ 0\le\alpha_j,\ \alpha_j \text{ integer}, \alpha_1+\cdots+\alpha_s\le m,\ j=1,\dots,s; \\ &\text{or } \alpha_j=0,\ j=1,\dots,k-1,k+1,\dots,s,\ m\le\alpha_k\le m_k,\ k=1,\dots,s\} \end{aligned} \tag{1.10}$$

was considered.

The MDMP, beside arising in a natural way, can be applied in several other fields, e.g., bounding expected utilities (Prékopa and Mádi-Nagy [35]), solving generalized s-dimensional transportation problems (Hou and Prékopa [16]) and approximating values of multivariate generating functions (Mádi-Nagy and Prékopa [22]). One of the most popular applications is to bound probabilities of Boolean functions of events. These results are based on the binomial MDMP. Let us introduce the notion of cross-binomial moments.

DEFINITION 1.2. *The* $(\alpha_1,\dots,\alpha_s)$-order cross-binomial moment *of the random vector* $(X_1,\dots,X_s)$*, with the support* $Z\subset\mathbb{N}^s$*, is defined as*

$$S_{\alpha_1\dots\alpha_s}=E\left[\binom{X_1}{\alpha_1}\cdots\binom{X_s}{\alpha_s}\right],$$

where $\alpha_1,\dots,\alpha_s$ *are nonnegative integers. The sum* $\alpha_1+\cdots+\alpha_s$ *will be called* the total order of the moment.

Assume that we have n arbitrary events. We can subdivide them into s subsequences. Let the jth subsequence be designated as $A_{j1},\dots,A_{jn_j}$, $j=1,\dots,s$.

Certainly, $n_1 + \cdots + n_s = n$. Let the random variable X_j with the support $Z_j = \{0, 1, \ldots, n_j\}$ be the number of events that occur in the jth sequence, $j = 1, \ldots, s$. In case of event sequences

$$E\left[\binom{X_1}{\alpha_1}\cdots\binom{X_s}{\alpha_s}\right] = \sum_{\substack{1\le i_{j1}<\cdots<i_{j\alpha_j}\le n_j,\\ j=1,\ldots,s}} P[A_{1i_{11}}\cap\cdots\cap A_{1i_{1\alpha_1}}\cap\cdots\cap A_{si_{s1}}\cap\cdots\cap A_{si_{s\alpha_s}}],$$

in accordance with the notation $S_{\alpha_1\ldots\alpha_s}$ of Definition 1.2. The binomial MDMP can be formulated by the following LP:

$$\begin{aligned} &\min(\max)\ \sum_{i_1=0}^{n_1}\cdots\sum_{i_s=0}^{n_s} f_{i_1\ldots i_s}p_{i_1\ldots i_s} \\ &\text{subject to} \\ &\sum_{i_1=0}^{n_1}\cdots\sum_{i_s=0}^{n_s}\binom{i_1}{\alpha_1}\cdots\binom{i_s}{\alpha_s}p_{i_1\ldots i_s} = S_{\alpha_1\ldots\alpha_s} \\ &\quad\text{for } (\alpha_1\ldots\alpha_s)\in H, \\ &p_{i_1\ldots i_s}\ge 0 \text{ for all } i_1,\ldots,i_s. \end{aligned} \tag{1.11}$$

If we would like to give bounds for the probability of the union or intersection of all the n events, then we have to consider

$$f(z_1,\ldots,z_s) = \begin{cases} 0 \text{ if } (z_1,\ldots,z_s) = (0,\ldots,0), \\ 1 \text{ otherwise} \end{cases} \tag{1.12}$$

or

$$f(z_1,\ldots,z_s) = \begin{cases} 1 \text{ if } (z_1,\ldots,z_s) = (n_1,\ldots,n_j), \\ 0 \text{ otherwise,} \end{cases} \tag{1.13}$$

respectively. If we assume in problem (1.7) that $Z_j = \{0, 1, \ldots, n_j\}$, $j = 1, \ldots, s$, then in the case of the set H (1.10) and (1.8), problems (1.7) and (1.11) can be transformed into each other by simple nonsingular transformations, similar to in the univariate case. Hence, *the dual feasible basis columns in the binomial MDMP and in its transformed power MDMP pair are the same*. The binomial MDMP gives a useful tool to approximate the unknown probabilities, e.g., in network reliability calculation (Habib and Szántai [11]) as well as in probabilistic constrained stochastic programming models (Prékopa [31], Fábián and Szőke [8]). It can also be a good alternative to the bounding techniques of Bukszár and Prékopa [4] and Bukszár and Szántai [5]. This type of probability bound is also useful in developing variance reduction Monte-Carlo simulation algorithms for estimating the exact probability values (Szántai [39], [40]).

Unfortunately, in case of the MDMP we cannot give all the dual feasible structures under any assumption on the function f; hence, we cannot generalize the numerically stable dual method of DMP for the multivariate case. However, under some circumstances, we can give some dual feasible bases which can be used as an initial basis in the dual method. This means, on one hand, that we can skip the first phase in the execution of the dual algorithm, which results in shorter running time and better numerical stability. On the other hand, the objective function value corresponding to a dual feasible basic solution yields a bound for the optimum value of problem (1.7). Hence, if we know a large variety of dual feasible bases, then the best bounds corresponding to those bases can give a good approximation to the optimum value

without any numerical difficulties. In case of (1.9) we know enough dual feasible bases for the approximation, but *in the case of* (1.10) *(and* (1.8)*)* a large number of dual feasible bases are known *only in the bivariate case (*i.e.*,* $s = 2$*).*

The main result of this paper is that we generalize the dual feasible basis structures corresponding to the bivariate case of the set H (1.10), for higher dimensions. We follow a similar way as in the Min and Max Algorithms of Mádi-Nagy and Prékopa [21], but we define a new set of H:

$$
\begin{aligned}
H \quad = \quad & \{(\alpha_1, 0, \ldots, 0, \alpha_j, 0, \ldots, 0) |\ 0 \le \alpha_1, \alpha_j,\ \alpha_1, \alpha_j \text{ integer},\ \alpha_1 + \alpha_j \le m, \\
& j = 2, \ldots, s; \} \\
& \cup \{(0, \ldots, 0, \alpha_j, 0, \ldots, 0) |\ m + 1 \le \alpha_j \le m_j,\ \alpha_j \text{ integer},\ \ j = 1, \ldots, s\}.
\end{aligned} \tag{1.14}
$$

It is easy to see that, in the bivariate case, set (1.14) and set (1.10) are the same. The choice of (1.14) allows us to give a large variety of dual feasible bases of higher dimensional MDMPs under some assumptions on the function f. *In case of* H (1.14) *the power and the binomial MDMP can be transformed into each other by simple nonsingular transformations, which means that the given dual feasible bases can be used in the binomial MDMP, too.* The advantage of a large set of directly given dual feasible bases is twofold. On one hand, we can give good bounds without numerical difficulties even if regular solvers cannot give any useful results. On the other hand, in case of the binomial MDMP, we can take binomial moments (i.e., sums of probabilities of intersections of events) of many subsets of the events into account. Hence, the dual feasible bases of the binomial MDMP can yield better bounds than the bounding methods based on the information of the individual probabilities of the intersection of events, e.g., the method of Bukszár and Prékopa [4]. These advantages will be illustrated by numerical examples.

Our bounding technique of the MDMP is based on multivariate Lagrange interpolation. In section 3 we give a multivariate Lagrange polynomial with prescribed degrees of the variables that approximates the function values of $f(z_1, \ldots, z_s)$, $(z_1, \ldots, z_s) \in Z$ from below or above under some assumptions on the function f.

This paper is organized as follows. In section 2 we present the connection between the MDMP and the multivariate Lagrange interpolation; in section 3 we prove a theorem on multivariate Lagrange interpolation corresponding to our new MDMP. In section 4 we give several dual feasible bases as well as bounds of the MDMP with the set H (1.14). In section 5 numerical examples are presented.

2. The MDMP and the multivariate Lagrange interpolation. In case of univariate Lagrange interpolation concerning arbitrary distinct points $z_0, \ldots, z_n \in \mathbb{R}$ a unique interpolation polynomial of degree n can be given. The multivariate case is much more difficult. On one hand, it is difficult to identify the geometric distribution of the interpolation points for which we can give a (unique) multivariate Lagrange polynomial with prescribed degrees of the variables. On the other hand, it is also difficult to give an appropriate reminder formula.

Regarding the degrees of the Lagrange polynomial we give the following definition.

DEFINITION 2.1. *Let* $H = \{(\alpha_1, \ldots, \alpha_s)\}$ *be a finite set of* s*-tuples of nonnegative integers* $(\alpha_1, \ldots, \alpha_s)$*, and* $\boldsymbol{z} = (z_1, \ldots z_s) \in \mathbb{R}^s$*. We say that* $p(\boldsymbol{z})$ is an H-type polynomial *if its variables have the degrees from the set* H*, i.e.,*

$$
p(\boldsymbol{z}) = \sum_{(\alpha_1, \ldots, \alpha_s) \in H} c_{\alpha_1 \cdots \alpha_s} z_1^{\alpha_1} \cdots z_s^{\alpha_s}, \tag{2.1}
$$

where all $c_{\alpha_1 \cdots \alpha_s}$ *are real.*

On the geometric distributions of the interpolation points we can give the following definition.

DEFINITION 2.2. *Let $U = \{\boldsymbol{u}_1, \dots, \boldsymbol{u}_M\}$ be a set of distinct points in $\mathbb{R}^s$ and $H = \{(\alpha_1, \dots, \alpha_s)\}$ a finite set of s-tuples of nonnegative integers $(\alpha_1, \dots, \alpha_s)$. We say that the set U admits an H-type Lagrange interpolation if for any real function $f(\boldsymbol{z})$, $\boldsymbol{z} \in U$, there exists an H-type polynomial $p(\boldsymbol{z})$, such that*

$$p(\boldsymbol{u}_i) = f(\boldsymbol{u}_i), \qquad i = 1, \dots, M. \tag{2.2}$$

In the literature of multivariate Lagrange interpolation the notion of poised set of points is usual; see, e.g., Definition 1.1 in Gasca and Sauer [9]. Regarding this, Definition 2.2 is equivalent to the notion that the set of points U is *poised* in the linear space of H-type Lagrange polynomials. However, in this paper our definition is more appropriate because, in contrast with other papers on interpolation, we do not deal explicitly with the linear space of polynomials.

The problem of H-type Lagrange interpolation in case of set H (1.8) and (1.9) was discussed in several papers. A good survey on this topic with results can be found in Gasca and Sauer [9]; an earlier historical background is presented in Gasca and Sauer [10]. The case of the set H (1.10) was first investigated in Mádi-Nagy and Prékopa [21]. The results of our case (1.14) will be presented in the following section.

Regarding the connection between the MDMP and the multivariate Lagrange interpolation we should consider the following. Let us use the following notation for the compact matrix form of (1.7), for a given set H:

$$\begin{array}{rl} & \min(\max) \quad \boldsymbol{f}^T\boldsymbol{p} \\ \text{subject to} & \\ & \begin{array}{rcl} A\boldsymbol{p} & = & \boldsymbol{b}, \\ \boldsymbol{p} & \geq & \boldsymbol{0}. \end{array} \end{array} \tag{2.3}$$

DEFINITION 2.3. *Let $\boldsymbol{b}(z_1, \dots, z_s)$ be defined in a similar way as $\boldsymbol{b}$, but we remove the expectation and replace z_j for X_j, $j = 1, \dots, s$; i.e., if we consider a component of the vector $\boldsymbol{b}$, which can be written as $\mu_{\alpha_1 \cdots \alpha_s} = E[X_1^{\alpha_1} \cdots X_s^{\alpha_s}]$ $((\alpha_1, \dots \alpha_s) \in H)$, then the corresponding component of the vector $\boldsymbol{b}(z_1, \dots, z_s)$ will be $z_1^{\alpha_1} \cdots z_s^{\alpha_s}$.*

THEOREM 2.1. *Let us consider a basis B of problem (2.3). Note that the term "basis" as well as the symbol B mean a matrix and, at the same time, the collection of its column vectors. Let I be the set of subscripts corresponding to the columns of B, i.e.,*

$$I = \{(i_1, \dots, i_s) |\ a_{i_1 \cdots i_s} \in B\}, \tag{2.4}$$

where $a_{i_1 \cdots i_s}$ indicates the column of the coefficient matrix A corresponding to the point $(z_{1i_1}, \dots, z_{si_s})$. If we consider the set of distinct points in $\mathbb{R}^s$

$$U = \{(z_{1i_1}, \dots, z_{si_s}) |\ (i_1, \dots, i_s) \in I\}, \tag{2.5}$$

then

$$L_I(z_1, \dots, z_s) \quad = \quad \boldsymbol{f}_B^T B^{-1} \boldsymbol{b}(z_1, \dots, z_s) \tag{2.6}$$

is the unique H-type Lagrange polynomial corresponding to the set U.

Proof. The proof of fitting the interpolation points is yielded by substitution. The uniqueness follows from the fact that if the vector $\boldsymbol{c}$ consists of the corresponding

coefficients of an H-type Lagrange polynomial fitting the set U, then it should be the solution of

$$\boldsymbol{c}^T B = \boldsymbol{f}_B^T. \tag{2.7}$$

Indeed, the only solution of (2.7) is $\boldsymbol{c}^T = \boldsymbol{f}_B^T B^{-1}$. □

THEOREM 2.2. *Let V_{min} (V_{max}) designate the minimum (maximum) value in problem* (2.3). *Further let B_1 (B_2) designate a dual feasible basis, i.e.,*

$$\begin{aligned} &\boldsymbol{f}_{B_1}^T {B_1}^{-1} \boldsymbol{a}_{i_1 \cdots i_s} \le f_{i_1 \cdots i_s} \text{ for all } (z_{i_1}, \ldots, z_{i_s}) \in Z, \\ &(\boldsymbol{f}_{B_2}^T {B_2}^{-1} \boldsymbol{a}_{i_1 \cdots i_s} \ge f_{i_1 \cdots i_s} \text{ for all } (z_{i_1}, \ldots, z_{i_s}) \in Z), \end{aligned} \tag{2.8}$$

for the minimization (maximization) problem. Relation (2.8) *is called the condition of optimality of the minimization (maximization) problem* (2.3). *Then*

$$\boldsymbol{f}_{B_1}^T \boldsymbol{p}_{B_1} \le V_{min} \le E\left[f(X_1, \ldots, X_s)\right] \le V_{max} \le \boldsymbol{f}_{B_2}^T \boldsymbol{p}_{B_2}. \tag{2.9}$$

If B_1 (B_2) is an optimal basis in the minimization (maximization) problem, then the first (last) inequality holds with equality sign. We say that V_{min} and V_{max} are the sharp lower and upper bounds, respectively, for the expectation of $f(X_1, \ldots, X_s)$.

Proof. The theorem follows from the basic results of linear programming theory. □

By the aid of dual feasible bases we can give a special multivariate Lagrange interpolation that approximates the function values of $f(z_1, \ldots, z_s)$ from below (above) in case of the min (max) problem of (2.3).

THEOREM 2.3. *If the basis B is dual feasible in the minimization (maximization) problem and the subscript set I is defined as in* (2.4), *then*

$$\begin{aligned} &f(z_1, \ldots, z_s) \ge L_I(z_1, \ldots, z_s) \text{ for all } (z_1, \ldots, z_s) \in Z, \\ &(f(z_1, \ldots, z_s) \le L_I(z_1, \ldots, z_s) \text{ for all } (z_1, \ldots, z_s) \in Z), \end{aligned} \tag{2.10}$$

where equality holds in case of $(z_1, \ldots, z_s) \in U$ of (2.5). *Regarding $E[f(X_1, \ldots, X_s)]$ we have the following bound:*

$$\begin{aligned} &E[f(X_1, \ldots, X_s)] \ge E[L_I(X_1, \ldots, X_s)], \\ &(E[f(X_1, \ldots, X_s)] \le E[L_I(X_1, \ldots, X_s)]). \end{aligned} \tag{2.11}$$

If the basis is also primal feasible, then it is optimal and, thus, the obtained bound is sharp.

Proof. The inequality (2.10) follows from (2.6) and (2.8). The bound (2.11) can be obtained if we replace $(X_1, \ldots, X_s)$ for $(z_1, \ldots, z_s)$ and take the expectations in (2.10). □

THEOREM 2.4. *Assume that $L_I(\boldsymbol{z})$ is an H-type Lagrange polynomial corresponding to the points Z_I and inequality* (2.10) *is satisfied. If the columns in the* min *(*max*) problem* (2.3) *corresponding to the interpolation points Z_I are linearly independent, then they form a dual feasible basis.*

Proof. Let B_1 (B_2) be the matrix that consists of the columns corresponding to the points of Z_I. Since B_1 (B_2) is basis, (2.10) is equivalent to the condition of optimality (2.8). Thus, B_1 (B_2) is dual feasible in the min (max) problem (2.3). □

In the following section we give a formula of an H-type Lagrange polynomial and its remainder, where the set H is the type of (1.14). We fix only the subscript set I of the interpolation points, which means that we can get several sets

$$Z_I = \{(z_{1i_1}, \ldots, z_{si_s}) |\ (i_1, \ldots, i_s) \in I\} \tag{2.12}$$

that admit an H-type Lagrange interpolation. Indeed, if we change the order of the elements in the sequence $\{z_{j0},\dots,z_{jn_j}\}$ (or, equivalently, we change the subscripts of the elements within the sets Z_j), $j=1,\dots,s$, then the set of interpolation points Z_I will be different.

The coefficients of the Lagrange polynomial will be given in terms of the multivariate divided differences of the function f. Hence, we give the following definition.

DEFINITION 2.4. *Let $f(z),\ z\in\{z_0,\dots,z_n\}$, be a univariate discrete function, where $z_0,\dots,z_n$ are distinct real numbers,*

$$[z_i;f]:=f(z_i),\ \textit{where } z_i\in\{z_0,\dots,z_n\}.$$

The kth order (univariate) divided differences *($k\geq 1$) are defined recursively:*

$$[z_i,\dots,z_{i+k};f]=\frac{[z_{i+1},\dots,z_{i+k};f]-[z_i,\dots,z_{i+k-1};f]}{z_{i+k}-z_i},\ \text{where } z_i\in\{z_0,\dots,z_n\}.$$

DEFINITION 2.5. *Let $f(\boldsymbol{z}),\ \boldsymbol{z}\in Z=Z_1\times\cdots\times Z_s$, be a multivariate discrete function and take the subset*

$$\begin{aligned} Z_{I_1\dots I_s} &= \{z_{1i},\ i\in I_1\}\times\cdots\times\{z_{si},\ i\in I_s\}\\ &= Z_{1I_1}\times\cdots\times Z_{sI_s}, \end{aligned} \tag{2.13}$$

*where $|I_j|=k_j+1,\ j=1,\dots,s$. Then we can define the $(k_1,\dots,k_s)$-*order (multivariate) divided difference of *f on the set* (2.13) *in an iterative way. First we take the k_1th divided difference with respect to the first variable, then the k_2th divided difference with respect to the second variable, etc. These operations can be executed in any order even in a mixed manner, the result always being the same. Let*

$$[z_{1i},\ i\in I_1;\cdots;z_{si},\ i\in I_s;f] \tag{2.14}$$

designate the $(k_1,\dots,k_s)$-order divided difference. The sum $k_1+\cdots+k_s$ is called the total order of the divided difference.

In order to make the definition easier to understand we present the following example.

Example 2.1.

$$[z_{10},z_{11};z_{20},z_{21};f]=\left[z_{20},z_{21};\frac{f(z_{11},z_2)-f(z_{10},z_2)}{z_{11}-z_{10}}\right]$$

$$=\frac{\frac{f(z_{11},z_{21})-f(z_{10},z_{21})}{z_{11}-z_{10}}-\frac{f(z_{11},z_{20})-f(z_{10},z_{20})}{z_{11}-z_{10}}}{z_{21}-z_{20}}.$$

3. A theorem on multivariate Lagrange interpolation. Let $f(\boldsymbol{z})$, $\boldsymbol{z}\in Z=Z_1\times\cdots\times Z_s$, where $Z_j=\{z_{j0},\dots,z_{jn_j}\}$ consists of distinct real values, $j=1,\dots,s$. In this section we present an H-type Lagrange polynomial and its remainder on the interpolation points $Z_I=\{(z_{1i_1},\dots,z_{si_s})|\ (i_1,\dots,i_s)\in I\}$, where the set H is defined by (1.14) and

$$I=\left(\bigcup_{j=1}^{s}I_j\right)\cup\left(\bigcup_{j=1}^{s}J_j\right), \tag{3.1}$$

where

$$I_1 = \{(i_1, 0, \dots, 0) |\ 0 \le i_1 \le m-1,\ \text{integers}\}, \tag{3.2}$$

$$I_j = \{(i_1, 0, \dots, 0, i_j, 0, \dots, 0) |\ 0 \le i_1,\ 1 \le i_j \le m-1,\ \text{integers},\ i_1 + i_j \le m\},\quad j = 2, \dots, s,$$

and

$$\begin{aligned} &J_j = \{(0, \dots, 0, i_j, 0, \dots, 0) |\ i_j \in K_j\},\\ &K_j = \{k_j^{(1)}, \dots, k_j^{(|K_j|)}\} \subset \{m, m+1, \dots, n_j\},\\ &|K_j| = mj + 1 - m,\ j = 1, \dots, s. \end{aligned} \tag{3.3}$$

In what follows we will use the notations

$$\begin{aligned} Z_{ji} &= \{z_{j0}, \dots, z_{ji}\},\\ Z'_{ji} &= \{z_{j0}, \dots, z_{ji}, z_j\},\\ &\quad i = 0, \dots, n_j,\ j = 1, \dots, s, \end{aligned}$$

and

$$\begin{aligned} K_{ji} &= \{k_j^{(1)}, \dots, k_j^{(i)}\},\\ Z_{jK_{ji}} &= \{z_{jk_j^{(1)}}, \dots, z_{jk_j^{(i)}}\},\\ &\quad i = 1, \dots, |K_j|,\ j = 1, \dots, s,\\ Z_{jK_j} &= Z_{jK_{j|K_j|}},\ j = 1, \dots, s. \end{aligned}$$

We assign the Lagrange polynomial, given by its Newton form

$$\begin{aligned} &L_I(z_1, \dots, z_s)\\ &= \sum_{i_1=0}^{m-1} [Z_{1i_1}; Z_{20}; \cdots; Z_{s0}; f] \prod_{k=0}^{i_1-1} (z_1 - z_{1k})\\ &+ \sum_{j=2}^{s} \sum_{\substack{i_1+i_j \le m\\ 1 \le i_j \le m-1}} \left[Z_{1i_1}; Z_{20}; \cdots; Z_{(j-1)0}; Z_{ji_j}; Z_{(j+1)0}; Z_{s0}; f\right]\\ &\qquad \times \prod_{k=0}^{i_1-1} (z_1 - z_{1k}) \prod_{k=0}^{i_j-1} (z_j - z_{jk})\\ &+ \sum_{j=1}^{s} \sum_{i=1}^{|K_j|} \left[Z_{10}; \cdots; Z_{(j-1)0}; Z_{j(m-1)} \cup Z_{jK_{ji}}; Z_{(j+1)0}; \cdots; Z_{s0}; f\right]\\ &\qquad \times \prod_{\substack{k \in \{0, \dots, m-1\} \cup K_{j(i-1)}\\ k \ne i_j - 1}} (z_j - z_{jk}), \end{aligned} \tag{3.4}$$

where, by definition, $\prod_{k=0}^{i_j-1} (z_j - z_{jk}) = 1$, for $i_j = 0$, and $K_{j0} = \emptyset$.

In (3.4) the function f is not necessarily restricted to the set Z as its domain of definition; it may be defined on any subset of $\mathbb{R}^s$ that contains Z.

Next, we define the residual function:

$$R_I(z_1, \dots, z_s) = R_{1I}(z_1, \dots, z_s) + R_{2I}(z_1, \dots, z_s) + R_{3I}(z_1, \dots, z_s), \tag{3.5}$$

where

$$
\begin{aligned}
&R_{1I}(z_1,\dots,z_s)\\
&=\sum_{j=1}^{s}\left[z_{10};\cdots;z_{(j-1)0};Z_{j(m-1)}\cup Z_{jK_j}\cup\{z_j\};z_{(j+1)0};\cdots;z_{s0};f\right]\\
&\qquad\times\prod_{k\in\{0,\dots,m-1\}\cup K_j}(z_j-z_{jk})
\end{aligned}
\tag{3.6}
$$

and

$$
\begin{aligned}
&R_{2I}(z_1,\dots,z_s)\\
&=\sum_{j=2}^{s}\Bigg(\sum_{\substack{i_1+i_j=m\\0\le i_1,i_j\le m-1}}\left[Z'_{1i_1};Z_{20};\cdots;Z_{(j-1)0};Z_{ji_j};Z_{(j+1)0};\cdots;Z_{s0};f\right]\\
&\qquad\times\prod_{l=0}^{i_1}(z_1-z_{1l})\prod_{k=0}^{i_j-1}(z_j-z_{jk})\\
&\qquad+[Z'_{10};Z_{20};\cdots;Z_{(j-1)0};Z'_{j(m-1)};Z_{(j+1)0};\cdots;Z_{s0};f]\\
&\qquad\times(z_1-z_{10})\prod_{k=0}^{m-1}(z_j-z_{jk})\Bigg)
\end{aligned}
\tag{3.7}
$$

and

$$
\begin{aligned}
&R_{3I}(z_1,\dots,z_s)\\
&=\sum_{h=2}^{s-1}\sum_{j=h+1}^{s}\left[z_1;\cdots;z_{h-1};Z'_{h0};Z_{(h+1)0};\cdots;Z_{(j-1)0};Z'_{j0};Z_{(j+1)0};\cdots;Z_{s0};f\right]\\
&\qquad\times(z_h-z_{h0})(z_j-z_{j0}).
\end{aligned}
\tag{3.8}
$$

We prove the following theorem.

THEOREM 3.1. *Consider the H-type Lagrange polynomial* (3.4)*, corresponding to the points Z_I. For any $\boldsymbol{z}=(z_1,\dots,z_s)$ for which the function f is defined, we have the equality*

$$
L_I(z_1,\dots,z_s)+R_I(z_1,\dots,z_s)=f(z_1,\dots,z_s).
\tag{3.9}
$$

Proof. For the sake of simplicity we assume that $m_j\le n_j$, $j=1,\dots,s$. The proof of the general case needs only slight modification. Now we consider the following lemma.

LEMMA 3.2. *We have the equality*

$$
\begin{aligned}
&L_I(z_1,\dots,z_s)+R_{1I}(z_1,\dots,z_s)\\
&=\sum_{i_1=0}^{m-1}[Z_{1i_1};Z_{20};\cdots;Z_{s0};f]\prod_{k=0}^{i_1-1}(z_1-z_{1k})\\
&+\sum_{j=2}^{s}\sum_{\substack{i_1+i_j\le m\\1\le i_j\le m-1}}\left[Z_{1i_1};Z_{20};\cdots;Z_{(j-1)0};Z_{ji_j};Z_{(j+1)0};Z_{s0};f\right]\\
&\times\prod_{l=0}^{i_1-1}(z_1-z_{1l})\prod_{k=0}^{i_j-1}(z_j-z_{jk})\\
&+\sum_{j=1}^{s}[Z_{10};\cdots;Z_{(j-1)0};Z'_{j(m-1)};Z_{(j+1)0};\cdots;Z_{s0};f]\prod_{k=0}^{m-1}(z_j-z_{jk}).
\end{aligned}
\tag{3.10}
$$

Proof of Lemma 3.2. The proof is the same as that of Lemma 3.2 in Mádi-Nagy and Prékopa [21]. □

LEMMA 3.3.

$$
\begin{aligned}
&L_I(z_1,\dots,z_s)+R_{1I}(z_1,\dots,z_s)+R_{2I}(z_1,\dots,z_s)\\
&=[z_1;Z_{20};\cdots;Z_{s0};f]\\
&+\sum_{j=2}^{s}\left[z_1;Z_{20};\cdots;Z_{(j-1)0};Z'_{j0};Z_{(j+1)0};\cdots;Z_{s0};f\right](z_j-z_{j0}).
\end{aligned}
\tag{3.11}
$$

Proof of Lemma 3.3. Let us look at the terms of (3.7) and (3.11) as univariate functions of z_1. Adding up the first term of (3.10) and the case of $j=1$ in the third term of (3.10) we get the first term of (3.11); i.e.,

$$
\sum_{i_1=0}^{m-1}[Z_{1i_1};Z_{20};\cdots;Z_{s0};f]\prod_{k=0}^{i_1-1}(z_1-z_{1k})+[Z'_{1(m-1)};Z_{20};\cdots;Z_{s0};f]\prod_{k=0}^{m-1}(z_1-z_{1k})
$$

$$
=[z_1;Z_{20};\cdots;Z_{s0};f]. \tag{3.12}
$$

Let us consider the terms of the second sum of (3.10) for a given value of j and i_j. Adding them to the terms of the first part of the sum of (3.7) with the same value of j and i_j we have

$$
\sum_{i_1\le m-i_j}\left[Z_{1i_1};Z_{20};\cdots;Z_{(j-1)0};Z_{ji_j};Z_{(j+1)0};Z_{s0};f\right]\prod_{k=0}^{i_j-1}(z_j-z_{jk})\prod_{l=0}^{i_1-1}(z_1-z_{1l})
$$

$$
+[Z'_{1(m-i_j)};Z_{20};\cdots;Z_{(j-1)0};Z_{ji_j};Z_{(j+1)0};\cdots;Z_{s0};f]\prod_{k=0}^{i_j-1}(z_j-z_{jk})\prod_{l=0}^{m-i_j}(z_1-z_{1l})
$$

$$
=\left[z_1;Z_{20};\cdots;Z_{(j-1)0};Z_{ji_j};Z_{(j+1)0};\cdots;Z_{s0};f\right]\prod_{k=0}^{i_j-1}(z_j-z_{jk}). \tag{3.13}
$$

Adding up all terms of (3.10) and (3.7) corresponding to a given $2\le j\le s$, by the use of the result (3.13) we have

$$
\sum_{i_j=1}^{m-1}\left[z_1;Z_{20};\cdots;Z_{(j-1)0};Z_{ji_j};Z_{(j+1)0};\cdots;Z_{s0};f\right]\prod_{k=0}^{i_j-1}(z_j-z_{jk})
$$

$$
+[Z_{10};\cdots;Z_{(j-1)0};Z'_{j(m-1)};Z_{(j+1)0};\cdots;Z_{s0};f]\prod_{k=0}^{m-1}(z_j-z_{jk})
$$

$$
+[Z'_{10};Z_{20};\cdots;Z_{(j-1)0};Z'_{j(m-1)};Z_{(j+1)0};\cdots;Z_{s0};f](z_1-z_{10})\prod_{k=0}^{m-1}(z_j-z_{jk})
$$

$$
\begin{aligned}
&= \sum_{i_j=1}^{m-1} \left[z_1; Z_{20}; \cdots; Z_{(j-1)0}; Z_{ji_j}; Z_{(j+1)0}; \cdots; Z_{s0}; f\right] \prod_{k=0}^{i_j-1} (z_j - z_{jk}) \\
&+ \left[z_1; Z_{20}; \cdots; Z_{(j-1)0}; Z'_{j(m-1)}; Z_{(j+1)0}; \cdots; Z_{s0}; f\right] \prod_{k=0}^{m-1} (z_j - z_{jk}) .
\end{aligned} \tag{3.14}
$$

Considering (3.14) as a function of z_j we get that this sum equals

$$
\left[z_1; Z_{20}; \cdots; Z_{(j-1)0}; Z'_{j0}; Z_{(j+1)0}; \cdots; Z_{s0}; f\right] (z_j - z_{j0}) . \tag{3.15}
$$

Adding up (3.12) and the terms (3.15) for $j = 2, \ldots, s$ we get (3.11). Thus, the lemma is proven. □

Proof of Theorem 3.1. We prove that

$$
\begin{aligned}
&(L_I(z_1, \ldots, z_s) + R_{1I}(z_1, \ldots, z_s) + R_{2I}(z_1, \ldots, z_s)) + R_{3I}(z_1, \ldots, z_s) \\
&= f(z_1, \ldots, z_s),
\end{aligned}
$$

where the brackets emphasize that we will use (3.11). Let us choose the $h = 2$ case in the sum of (3.8) and add it to (3.11). Considering them as univariate functions of z_2 the result is

$$
\begin{aligned}
&\sum_{j=3}^{s} \left[z_1; Z'_{20}; Z_{30}; \cdots; Z_{(j-1)0}; Z'_{j0}; Z_{(j+1)0}; \cdots; Z_{s0}; f\right] (z_2 - z_{20}) (z_j - z_{j0}) \\
&+ [z_1; Z_{20}; \cdots; Z_{s0}; f] + \sum_{j=2}^{s} \left[z_1; Z_{20}; \cdots; Z_{(j-1)0}; Z'_{j0}; Z_{(j+1)0}; \cdots; Z_{s0}; f\right] (z_j - z_{j0}) \\
&= ([z_1; Z_{20}; \cdots; Z_{s0}; f] + [z_1; Z'_{20}; Z_{30}; \cdots; Z_{s0}; f] (z_2 - z_{20})) \\
&+ \left(\sum_{j=3}^{s} \left[z_1; Z_{20}; \cdots; Z_{(j-1)0}; Z'_{j0}; Z_{(j+1)0}; \cdots; Z_{s0}; f\right] (z_j - z_{j0}) \right. \\
&\left. + \sum_{j=3}^{s} \left[z_1; Z'_{20}; Z_{30}; \cdots; Z_{(j-1)0}; Z'_{j0}; Z_{(j+1)0}; \cdots; Z_{s0}; f\right] (z_2 - z_{20}) (z_j - z_{j0}) \right)
\end{aligned}
$$

$$
\begin{aligned}
&= [z_1; z_2; Z_{30}; \cdots; Z_{s0}; f] \\
&+ \sum_{j=3}^{s} \left[z_1; z_2; Z_{30}; \cdots; Z_{(j-1)0}; Z'_{j0}; Z_{(j+1)0}; \cdots; Z_{s0}; f\right] (z_j - z_{j0}) .
\end{aligned} \tag{3.16}
$$

Adding the $h = 3$ case in the sum of (3.8) to (3.16) we get a similar formula with z_3 in the places of the third variable. Finally, after choosing all the cases $h = 2, 3, \ldots, s-1$ in (3.8) and adding them in this order to (3.16) we have

$$
[z_1; z_2; \cdots, z_{s-1}; Z_{s0}; f] + \sum_{j=s}^{s} [z_1; z_2; \cdots; z_{s-1}; Z'_{s0}; f] (z_j - z_{j0}) .
$$

Considering this as the function of z_s, the result is

$$[z_1; z_2; \cdots, z_{s-1}; z_s; f] = f(z_1, \ldots, z_s).$$

Thus, the theorem is proven. □

THEOREM 3.4. *The H-type Lagrange polynomial* (3.4) *of the function f corresponding to the points Z_I is unique (for a given set Z). Furthermore, the set of columns B of A in problem* (2.3)*, corresponding to the subscript set I, forms a basis of A.*

Proof. The formula (3.4) gives a Lagrange polynomial on Z_I for any function f on the set Z. This means that we can obtain the coefficients of the H-type Lagrange polynomial in case of *any* function $f(\boldsymbol{z})$, $\boldsymbol{z} \in Z$. Let the matrix B consist of the columns corresponding to the points of Z_I. The coefficient vector $\boldsymbol{c}_f$ of the Lagrange polynomial corresponding to the function f is the solution of the equation

$$\boldsymbol{c}_f^T B = \boldsymbol{f}_B^T. \tag{3.17}$$

Since we have Lagrange interpolation for *any* f, (3.17) has a solution $\boldsymbol{c}_f$ in case of any vector $\boldsymbol{f}_B$. From this follows that the square matrix B is nonsingular, hence *the columns corresponding to the points of Z_I form a basis.* Thus, the second assertion is proven. Then the uniqueness is the corollary of Theorem 2.1. □

4. Bounds for the MDMP with the set H (1.14). The problem in the title of this section is the following LP:

$$\begin{aligned}
&\min(\max) \sum_{i_1=0}^{n_1} \cdots \sum_{i_s=0}^{n_s} f_{i_1 \ldots i_s} p_{i_1 \ldots i_s} \\
&\text{subject to} \\
&\sum_{i_1=0}^{n_1} \cdots \sum_{i_s=0}^{n_s} z_{1i_1}^{\alpha_1} z_{ji_j}^{\alpha_j} p_{i_1 \ldots i_s} = \mu_{\alpha_1 0 \ldots 0 \alpha_j 0 \ldots 0} \\
&\quad \text{for } \alpha_1, \alpha_j \geq 0,\ \alpha_1 + \alpha_j \leq m,\ j = 2, \ldots, s; \\
&\sum_{i_1=0}^{n_1} \cdots \sum_{i_s=0}^{n_s} z_{ji_j}^{\alpha_j} p_{i_1 \ldots i_s} = \mu_{0 \ldots 0 \alpha_j 0 \ldots 0} \\
&\quad \text{for } m+1 \leq \alpha_j \leq m_j,\ j = 1, \ldots, s; \\
&p_{i_1 \ldots i_s} \geq 0 \text{ for all } i_1, \ldots, i_s.
\end{aligned} \tag{4.1}$$

In the following theorems let us designate the coefficient matrix of (4.1) by A, the right-hand side vector by $\boldsymbol{b}$, and the coefficient vector of the objective function by $\boldsymbol{f}$, in agreement with (2.3).

We present several dual feasible bases for the min problem of (4.1), which give lower bounds for the objective function. As at the numerical examples we shall see, the best bounds, among the given ones, are usually close to the values of the sharp bounds. This means that we can give good bounds without using LP solvers, which work numerically unstably because of the Vandermonde systems in the coefficient matrix.

First of all, we should introduce some assumptions. They are as follows.

Assumption 1. The function $f(\boldsymbol{z})$, $\boldsymbol{z} \in Z$,

(a) has nonnegative univariate divided differences of order $m_j + 1$ regarding z_j, $j = 1, \ldots, s$,

(b) has nonnegative bivariate divided differences of order $m + 1$,

(c) has nonnegative mixed second order divided differences.

If $f(\boldsymbol{z}),\ \boldsymbol{z} \in Z$, is derived from a function $\overline{f}(\boldsymbol{z})$ defined in $\overline{Z} = [z_{10}, z_{1n_1}] \times \cdots \times [z_{s0}, z_{sn_s}]$ by taking $f(\boldsymbol{z}) = \overline{f}(\boldsymbol{z}),\ \boldsymbol{z} \in Z$, and $\overline{f}(\boldsymbol{z})$ has continuous, nonnegative derivatives of order $(k_1, \dots, k_s)$ in the interior of $\overline{Z}$, then all divided differences of $f(\boldsymbol{z}),\ \boldsymbol{z} \in Z$, of order $(k_1, \dots, k_s)$ are nonnegative. For further results in this respect, see Popoviciu [24]. This fact helps one find out whether a function satisfies Assumption 4. If in the assumption we require nonpositivity, then we shall see that the following bases will be dual feasible for the max problem and will give upper bounds for (4.1).

We introduce four different structures for K_j, defined in (3.4), as follows:

$$\begin{array}{cll} & |K_j| \text{ even} & |K_j| \text{ odd} \\ \min & u^{(j)}, u^{(j)}+1, \dots, v^{(j)}, v^{(j)}+1 & m, u^{(j)}, u^{(j)}+1, \dots, v^{(j)}, v^{(j)}+1 \\ \max & m, u^{(j)}, u^{(j)}+1, \dots, v^{(j)}, v^{(j)}+1, n_j & u^{(j)}, u^{(j)}+1, \dots, v^{(j)}, v^{(j)}+1, n_j; \end{array} \tag{4.2}$$

i.e., K_j (consists of distinct elements of the subset of $\{m, \dots, n_j\}$) is a set of pairs of consecutive elements completed by m and n_j depending on its parity and its (min or max) type.

As regards the ordering of the elements in the sets $Z_1, \dots, Z_s$, we mention separately in each theorem of this section what our assumption is about.

THEOREM 4.1. *Let $z_{j0} < z_{j1} < \cdots < z_{jn_j},\ j = 1, \dots, s$. Suppose that K_j follows the min structure of* (4.2)*, $j = 1, \dots, s$.*

Under Assumption 4*, $L_I(z_1, \dots, z_s)$, defined by* (3.4)*, is a unique H-type Lagrange polynomial on Z_I and satisfies the relations*

$$f(z_1, \dots, z_s) \geq L_I(z_1, \dots, z_s),\ (z_1, \dots, z_s) \in Z. \tag{4.3}$$

The set of columns B of A in problem (4.1)*, with the subscript set I, is a dual feasible basis in the minimization problem* (4.1)*, and*

$$E[f(X_1, \dots, X_s)] \geq E[L_I(X_1, \dots, X_s)]. \tag{4.4}$$

If B is also a primal feasible basis in problem (4.1)*, then the inequality* (4.4) *is sharp.*

Proof. We have only to prove (4.3). The proof of the other parts of the theorem is straightforward from Theorems 3.4, 2.4, and 2.3.

In order to prove (4.3) it is sufficient to show that

$$R_I(z_1, \dots, z_s) \geq 0 \text{ for all } (z_1, \dots, z_s) \in Z. \tag{4.5}$$

In fact, we show that all terms in the sum $R_I(z_1, \dots, z_s)$ are nonnegative. Considering any term in $R_{1I}(z_1, \dots, z_s)$ the first factor of the product is a divided difference, which is nonnegative because of Assumption 4, while the last part is

$$\prod_{k \in \{0, \dots, m-1\} \cup K_j} (z_j - z_{jk}) > 0 \text{ for } z_j \notin \{z_{j0}, \dots, z_{j(m-1)}\} \cup Z_{jK_j} \tag{4.6}$$

because there are even numbers of negative factors, due to the special structure of K_j. If $z_j \in \{z_{j0}, \dots, z_{j(m-1)}\} \cup Z_{jK_j}$, the above product is 0. This means that any term of $R_{1I}(z_1, \dots, z_s)$ is nonnegative. Considering a term of the sum $R_{2I}(z_1, \dots, z_s)$ or $R_{3I}(z_1, \dots, z_s)$ the first factor of the product is a divided difference, which is nonnegative because of Assumption 4. In the other part of the product at least one of the factors is zero or all factors are positive because of the ordering of Z_j's.

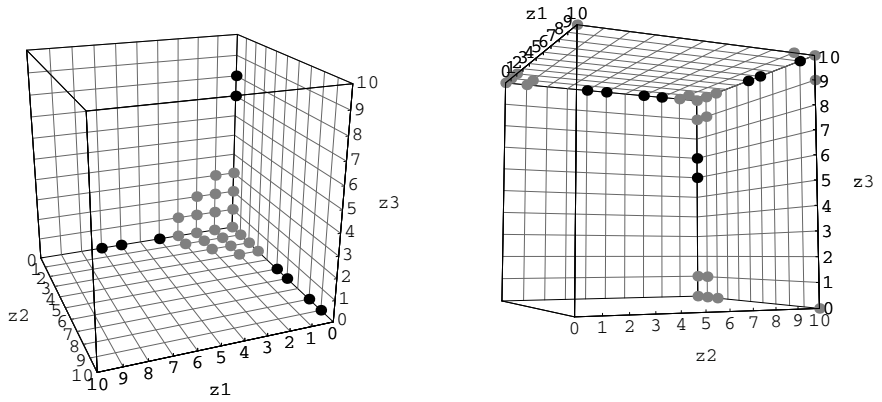


FIG. 1. *Assume* $Z_j = \{0, \dots, 10\}$, $j = 1, 2, 3$, $m = 4$, $m_1 = 7$, $m_2 = 8$, $m_3 = 6$. (a): *Dual feasible basis corresponding to Theorem* 4.1, *where* $K_1 = \{4, 6, 7\}$, $K_2 = \{5, 6, 8, 9\}$, $K_3 = \{7, 8\}$. (b): *Dual feasible basis of the Min Algorithm, where* $(z_{10}, z_{11}, z_{12}, z_{13}) = (0, 1, 10, 2)$, $(z_{j0}, z_{j1}, z_{j2}, z_{j3}) = (10, 0, 9, 1)$, $j = 2, 3$, $K_1 = \{6, 7, 10\}$ *(*$Z_{1K_1} = \{5, 6, 9\}$*)*, $K_2 = \{6, 7, 9, 10\}$ *(*$Z_{2K_2} = \{4, 5, 7, 8\}$*)*, $K_3 = \{8, 9\}$ *(*$Z_{3K_3} = \{6, 7\}$*). Elements of* $\bigcup_{j=1}^{s} I_j$ *are colored by gray while elements of* $\bigcup_{j=1}^{s} J_j$*'s are black.*

This means that all factors of the product are nonnegative and hence the term is nonnegative as well. All terms of $R_I(z_1, \dots, z_s)$ are nonnegative, hence the sum of them, i.e., $R_I(z_1, \dots, z_s)$, is also nonnegative on Z. Thus, the theorem is proven. □

The basis of Theorem 4.1 is illustrated in Figure 1(a).

In the following algorithm we will consider more orders of the elements of the set Z_j in the sequence $(z_{j0}, \dots, z_{jn_j})$, $j = 1, \dots, s$, for which all terms of $R_I(z_1, \dots, z_s)$ are nonnegative. Note that transposing the elements of the sequence $(z_{j0}, \dots, z_{jn_j})$ is equivalent to exchanging subscripts of the elements; however, the proof of the validity of the algorithm is based on the orders of the elements. At each order of the elements, the set of interpolation points Z_I as well as the corresponding dual feasible basis are different, which means that we can give more lower bounds on (4.1) by the objective function value of the dual feasible basis corresponding to Z_I.

Min Algorithm.

At first we assume, without loss of generality, that $Z_j = \{0, 1, \dots, n_j\}$, $j = 1, \dots, s$.

Step 0. Let

$$z_{20} = z_{30} = \cdots = z_{s0} = 0 \tag{4.7}$$

or

$$z_{20} = n_2, z_{30} = n_3, \dots, z_{s0} = n_s. \tag{4.8}$$

If (4.7) holds, then let $0 \le q_1 \le m$ be an even number, else (if (4.8) holds) let $0 \le q_1 \le m$ be an odd number. $L := (0, 1, \dots, (m-1)-q_1)$, $U := (n_1, n_1-1, \dots, n_1-(q_1-1))$, $V^0 := \{\text{arbitrary merger of the sequences } L, U\} = (v^0, v^1, \dots, v^{m-1})$.

$$(z_{10}, \dots, z_{1(m-1)}) := V^0.$$

Let $j = 2$. Goto Step 1.

Step 1. Initialize $t = 0$. If (4.7) holds, then let $l_0 = 1$ and $u_0 = n_j$, else let $l_0 = 0$ and $u_0 = n_j - 1$. Goto Step 2.

Step 2. Let $V^t = \{v^0, v^1, \dots, v^{m-1-t}\}$, $H^t = \{h^1, \dots, h^t\}$. If $v^{m-1-t} \in L$, then let $h^{t+1} = l^t$, $l^{t+1} = l^t + 1$, $u^{t+1} = u^t$, and if $v^{m-1-t} \in U$, then let $h^{t+1} = u^t$, $u^{t+1} = u^t - 1$, $l^{t+1} = l^t$. Set $t \leftarrow t+1$. If $t = m$, then goto Step 3, else repeat Step 2.

Step 3. Let

$$(z_{j1}, \dots, z_{j(m-1)}) = H^{m-1}.$$

Set $j \leftarrow j + 1$. If $j = s + 1$, then goto Step 4, else goto Step 1.

Step 4. Let $0, 1, \dots, r_j, n_j, \dots, n_j - (m - r_j - 2)$ be the numbers used to construct $z_{j0}, z_{j1}, \dots, z_{j(m-1)}$. Then let

$$(z_{jm}, z_{j(m+1)}, \dots, z_{jn_j}) = (r_j + 1, r_j + 2, \dots, n_j - (m - r_j - 1)) \tag{4.9}$$

as ordered sets, $j = 1, \dots, s$. Note that these subsets of the sets Z_j, $j = 1, \dots, s$, remain intact. If $m - r_j - 1$ is even, then K_j should follow a minimum structure in (4.2), and if $m - r_j - 1$ is odd, then K_j should follow a maximum structure. Stop, we have completed the construction of the dual feasible basis related to the subscript set I.

In the general case, where Z_j is not necessarily $\{0, 1, \dots, n_j\}$, $j = 1, \dots, s$, we do the following. First we order the elements in each Z_j in increasing order. Then we establish one-to-one correspondences between the elements of Z_j and the elements of the ordered set $(0, 1, \dots, n_j)$. After that, we carry out the Min Algorithm to find a dual feasible basis, using the sets $\{0, 1, \dots, n_j\}$, $j = 1, \dots, s$. Finally, we create the set Z_I, by the use of the above mentioned one-to-one correspondences.

THEOREM 4.2. *Let the elements of the sequence* $(z_{j0}, \dots, z_{jn_j})$, $j = 1, \dots, s$, *be in one of the orders of the Min Algorithm, and also let* K_j *follow the* min *(*max*) structure if* $m - r_j - 1$ *is even (odd) in the Min Algorithm,* $j = 1, \dots, s$.

Under Assumption 4, $L_I(z_1, \dots, z_s)$, *defined by* (3.4), *is a unique* H*-type Lagrange polynomial on* Z_I *and satisfies the relations*

$$f(z_1, \dots, z_s) \geq L_I(z_1, \dots, z_s), \qquad (z_1, \dots, z_s) \in Z. \tag{4.10}$$

The set of columns B *of* A *in problem* (4.1), *with the subscript set* I, *is a dual feasible basis in the minimization problem* (4.1), *and*

$$E[f(X_1, \dots, X_s)] \geq E[L_I(X_1, \dots, X_s)]. \tag{4.11}$$

If B *is also a primal feasible basis in problem* (4.1), *then the inequality* (4.11) *is sharp.*

Proof. We have only to prove that $R_I(z_1, \dots, z_s) \geq 0$ for all $(z_1, \dots, z_s) \in Z$ and then we can follow the proof of Theorem 4.1. We can restrict ourselves to the case $Z_j = \{0, 1, \dots, n_j\}$, $j = 1, \dots, s$, because in the following inequalities only the orders of the elements play a role.

First, let us consider $R_{3I}(z_1, \dots, z_s)$ of (3.8). All divided differences are nonnegative, and the set of the second parts of the products equals

$$(z_h - z_{h0})(z_j - z_{j0}), \qquad 2 \leq h < j \leq s. \tag{4.12}$$

These products can be nonnegative if and only if both factors are nonnegative or both factors are nonpositive. Because we have pairs for all $2 \leq h < j \leq s$ this means that

either all factors $(z_j - z_{j0})$, $2 \le j \le s$, are nonnegative, or all of them are nonpositive. The first case is provided by (4.7) while the second by (4.8). Hence, $R_{3I}(\boldsymbol{z}) \ge 0$, $\boldsymbol{z} \in Z$.

If we consider $R_{2I}(z_1, \ldots, z_s)$ of (3.7), we will see that the factors of products after the nonnegative divided differences are associated with the following arrays:

$$\begin{array}{cccccccc} z_{10} & z_{11} & z_{12} & \cdots & z_{1(m-2)} & z_{1(m-1)} & & z_{j0} \\ z_{10} & z_{11} & z_{12} & \cdots & z_{1(m-2)} & & z_{j0} & z_{j1} \\ & & & & \vdots & & & \\ z_{10} & z_{11} & z_{j0} & \cdots & z_{j(m-4)} & z_{j(m-3)} & & z_{j(m-2)} \\ z_{10} & z_{j0} & z_{j1} & \cdots & z_{j(m-3)} & z_{j(m-2)} & & z_{j(m-1)} \end{array} \tag{4.13}$$

$$j = 2, \ldots, s.$$

A sufficient condition for the nonnegativity of all products in (3.7) is that

$$|\{i|\ 0 \le i \le i_1,\ z_{1i} > z_1\}| + |\{i|\ 0 \le i \le i_j,\ z_{ji} > z_j\}| = \text{even number} \tag{4.14}$$

for every $0 \le i_j \le m-1$ integers satisfying $i_1 + i_j = m-1$, and for all $(z_1, z_j) \in Z_1 \times Z_j$, $j = 2, \ldots, s$. The first m elements of the first row in (4.13) are the elements of V^0; the $m+1$st element of the same row is z_{j0}. In Step 0 the parity of q_1 provides that (4.14) is satisfied for the product of (3.7) corresponding to the first row of (4.13). The elements of V^t, z_{j0}, H^t, in that order, constitute the tth row of tableau (4.13). In Steps 1 and 2 we define the following element h^{t+1} such that (4.14) is still satisfied for the product of (3.7) corresponding to the $t+1$st row of (4.13). From this follows that $R_{2I}(\boldsymbol{z}) \ge 0$, $\boldsymbol{z} \in Z$.

Regarding $R_{1I}(z_1, \ldots, z_s)$ the divided differences are nonnegative in terms of (3.6). Hence, after the assignments (4.9) we have only to choose the subscript sets K_j such that

$$\prod_{k \in \{0, \ldots, m-1\} \cup K_j} (z_j - z_{jk}) \ge 0, \qquad \boldsymbol{z} \in Z,$$

for $j = 1, \ldots, s$. For each j the products

$$\prod_{k=0}^{m-1} (z_j - z_{jk}),\ (z_1, \ldots, z_s) \in Z \tag{4.15}$$

are nonnegative (nonpositive) if $m - r_j - 1$ is even (odd). (Note that $r_2 = \cdots = r_s$ by construction.) Then the choice of K_j (with assignment (4.9)), $j = 1, \ldots, s$, provides that $R_{1I}(\boldsymbol{z}) \ge 0$ for all $\boldsymbol{z} \in Z$.

Thus, $R_I(\boldsymbol{z}) = R_{1I}(\boldsymbol{z}) + R_{2I}(\boldsymbol{z}) + R_{3I}(\boldsymbol{z}) \ge 0$ for all $\boldsymbol{z} \in Z$. □

The basis of the algorithm is illustrated in Figure 1(b).

The above algorithm allows for the construction of a variety of dual feasible bases. However, we do not have a simple criterion, like in the dual method, to decide which of the bases, that we can obtain by the above Min Algorithm, would improve on the bound (on the value of the objective function). Hence, in order to give the best bound for $E[f(X_1, \ldots, X_s)]$ we have to calculate the objective function values for all dual feasible bases yielded by the Min Algorithm and then choose the highest value from among them. For a given order of $(z_{j0}, \ldots, z_{js})$, $j = 1, \ldots, s$, we can find the

best sets K_j, $j = 1, \ldots, s$, independently. The argument of this can be found in Mádi-Nagy [20].

Numerical experiments show that this method produces very good results much faster than the execution of the dual algorithm. Its other main advantage is that it is not sensitive to the numerical difficulties that arise from the bad numerical property of the matrix A.

5. Numerical examples. In the following we consider three numerical examples. In the first one we give bounds for the expected value of a three-variate utility function by the aid of the Min Algorithm of section 4 and we also give the sharp bounds calculated by the LP solver CPLEX (www.ilog.com). The method of the Min Algorithm is written in Wolfram's Mathematica (www.wolfram.com). This means that its running times are not really comparable with the running times of the dual method of CPLEX; however, we give these results, too. Because of the numerical instability of the MDMP, CPLEX sometimes cannot give appropriate results (it reports infeasibility; however, the problem is feasible by construction). These cases are indicated by "−" sign in the tables. We can conclude that our Min Algorithm yields useful bounds without numerical difficulties even if the CPLEX cannot give the right answer.

In the last two examples we give bounds for the union of events by the application of the Min Algorithm for the binomial MDMP (1.11). Comparing the results to the cherry tree bounds of Bukszár and Prékopa [4] and the bounds of univariate and bivariate MDMPs, the bounds of our Min Algorithm turn out to be the best.

Example 5.1. In this example we solve a problem similar to Example 4.2 in Prékopa and Mádi-Nagy [35]. Consider the utility function

$$u(z_1, z_2, z_3) = \log\left[(e^{\alpha_1 z_1 + a_1} - 1)(e^{\alpha_2 z_2 + a_2} - 1)(e^{\alpha_3 z_3 + a_3} - 1) - 1\right], \quad (z_1, z_2, z_3) \in Z, \tag{5.1}$$

where Z is specialized as follows:

$$Z = (0, 1, 2, 3, 4, 5, 6, 7, 8, 9) \times (0, 1, 2, 3, 4, 5, 6, 7, 8, 9) \times (0, 1, 2, 3, 4, 5, 6, 7, 8, 9).$$

Let the parameters be $\alpha_1 = 1.75,\ \alpha_2 = 1.25,\ \alpha_3 = 0.75,\ a_1 = 3,\ a_2 = 2,\ a_3 = 1$. In the cited paper it was proven that the even (odd) order derivatives of the function are nonpositive (nonnegative). We use our Min Algorithm for the MDMP (1.7) with set H (1.14) to give an upper bound for the expected utility

$$E[u(X_1, X_2, X_3)]. \tag{5.2}$$

Since the even order divided differences of the function $-u(z_1, z_2, z_3)$ on Z are nonnegative, we can apply the Min Algorithm to give a lower bound for $E[-u(X_1, X_2, X_3)]$ that yields, indeed, an upper bound for (5.2).

Regarding the moments taken into account they are generated from the distribution of (X_1, X_2, X_3), defined by

$$X_1 = \min(X + Y_1, 9),$$

$$X_2 = \min(X + Y_2, 9),$$

$$X_3 = \min(X + Y_3, 9),$$

where X, Y_1, Y_2, Y_3 have independent Poisson distributions with parameters $1, 2, 2.5, 3$, respectively. Note that X_1, X_2, X_3 are stochastically dependent.

We calculate the bounds for several values of the parameters m, m_1, m_2, m_3 of the set H (1.14). We also give the sharp upper bounds calculated by the dual method of CPLEX (where it is possible). In order to see the closeness of our (not sharp) upper bounds we also calculate the minimum of the objective function value. That gives the possibility to compare the difference between our upper bound and the dual maximum to the gap between the sharp upper and lower bounds. The results are summarized below.

m	m_1	m_2	m_3	Upper bound	CPU	Dual max	CPU	Dual min	CPU
3	3	3	3	18.586715186	0.14	18.557668477	0.08	18.510511279	0.11
3	5	5	5	18.549493662	0.41	18.546562878	0.05	18.539798360	0.40
3	7	7	7	18.543584031	0.69	18.543580944	0.11	18.542882716	0.69
3	9	9	9	18.543304428	0.36	–	–	–	–

m	m_1	m_2	m_3	Upper bound	CPU	Dual max	CPU	Dual min	CPU
5	5	5	5	18.562168728	1.50	18.546562667	0.14	18.539911300	0.65
5	7	7	7	18.543588614	2.61	18.543580474	0.44	18.542921460	1.16
5	9	9	9	18.543304428	2.31	–	–	–	–

We can see that our bounds are very close to the sharp bounds except cases $m = m_1 = \cdots = m_3 = 3$ and $m = m_1 = \cdots = m_3 = 5$. On the other hand, we can give better approximation taking the marginal moments up to the order 9 into account, despite the fact that CPLEX cannot solve the MDMP.

We also check whether we do not lose too much information using the set H (1.14) instead of the set H (1.10) (set H (1.14) is a subset of H (1.10)). The sharp bounds for set H (1.10) are summarized below.

m	m_1	m_2	m_3	Dual max	CPU	Dual min	CPU
3	3	3	3	18.557668337	0.28	1.8510552138	0.12
3	5	5	5	18.546562783	0.53	1.8539798515	0.42
3	7	7	7	18.543580810	1.54	1.8542882927	0.11
3	9	9	9	–	–	–	–

m	m_1	m_2	m_3	Dual max	CPU	Dual min	CPU
5	5	5	5	18.546560902	2.13	1.8539941923	1.52
5	7	7	7	18.543579438	6.62	1.8542948216	3.69
5	9	9	9	–	–	–	–

We can conclude that these bounds are not significantly better. Hence, we do not lose as much as we can gain on the structure of the set H (1.14), which makes the Min Algorithm applicable for higher dimensions.

Example 5.2. In this example we shall give upper bounds for the probability of the union of events of the following two systems. In both systems there are 12 events $A_1, A_2, \ldots, A_{12}$, and 16 outcomes, $x_1, x_2, \ldots x_{16}$ with probabilities $P(x_1), P(x_2), \ldots, P(x_{16})$, respectively. The events of System I are defined by the matrix $R^I = (r^I_{ij})$, where $r^I_{ij} = 1$, if $x_i \in A_j$; otherwise $r^I_{ij} = 0$. The events of System II are defined by the matrix R^{II}, similarly. (The matrices R^I and R^{II} were randomly

generated with different densities.)

$$R^I = \begin{pmatrix} 0&0&0&0&0&0&1&0&0&0&0&1\\ 0&0&1&0&1&0&0&0&0&0&1&0\\ 0&0&1&0&0&0&0&0&0&0&0&0\\ 0&1&0&0&1&1&0&0&0&0&0&0\\ 1&1&0&1&0&0&0&1&0&1&0&0\\ 0&1&0&0&1&1&0&1&0&0&0&1\\ 0&0&1&1&0&0&0&0&0&0&0&0\\ 0&0&0&0&1&0&0&0&0&0&0&0\\ 1&0&0&0&0&0&0&0&0&1&1&0\\ 1&0&0&0&0&1&0&0&1&1&0&1\\ 0&0&1&0&0&1&0&0&0&0&1&0\\ 0&0&1&1&0&1&1&0&0&0&0&1\\ 0&1&0&0&0&1&1&0&0&1&1&0\\ 1&0&0&0&1&0&0&1&0&0&1&0\\ 0&1&0&0&0&1&0&0&0&1&0&0\\ 0&0&0&0&0&0&0&0&0&0&0&0 \end{pmatrix}, \; R^{II} = \begin{pmatrix} 1&1&0&1&1&1&0&1&1&1&1&0\\ 0&1&1&0&1&1&0&0&1&0&1&0\\ 1&0&1&1&0&1&1&1&1&1&1&0\\ 1&1&1&0&1&1&1&0&1&1&1&1\\ 1&1&1&1&1&1&1&0&1&1&1&1\\ 1&1&1&1&1&1&1&1&0&0&1&1\\ 1&1&1&1&1&1&1&1&1&1&1&1\\ 1&1&1&0&0&1&1&1&0&1&1&1\\ 1&1&1&1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&0&1&0&1&1&1\\ 1&1&1&1&1&0&1&1&1&1&0&1\\ 1&1&0&1&0&1&0&1&0&1&1&1\\ 1&1&1&1&1&1&1&1&1&1&1&1\\ 1&1&1&0&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&0&0&1&0&1&1&0\\ 0&0&0&0&0&0&0&0&0&0&0&0 \end{pmatrix}. \tag{5.3}$$

It is easy to see that at least one event occurs at each outcome except x_{16}. From this

$$P(\cup A_i) = 1 - P(x_{16}).$$

Regarding the probability of the outcomes we consider the following three cases.

	$P(x_1)$	$P(x_2)$	$P(x_3)$	$P(x_4)$	$P(x_5)$	$P(x_6)$	$P(x_7)$	$P(x_8)$
Case 1:	0.023	0.034	0.045	0.056	0.067	0.078	0.067	0.056
Case 2:	0.012	0.022	0.023	0.033	0.034	0.044	0.045	0.055
Case 3:	0.0329	0.1076	0.0599	0.1108	0.042	0.0055	0.0508	0.1142

(5.4)

	$P(x_9)$	$P(x_{10})$	$P(x_{11})$	$P(x_{12})$	$P(x_{13})$	$P(x_{14})$	$P(x_{15})$	$P(x_{16})$
Case 1:	0.045	0.038	0.011	0.022	0.033	0.044	0.055	0.326
Case 2:	0.056	0.066	0.067	0.077	0.078	0.088	0.089	0.211
Case 3:	0.048	0.0235	0.0676	0.0295	0.0441	0.1265	0.1058	0.0313

We use several bounding techniques depending on the known information about the systems. On one hand, we give the (sharp) upper bounds of the univariate binomial moment problem of (1.4) with the first function f in (1.5), based on the information on S_1, S_2, S_3. We use the dual method of Prékopa [27].

On the other hand, we subdivide the sequence of $\{A_1, A_2, \ldots, A_{12}\}$ into subsequences

$$\{A_1, A_2, \ldots, A_6\},\ \{A_7, A_8, \ldots, A_{12}\} \tag{5.5}$$

and then into subsequences

$$\{A_1, A_2, A_3, A_4\},\ \{A_5, A_6, A_7, A_8\},\ \{A_9, A_{10}, A_{11}, A_{12}\} \tag{5.6}$$

and then into

$$\{A_1, A_2, A_3\},\ \{A_4, A_5, A_6\},\ \{A_7, A_8, A_9\},\ \{A_{10}, A_{11}, A_{12}\}. \tag{5.7}$$

We consider the multivariate binomial moment problem (1.11) with the function f in (1.12) with the appropriate sets H (1.14), respectively. Prékopa [30] has shown that the even (odd) order divided differences of f in (1.12) are nonpositive (nonnegative). This means that in cases where $m+1$ and m_j+1, $j = 1, \ldots, s$ (where s is the number of subsequences), are even we can give upper bounds by the application of the Min

TABLE 1
Results of System I.

	Univariate S_1, S_2, S_3	(5.5) $m = 3$, $m_1 = 3$, $m_2 = 3$	(5.6) $m = 3$, $m_1 = 3$, $m_2 = 3$, $m_3 = 3$	(5.7) $m = 3$, $m_1 = m_2 = 3$, $m_3 = m_4 = 3$	Cherry tree bound
Case 1:	0.7412	1	0.969	0.697	0.785
Case 2:	0.8595	1	1	0.801	0.900
Case 3:	1	1	1	1	1

TABLE 2
Results of System II.

	Univariate S_1, S_2, S_3	(5.5) $m = 3$, $m_1 = 3$, $m_2 = 3$	(5.6) $m = 3$, $m_1 = 3$, $m_2 = 3$, $m_3 = 3$	(5.7) $m = 3$, $m_1 = m_2 = 3$, $m_3 = m_4 = 3$	Cherry tree bound
Case 1:	0.781345	0.749467	0.702333	0.674	0.685
Case 2:	0.921273	0.873867	0.807333	0.789	0.789
Case 3:	1	1	1	0.9687	0.9687

Algorithm for $-f(\boldsymbol{z})$. In this example we use the Min Algorithm with parameters $m = 3$, $m_j = 3$, $j = 1, \ldots, s$.

Finally, we give the so-called cherry tree bounds of Bukszár and Prékopa [4]. These bounds are based on the knowledge of the individual probabilities of the occurrences of the events, of the intersections of pairs of events, and of the intersections of three events. These bounds are always at least as good as the Hunter–Worsley second order bounds (see Hunter [17]).

The results for System I and System II are summarized in Tables 1 and 2, respectively. Comparing the bounds, we can see that we get the best bound in case of the subsequences (5.7).

Regarding the order of the columns in Tables 1 and 2, we took more and more information on the system into account; i.e., in case of the univariate binomial moments we just considered the sums of probabilities of events, intersections of pairs of events, and triples of events. In the case of multivariate binomial moments we considered the sums of those probabilities of smaller groups of the events; i.e., we separated the sums of the univariate moments into subsums. Finally, in the case of cherry trees we used individual probabilities. This would imply that we should get better and better bounds. However, the bounds of our paper as well as the cherry tree bounds are not sharp; that is why it could happen that *the Min Algorithm gave better bounds, using less information, than the cherry tree bounds of Bukszár and Prékopa* [4].

The numerical results confirm that our method yields a new effective tool for bounding the probability of the union of events based on the knowledge of probabilities of the intersection up to three events. On the other hand, we can see that the bivariate bounds corresponding to the subsequences (5.5) are much weaker than the bounds corresponding to the binomial MDMPs of higher dimensions. *Since our generalization of the bivariate Min Algorithm gives the possibility of taking more detailed information into account, it can give substantially better bounds.*

Example 5.3. We shall also give upper bounds for the probability of the union of events. There are 20 events $A_1, A_2, \dots, A_{20}$, and 16 outcomes, $x_1, x_2, \dots, x_{16}$ with probabilities $P(x_1), P(x_2), \dots, P(x_{16})$, respectively. The events of the system are defined by the matrix $R = (r_{ij})$, where $r_{ij} = 1$, if $x_i \in A_j$; otherwise $r_{ij} = 0$.

$$R = \begin{pmatrix}
1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\
0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\
1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}. \tag{5.8}$$

The matrix R is taken from System 1 of Example 2 in Prékopa and Gao [34], but we use the probabilities of outcomes of (5.4) instead of their probabilities because in their system the probability of the union is nearly 1. We consider the univariate binomial DMP for S_1, S_2, S_3 and for $S_1, \dots, S_5$. We also give the results of the multivariate binomial DMP (1.11) with the function f in (1.12) with the appropriate sets H (1.14). We consider the subdivisions

$$\{A_1, \dots, A_7\},\ \{A_8, \dots, A_{14}\},\ \{A_{15}, \dots, A_{20}\}, \tag{5.9}$$

with $m = m_j = 5$ $(j = 1, 2, 3)$, and

$$\{A_1, \dots, A_5\},\ \{A_6, \dots, A_{10}\},\ \{A_{11}, \dots, A_{15}\},\ \{A_{16}, \dots, A_{20}\}, \tag{5.10}$$

with $m = m_j = 5$ $(j = 1, 2, 3, 4)$. (The MDMPs with third order binomial moments yield trivial bounds.) We also give the cherry tree bounds. The results are summarized in Table 3. Among the third order bounds once the univariate binomial DMP bound, once the cherry tree bound was the best, however, none of them were useful. However, *using moments up to the order* 5 *our binomial MDMP, with subdivision* (5.10)*, yields the best (in these cases the sharp) bounds.*

TABLE 3
Results of the system of Example 5.3.

	Univariate S_1, S_2, S_3	Cherry tree bound	Univariate $S_1, \dots, S_5$	(5.9) $m = 5$, $m_j = 5$	(5.10) $m = 5$, $m_j = 5$
Case 1:	0.794036	0.823	0.6882	0.96691	0.674
Case 2:	0.960964	0.901	0.810878	1	0.789
Case 3:	1	1	0.99522	1	0.9687

REFERENCES

[1] N. I. Akhiezer, *The Classical Moment Problem and Some Related Questions in Analysis*, Hafner Publishing, New York, 1965.

[2] I. Bienaymé, *Considérations a l'appui de la découverte de Laplace sur la loi de probabilité dans la méthode des moindres carré*, C.R. Acad. Sci. Paris, 37 (1853), pp. 309–326.

[3] E. Boros and A. Prékopa, *Closed form two-sided bounds for probabilities that at least r and exactly r out of n events occur*, Math. Oper. Res., 14 (1989), pp. 317–342.

[4] J. Bukszár and A. Prékopa, *Probability bounds with cherry trees*, Math. Oper. Res., 26 (2001), pp. 174–192.

[5] J. Bukszár and T. Szántai, *Probability bounds given by hypercherry trees*, Optim. Methods Softw., 17 (2002), pp. 409–422.

[6] P. Chebyshev, *Sur les valeurs limites des intégrales*, Journal de Mathématiques Pures et Appliquées, 19 (1874), pp. 157–160.

[7] P. Chebyshev, *Sur deux théorèmes relatifs aux probabilités*, Acta Math., 14 (1890), pp. 305–315.

[8] C. I. Fábián and Z. Szőke, *Solving two-stage stochastic programming problems with level decomposition*, Comput. Manag. Sci., 4 (2007), pp. 313–353.

[9] M. Gasca and T. Sauer, *Polynomial interpolation in several variables*, Adv. Comput. Math., 12 (2000), pp. 377–410.

[10] M. Gasca and T. Sauer, *On the history of multivariate polynomial interpolation*, J. Comput. Appl. Math., 122 (2000), pp. 23–35.

[11] A. Habib and T. Szántai, *New bounds on the reliability of the consecutive k-out-of-r-from-n: F system*, Reliability Engineering and System Safety, 68 (2000), pp. 97–106.

[12] H. Hamburger, *Beiträge zur Konvergenztheorie der Stieltjesschen Kettenbrüche*, Math. Z., 4 (1919), pp. 186–222.

[13] H. Hamburger, *Uber eine Erweiterung des Stieltjesschen Momentproblems*, Math. Ann., 81 (1920), pp. 235–319; 82 (1921), pp. 120–164 and pp. 168–187.

[14] F. Hausdorff, *Summationsmethoden und Momentfolgen*, I *and* II, Math. Z., 16 (1921), pp. 74–109 and pp. 280–299.

[15] F. Hausdorff, *Momentprobleme für ein endliches Intervall*, Math. Z., 16 (1923), pp. 220–248.

[16] X. Hou and A. Prékopa, *Monge property and bounding multivariate probability distribution functions with given marginals and covariances*, SIAM J. Optim., 18 (2007), pp. 138–155.

[17] D. Hunter, *An upper bound for the probability of a union*, J. Appl. Prob., 13 (1976), pp. 597–603.

[18] T. H. Kjeldsen, *The early history of the moment problem*, Historia Math., 20 (1993), pp. 19–44.

[19] M. G. Krein and A. A. Nudelman, *The Markov Moment Problem and Extremal Problems*. Translations of Mathematical Monographs 50, AMS, Providence, RI, 1977.

[20] G. Mádi-Nagy, *A method to find the best bounds in a multivariate discrete moment problem if the basis structure is given*, Studia Sci. Math. Hungar., 42 (2005), pp. 207–226.

[21] G. Mádi-Nagy and A. Prékopa, *On multivariate discrete moment problems and their applications to bounding expectations and probabilities*, Math. Oper. Res., 29 (2004), pp. 229–258.

[22] G. Mádi-Nagy and A. Prékopa, *Bounding Expectations of Functions of Random Vectors with Given Marginals and Some Moments: Applications of the Multivariate Discrete Moment Problem*, RUTCOR Research Report, 11–2007, Piscataway, NJ, 2007.

[23] A. Markov, *On Certain Applications of Algebraic Continued Fractions*, Ph.D. thesis, St. Petersburg, Russia, 1884.

[24] T. Popoviciu, *Les Fonctions Convexes*, Actualités Scientifiques et Industrielles 992, Hermann et Cie, Paris, 1944.

[25] A. Prékopa, *Boole-Bonferroni inequalities and linear programming*, Oper. Res., 36 (1988), pp. 145–162.

[26] A. Prékopa, *Sharp bounds on probabilities using linear programming*, Oper. Res., 38 (1990), pp. 227–239.

[27] A. Prékopa, *The discrete moment problem and linear programming*, Discrete Appl. Math., 27 (1990), pp. 235–254.

[28] A. Prékopa, *Inequalities on expectations based on the knowledge of multivariate moments*, in Stochastic Inequalities, M. Shaked and Y. L. Tong, eds., Lecture Notes Monogr. Ser. 22, Inst. Math. Statist., Hayward, CA, 1992, pp. 309–331.

[29] A Prékopa, *Stochastic Programming*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1995.

[30] A. Prékopa, *Bounds on probabilities and expectations using multivariate moments of discrete distributions*, Studia Sci. Math. Hungar., 34 (1998), pp. 349–378.

[31] A. Prékopa, *The use of discrete moment bounds in probabilistic constrained stochastic programming models*, Ann. Oper. Res., 85 (1999), pp. 21–38.
[32] A. Prékopa, *On Multivariate Discrete Higher Order Convex Functions and Their Applications*, RUTCOR Research Report, 39-2000, Piscataway, NJ, 2000. Also in, Proceedings of the Sixth International Conference on Generalized Convexity and Monotonicity, Karlovasi, Samos, Greece, to appear.
[33] A. Prékopa and G. Alexe, *Dual Methods for the Numerical Solution of the Univariate Power Moment Problem*, RUTCOR Research Report, 14–2003, Piscataway, NJ, 2003.
[34] A. Prékopa and L. Gao, *Bounding the probability of the union of events by aggregation and disaggregation in linear programs*, Discrete Appl. Math., 145 (2005), pp. 444–454.
[35] A. Prékopa and G. Mádi-Nagy, *A class of multiattribute utility functions*, Econom. Theory, 34 (2008), pp. 591–602.
[36] F. Riesz, *Sur le probléme des moments*, Arkiv for Matematik Astronomi och Fysik, 17 (1923), pp. 1–52.
[37] S. M. Samuels and W. J. Studden, *Bonferroni-type probability bounds as an application of the theory of Tchebycheff system*, in Probability, Statistics and Mathematics, Papers in Honor of Samuel Karlin, Academic Press, Boston, 1989, pp. 271–289.
[38] T. J. Stieltjes, *Recherches sur les fractions continues*, Ann. Fac. Sci. Univ. Toulouse, 8 (1894), pp. J76–J122; 9 (1895), pp. A5–A47.
[39] T. Szántai, *Evaluation of a special multivariate gamma distribution function*, Math. Programming Stud., 27 (1986), pp. 1–16.
[40] T. Szántai, *Improved bounds and simulation procedures on the value of the multivariate normal probability distribution function*, Ann. Oper. Res., 100 (2000), pp. 85–101.

SIAM J. OPTIM.
Vol. 19, No. 4, pp. 1807–1827

SMOOTH OPTIMIZATION APPROACH FOR SPARSE COVARIANCE SELECTION*

ZHAOSONG LU†

Abstract. In this paper we first study a smooth optimization approach for solving a class of nonsmooth *strictly* concave maximization problems whose objective functions admit smooth convex minimization reformulations. In particular, we apply Nesterov's smooth optimization technique [Y. E. Nesterov, *Dokl. Akad. Nauk SSSR*, 269 (1983), pp. 543–547; Y. E. Nesterov, *Math. Programming*, 103 (2005), pp. 127–152] to their dual counterparts that are smooth convex problems. It is shown that the resulting approach has $\mathcal{O}(1/\sqrt{\epsilon})$ iteration complexity for finding an ϵ-optimal solution to both primal and dual problems. We then discuss the application of this approach to sparse covariance selection that is approximately solved as an l_1-norm penalized maximum likelihood estimation problem, and also propose a variant of this approach which has substantially outperformed the latter one in our computational experiments. We finally compare the performance of these approaches with other first-order methods, namely, Nesterov's $\mathcal{O}(1/\epsilon)$ smooth approximation scheme and block-coordinate descent method studied in [A. d'Aspremont, O. Banerjee, and L. El Ghaoui, *SIAM J. Matrix Anal. Appl.*, 30 (2008), pp. 56–66; J. Friedman, T. Hastie, and R. Tibshirani, *Biostatistics*, 9 (2008), pp. 432–441] for sparse covariance selection on a set of randomly generated instances. It shows that our smooth optimization approach substantially outperforms the first method above, and moreover, its variant substantially outperforms both methods above.

Key words. sparse covariance selection, nonsmooth strictly concave maximization, smooth minimization

AMS subject classifications. 90C22, 90C25, 90C47, 65K05, 62J10

DOI. 10.1137/070695915

1. Introduction.

In [19, 21], Nesterov proposed an efficient smooth optimization method for solving convex programming problems of the form

$$\min\{f(u) : u \in U\}, \tag{1}$$

where f is a convex function with Lipschitz continuous gradient, and U is a closed convex set. It is shown that his method has $\mathcal{O}(1/\sqrt{\epsilon})$ iteration complexity bound, where $\epsilon > 0$ is the absolute precision of the final objective function value. A proximal-point-type algorithm for (1) having the same complexity as above has also been proposed more recently by Auslender and Teboulle [2].

Motivated by [10], we are particularly interested in studying the use of a smooth optimization approach for solving a class of nonsmooth *strictly* concave maximization problems whose objective functions admit smooth convex minimization reformulations in this paper. Our key idea is to apply Nesterov's smooth optimization technique [19, 21] to their dual counterparts that are smooth convex problems. It is shown that the resulting approach has $\mathcal{O}(1/\sqrt{\epsilon})$ iteration complexity for finding an ϵ-optimal solution to both primal and dual problems.

One interesting application of the above approach is for sparse covariance selection. Given a set of random variables with Gaussian distribution for which the

*Received by the editors June 30, 2007; accepted for publication (in revised form) October 3, 2008; published electronically February 20, 2009.
http://www.siam.org/journals/siopt/19-4/69591.html

†Department of Mathematics, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada (zhaosong@sfu.ca). This author was supported in part by NSERC Discovery Grant and SFU President's Research Grant.

true covariance matrix is unknown, covariance selection is a procedure used to estimate true covariance from a sample covariance matrix by maximizing its likelihood while imposing a certain sparsity on the inverse of the covariance estimation (e.g., see [11]). Therefore, it can be applied to determine a robust estimate of the true variance matrix, and simultaneously to discover the sparse structure in the underlying model. Despite its popularity in numerous real-world applications (e.g., see [3, 10, 25] and the references therein), sparse covariance selection itself is a challenging NP-hard combinatorial optimization problem. By an argument that is often used in regression techniques such as LASSO [23], Yuan and Lin [25] and d'Aspremont et al. [10] (see also [3]) showed that it can be approximately solved as an l_1-norm penalized maximum likelihood estimation problem. Moreover, the authors of [10] studied two efficient first-order methods for solving this problem, that is, Nesterov's smooth approximation scheme and block-coordinate descent method. It was shown in [10] that their first method has $\mathcal{O}(1/\epsilon)$ iteration complexity for finding an ϵ-optimal solution. For their second method, each iterate requires solving a box constrained quadratic programming, and it has a local linear convergence rate. However, its global iteration complexity for finding an ϵ-optimal solution is theoretically unknown. After the first release of our paper, Friedman, Hastie, and Tibshirani [16] studied a slight variant of the block-coordinate descent method proposed in [10]. At each iteration of their method, a coordinate descent approach is applied to solve a lasso (l_1-regularized) least-squares problem, which is the dual of the box constrained quadratic programming appearing in the block-coordinate descent method [10]. In contrast with these methods, the smooth optimization approach proposed in this paper has a more attractive iteration complexity that is $\mathcal{O}(1/\sqrt{\epsilon})$ for finding an ϵ-optimal solution. In addition, we propose a variant of the smooth optimization approach which has substantially outperformed the latter one in our computational experiments. We also compare the performance of our approaches with their methods for sparse covariance selection on a set of randomly generated instances. It shows that our smooth optimization approach substantially outperforms their first method above (i.e., Nesterov's smooth approximation scheme) and, moreover, its variant substantially outperforms their methods [10, 16] mentioned above.

The paper is organized as follows. In section 2, we introduce a class of nonsmooth concave maximization problems in which we are interested and propose a smooth optimization approach to them. In section 3, we briefly introduce sparse covariance selection and show that it can be approximately solved as an l_1-norm penalized maximum likelihood estimation problem. We also discuss the application of the smooth optimization approach for solving this problem and propose a variant of this approach. In section 4, we compare the performance of our smooth optimization approach and its variant with two other first-order methods studied in [10, 16] for sparse covariance selection on a set of randomly generated instances. Finally, we present some concluding remarks in section 5.

1.1. Notation. In this paper, all vector spaces are assumed to be finite-dimensional. The space of symmetric $n \times n$ matrices will be denoted by $\mathcal{S}^n$. If $X \in \mathcal{S}^n$ is positive semidefinite, we write $X \succeq 0$. Also, we write $X \preceq Y$ to mean $Y - X \succeq 0$. The cone of positive semidefinite (resp., definite) matrices is denoted by $\mathcal{S}^n_+$ (resp., $\mathcal{S}^n_{++}$). Given matrices X and Y in $\Re^{p\times q}$, the standard inner product is defined by $\langle X, Y\rangle := \mathrm{Tr}(XY^T)$, where $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix. $\|\cdot\|$ denotes the Euclidean norm and its associated operator norm unless it is explicitly stated otherwise. The Frobenius norm of a real matrix X is defined as $\|X\|_F := \sqrt{\mathrm{Tr}(XX^T)}$.

We denote by e the vector of all ones, and by I the identity matrix. Their dimensions should be clear from the context. For a real matrix X, we denote by $\mathrm{Card}(X)$ the cardinality of X, that is, the number of nonzero entries of X, and denote by $|X|$ the absolute value of X, that is, $|X|_{ij} = |X_{ij}|$ for all i, j. The determinant and the minimal (resp., maximal) eigenvalue of a real symmetric matrix X are denoted by $\det X$ and $\lambda_{\min}(X)$ (resp., $\lambda_{\max}(X)$), respectively. For an n-dimensional vector w, $\mathrm{diag}(w)$ denotes the diagonal matrix whose ith diagonal element is w_i for $i = 1, \ldots, n$. We denote by $\mathcal{Z}_+$ the set of all nonnegative integers.

Let the space $\mathcal{F}$ be endowed with an arbitrary norm $\|\cdot\|$. The dual space of $\mathcal{F}$, denoted by $\mathcal{F}^*$, is the normed real vector space consisting of all linear functionals of $s : \mathcal{F} \to \Re$, endowed with the dual norm $\|\cdot\|^*$ defined as

$$\|s\|^* := \max_u \{\langle s, u\rangle : \ \|u\| \le 1\} \quad \forall s \in \mathcal{F}^*,$$

where $\langle s, u\rangle := s(u)$ is the value of the linear functional s at u. Finally, given an operator $\mathcal{A} : \mathcal{F} \to \mathcal{F}^*$, we define

$$\mathcal{A}[H, H] := \langle \mathcal{A}H, H\rangle$$

for any $H \in \mathcal{F}$.

2. Smooth optimization approach. In this section, we consider a class of concave nonsmooth maximization problems:

$$\max_{x \in X} \{g(x) := \min_{u \in U} \phi(x, u)\}, \tag{2}$$

where X and U are nonempty convex compact sets in finite-dimensional real vector spaces $\mathcal{E}$ and $\mathcal{F}$, respectively, and $\phi(x, u) : X \times U \to \Re$ is a continuous function which is *strictly* concave in $x \in X$ for every fixed $u \in U$, and convex differentiable in $u \in U$ for every fixed $x \in X$. Therefore, for any $u \in U$, the function

$$f(u) := \max_{x \in X} \phi(x, u) \tag{3}$$

is well-defined. We also easily conclude that $f(u)$ is convex differentiable on U, and its gradient is given by

$$\nabla f(u) = \nabla_u \phi(x(u), u) \ \ \forall u \in U, \tag{4}$$

where $x(u)$ denotes the unique solution of (3).

Let the space $\mathcal{F}$ be endowed with an arbitrary norm $\|\cdot\|$. We further assume that $\nabla f(u)$ is Lipschitz continuous on U with respect to $\|\cdot\|$, i.e., there exists some $L > 0$ such that

$$\|\nabla f(u) - \nabla f(\tilde{u})\|^* \le L\|u - \tilde{u}\| \quad \forall u, \tilde{u} \in U.$$

Under the above assumptions, we easily observe that (i) problem (2) and its dual, that is,

$$\min_u \{f(u) : \ u \in U\}, \tag{5}$$

are both solvable and have the same optimal value; and (ii) the dual problem (5) can be suitably solved by Nesterov's smooth minimization approach [19, 21].

Denote by $d(u)$ a prox-function of the set U. We assume that $d(u)$ is continuous and strongly convex on U with modulus $\sigma > 0$. Let u_0 be the center of the set U defined as

$$u_0 = \arg\min\{d(u) : \ u \in U\}. \tag{6}$$

Without loss of generality assume that $d(u_0) = 0$. We now describe Nesterov's smooth minimization approach [19, 21] for solving the dual problem (5), and we will show that it simultaneously solves the nonsmooth concave maximization problem (2).

SMOOTH MINIMIZATION ALGORITHM.

Let $u_0 \in U$ be given in (6). Set $x_{-1} = 0$ and $k = 0$.

(1) Compute $\nabla f(u_k)$ and $x(u_k)$. Set $x_k = \frac{k}{k+2}x_{k-1} + \frac{2}{k+2}x(u_k)$.

(2) Find $u_k^{sd} \in \mathrm{Argmin}\big\{\langle \nabla f(u_k), u - u_k\rangle + \frac{L}{2}\|u - u_k\|^2 : \ u \in U\big\}$.

(3) Find $u_k^{ag} = \mathrm{argmin}\big\{\frac{L}{\sigma}d(u) + \sum_{i=0}^{k}\frac{i+1}{2}[f(u_i) + \langle \nabla f(u_i), u - u_i\rangle] : \ u \in U\big\}$.

(4) Set $u_{k+1} = \frac{2}{k+3}u_k^{ag} + \frac{k+1}{k+3}u_k^{sd}$.

(5) Set $k \leftarrow k + 1$ and go to step 1.

end

The following property of the above algorithm is established in Theorem 2 of Nesterov [21].

THEOREM 2.1. *Let the sequence $\{(u_k, u_k^{sd})\}_{k=0}^{\infty} \subseteq U \times U$ be generated by the smooth minimization algorithm. Then for any $k \geq 0$ we have*

$$\frac{(k+1)(k+2)}{4}f(u_k^{sd}) \leq \min\left\{\frac{L}{\sigma}d(u) + \sum_{i=0}^{k}\frac{i+1}{2}[f(u_i) + \langle \nabla f(u_i), u - u_i\rangle] : \ u \in U\right\}. \tag{7}$$

We are ready to establish the main convergence result of the smooth minimization algorithm for solving the nonsmooth concave maximization problem (2) and its dual (5). Its proof is a generalization of the one given in a more special context in [21].

THEOREM 2.2. *After k iterations, the smooth minimization algorithm generates a pair of approximate solutions (u_k^{sd}, x_k) to problem* (2) *and its dual* (5)*, respectively, which satisfy the following inequality:*

$$0 \leq f(u_k^{sd}) - g(x_k) \leq \frac{4LD}{\sigma(k+1)(k+2)}. \tag{8}$$

Thus if the termination criterion $f(u_k^{sd}) - g(x_k) \leq \epsilon$ is applied, the iteration complexity of finding an ϵ-optimal solution to problem (2) *and its dual* (5) *by the smooth minimization algorithm does not exceed $2\sqrt{LD/(\sigma\epsilon)}$, where*

$$D = \max\{d(u) : \ u \in U\}. \tag{9}$$

Proof. In view of (3), (4), and the notation $x(u)$, we have

$$f(u_i) + \langle \nabla f(u_i), u - u_i\rangle = \phi(x(u_i), u_i) + \langle \nabla_u \phi(x(u_i), u_i), u - u_i\rangle. \tag{10}$$

Invoking the fact that the function $\phi(x, \cdot)$ is convex on U for every fixed $x \in X$, we obtain

$$\phi(x(u_i), u_i) + \langle \nabla_u \phi(x(u_i), u_i), u - u_i\rangle \leq \phi(x(u_i), u). \tag{11}$$

Notice that $x_{-1} = 0$, and $x_k = \frac{k}{k+2}x_{k-1} + \frac{2}{k+2}x(u_k)$ for any $k \geq 0$, which imply

$$x_k = \sum_{i=0}^{k} \frac{2(i+1)}{(k+1)(k+2)} x(u_i). \tag{12}$$

Using (10), (11), (12), and the fact that the function $\phi(\cdot, u)$ is concave on X for every fixed $u \in U$, we have

$$\begin{aligned}\sum_{i=0}^{k}(i+1)[f(u_i) + \langle \nabla f(u_i), u - u_i\rangle] &\leq \sum_{i=0}^{k}(i+1)\phi(x(u_i), u) \\ &\leq \frac{1}{2}(k+1)(k+2)\phi(x_k, u)\end{aligned}$$

for all $u \in U$. It follows from this relation, (7), (9), and (2) that

$$\begin{aligned}f(u_k^{sd}) &\leq \frac{4LD}{\sigma(k+1)(k+2)} \\ &\quad + \min_u \left\{ \sum_{i=0}^{k} \frac{2(i+1)}{(k+1)(k+2)}[f(u_i) + \langle \nabla f(u_i), u - u_i\rangle] : \ u \in U \right\} \\ &\leq \frac{4LD}{\sigma(k+1)(k+2)} + \min_{u \in U} \phi(x_k, u) = \frac{4LD}{\sigma(k+1)(k+2)} + g(x_k),\end{aligned}$$

and hence the inequality (8) holds. The remaining conclusion directly follows from (8). ∎

Remark. We shall mention that Nesterov [20] developed the excessive gap technique for solving problem (2) and its dual (5) in a special context, which enjoys the same iteration complexity as the smooth minimization algorithm described above. In addition, it is not hard to observe that the technique proposed in [20] can be extended to solve problem (2) and its dual (5) in the aforementioned general framework, provided that the subproblem

$$\min_{u \in U} \phi(x, u) + \mu d(u) \tag{13}$$

can be suitably solved for any given $\mu > 0$ and $x \in X$. The computation of each iterate of Nesterov's excessive gap technique [20] is similar to that of the smooth minimization algorithm except that the former method requires solving a prox subproblem in the form of (13), but the latter one needs to solve the prox subproblem described in step 3 above. When the function $\phi(x, \cdot)$ is affine for every fixed $x \in X$, these two prox subproblems have the same form, and thus the computational cost of Nesterov's excessive gap technique [20] is almost the same as that of the smooth minimization algorithm; however, for a more general function $\phi(\cdot, \cdot)$, the computational cost of the former method can be more expensive than that of the latter method.

The following results will be used to develop a variant of the smooth minimization algorithm for sparse covariance selection in subsection 3.4.

LEMMA 2.3. *Problem* (2) *has a unique optimal solution, denoted by* x^*. *Moreover, for any* $u^* \in \mathrm{Argmin}\{f(u) : \ u \in U\}$, *we have*

$$x^* = \arg\max_{x \in X} \phi(x, u^*). \tag{14}$$

Proof. We clearly know that problem (2) has an optimal solution. To prove its uniqueness, it suffices to show that $g(x)$ is strictly concave on X. Indeed, since $X \times U$ is a convex compact set and $\phi(x,u)$ is continuous on $X \times U$, it follows that for any $t \in (0,1)$, $x^1 \neq x^2 \in X$, there exist some $\tilde{u} \in U$ such that

$$\phi(tx^1 + (1-t)x^2, \tilde{u}) = \min_{u \in U} \phi(tx^1 + (1-t)x^2, u).$$

Recall that $\phi(\cdot, u)$ is strictly concave on X for every fixed $u \in U$. Therefore, we have

$$\begin{aligned}\phi(tx^1 + (1-t)x^2, \tilde{u}) &> t\phi(x^1, \tilde{u}) + (1-t)\phi(x^2, \tilde{u}),\\ &\geq t \min\nolimits_{u \in U} \phi(x^1, u) + (1-t) \min\nolimits_{u \in U} \phi(x^2, u),\end{aligned}$$

which together with (2) implies that

$$g(tx^1 + (1-t)x^2) > tg(x^1) + (1-t)g(x^2)$$

for any $t \in (0,1)$, $x^1 \neq x^2 \in X$, and hence $g(x)$ is strictly concave on X as desired.

Note that x^* is the optimal solution of problem (2). We clearly know that for any $u^* \in \mathrm{Argmin}\{f(u) : u \in U\}$, (u^*, x^*) is a saddle point for problem (2), that is,

$$\phi(x^*, u) \geq \phi(x^*, u^*) \geq \phi(x, u^*) \quad \forall (x,u) \in X \times U,$$

and hence we have

$$x^* \in \mathrm{Arg}\max_{x \in X} \phi(x, u^*).$$

This, together with the fact that $\phi(\cdot, u^*)$ is strictly concave on X, immediately yields (14). ∎

THEOREM 2.4. *Let x^* be the unique optimal solution of (2), and let f^* be the optimal value of problems (2) and (5). Assume that the sequences $\{u_k\}_{k=0}^{\infty}$ and $\{x(u_k)\}_{k=0}^{\infty}$ are generated by the smooth minimization algorithm. Then the following statements hold:*

(1) *$f(u_k) \to f^*$, $x(u_k) \to x^*$ as $k \to \infty$;*

(2) *$f(u_k) - g(x(u_k)) \to 0$ as $k \to \infty$.*

Proof. Recall from the smooth minimization algorithm that

$$u_{k+1} = \left(2u_k^{ag} + (k+1)u_k^{sd}\right)/(k+3) \ \ \forall k \geq 0.$$

Since $u_k^{sd}, u_k^{ag} \in U$ for all $k \geq 0$, and U is a compact set, we have $u_{k+1} - u_k^{sd} \to 0$ as $k \to \infty$. Notice that $f(u)$ is continuous on the compact set U, and hence it is uniformly continuous on U. Then we further have $f(u_{k+1}) - f(u_k^{sd}) \to 0$ as $k \to \infty$. Also, it follows from Theorem 2.2 that $f(u_k^{sd}) \to f^*$ as $k \to \infty$. Therefore, we conclude that $f(u_k) \to f^*$ as $k \to \infty$.

Note that X is a compact set, and $x(u_k) \subseteq X$ for all $k \geq 0$. To prove that $x(u_k) \to x^*$ as $k \to \infty$, it suffices to show that every convergent subsequence of $\{x(u_k)\}_{k=0}^{\infty}$ converges to x^* as $k \to \infty$. Indeed, assume that $\{x(u_{n_k})\}_{k=0}^{\infty}$ is an arbitrary convergent subsequence, and $x(u_{n_k}) \to \tilde{x}^*$ as $k \to \infty$ for some $\tilde{x}^* \in X$. Without loss of generality, assume that the sequence $\{u_{n_k}\}_{k=0}^{\infty} \to \tilde{u}^*$ as $k \to \infty$ for some $\tilde{u}^* \in U$ (otherwise, one can consider any convergent subsequence of $\{u_{n_k}\}_{k=0}^{\infty}$). Using the result that $f(u_k) \to f^*$, we obtain that

$$\phi\left(x(u_{n_k}), u_{n_k}\right) = f(u_{n_k}) \to f^* \quad \text{as } k \to \infty.$$

Upon letting $k \to \infty$ and using the continuity of $\phi(\cdot,\cdot)$, we have $\phi(\tilde{x}^*, \tilde{u}^*) = f(\tilde{u}^*) = f^*$. Hence, it follows that

$$\tilde{u}^* \in \operatorname{Arg}\min_{u \in U} f(u), \quad \tilde{x}^* = \arg\max_{x \in X} \phi(x, \tilde{u}^*),$$

which together with Lemma 2.3 implies that $\tilde{x}^* = x^*$. Hence, as desired, $x(u_{n_k}) \to x^*$ as $k \to \infty$.

As shown in Lemma 2.3, the function $g(x)$ is continuous on X. This result, together with statement 1, immediately implies that statement 2 holds. □

3. Sparse covariance selection. In this section, we discuss the application of the smooth optimization approach proposed in section 2 to sparse covariance selection. More specifically, we briefly introduce sparse covariance selection in subsection 3.1 and show that it can be approximately solved as an l_1-norm penalized maximum likelihood estimation problem in subsection 3.2. In subsection 3.3, we address some implementation details of the smooth optimization approach for solving this problem and propose a variant of this approach in subsection 3.4.

3.1. Introduction of sparse covariance selection. In this subsection, we briefly introduce sparse covariance selection. For more details, see d'Aspremont, Banerjee, and El Ghaoui [10] and the references therein.

Given n variables with a Gaussian distribution $\mathcal{N}(0, C)$ for which the true covariance matrix C is unknown, we are interested in estimating C from a sample covariance matrix Σ by maximizing its likelihood while imposing a certain number of components in the inverse of the estimation of C to zero. This problem is commonly known as *sparse covariance selection* (see [11]). Since zeros in the inverse of covariance matrix correspond to conditional independence in the model, sparse covariance selection can be used to determine a robust estimate of the covariance matrix, and simultaneously discover the sparse structure in the underlying graphical model.

Several approaches have been proposed for sparse covariance selection in literature. For example, Bilmes [4] proposed a method based on choosing statistical dependencies according to conditional mutual information computed from training data. The recent works [18, 12] involve identifying the Gaussian graphical models that are best supported by the data and any available prior information on the covariance matrix. Given a sample covariance matrix $\Sigma \in \mathcal{S}_+^n$, d'Aspremont et al. [10] recently formulated sparse covariance selection as the following estimation problem:

$$\begin{array}{ll} \max_X & \log\det X - \langle \Sigma, X\rangle - \rho \mathrm{Card}(X) \\ \text{s.t.} & \tilde{\alpha} I \preceq X \preceq \tilde{\beta} I, \end{array} \tag{15}$$

where $\rho > 0$ is a parameter controlling the trade-off between likelihood and cardinality, and $0 \le \tilde{\alpha} < \tilde{\beta} \le \infty$ are the fixed bounds on the eigenvalues of the solution. For some specific choices of ρ, the formulation (15) has been used for model selection in [1, 5], and applied to speech recognition and gene network analysis (see [4, 13]).

Note that the estimation problem (15) itself is an NP-hard combinatorial problem because of the penalty term $\mathrm{Card}(X)$. To overcome the computational difficulty, d'Aspremont et al. [10] used an argument that is often used in regression techniques (e.g., see [23, 6, 14]), where sparsity of the solution is concerned, to relax $\mathrm{Card}(X)$ to $e^T|X|e$, and obtained the following l_1-norm penalized maximum likelihood estimation

problem:

$$\begin{array}{ll} \max_X & \log\det X - \langle \Sigma, X\rangle - \rho e^T|X|e \\ \text{s.t.} & \tilde{\alpha} I \preceq X \preceq \tilde{\beta} I. \end{array} \tag{16}$$

Recently, Yuan and Lin [25] proposed a similar estimation problem for sparse covariance selection given as follows:

$$\begin{array}{ll} \max_X & \log\det X - \langle \Sigma, X\rangle - \rho \sum_{i\neq j} |X_{ij}| \\ \text{s.t.} & \tilde{\alpha} I \preceq X \preceq \tilde{\beta} I, \end{array} \tag{17}$$

with $\tilde{\alpha} = 0$ and $\tilde{\beta} = \infty$. They showed that problem (17) can be suitably solved by the interior point algorithm developed in Vandenberghe, Boyd, and Wu [24]. A few other approaches have also been studied for sparse covariance selection by solving some related maximum likelihood estimation problems in the literature. For example, Huang, Liu, and Pourahmadi [17] proposed an iterative (heuristic) algorithm to minimize a nonconvex penalized likelihood. Dahl et al. [8, 7] applied Newton's method, the coordinate steepest descent method, and the conjugate gradient method for the problems for which the conditional independence structure is partially known.

As shown in d'Aspremont et al. [10] (see also [3]) and Yuan and Lin [25], the l_1-norm penalized maximum likelihood estimation problems (16) and (17) are capable of discovering effectively the sparse structure or, equivalently, the conditional independence in the underlying graphical model. Also, it is not hard to see that the estimation problem (17) becomes a special case of problem (16) if replacing Σ by $\Sigma + \rho I$ in (17). For these reasons, in the remainder of this paper we focus on problem (16) only.

3.2. Nonsmooth strictly concave maximization reformulation. In this subsection, we show that problem (16) can be reformulated as a nonsmooth strictly concave maximization problem of the form (2).

Recall from subsection 3.1 that $\Sigma \in \mathcal{S}^n_+$, and keep in mind that the notation $|\cdot|$, $\|\cdot\|$, and $\|\cdot\|_F$ is defined in subsection 1.1. We first provide some tighter bounds on the optimal solution of problem (16) for the case where $\tilde{\alpha} = 0$ and $\tilde{\beta} = \infty$.

PROPOSITION 3.1. *Assume that $\tilde{\alpha} = 0$ and $\tilde{\beta} = \infty$. Let $X^* \in \mathcal{S}^n_{++}$ be the unique optimal solution of problem* (16)*. Then we have $\alpha I \preceq X^* \preceq \beta I$, where*

$$\alpha = \frac{1}{\|\Sigma\| + n\rho}, \quad \beta = \min\left\{\frac{n - \alpha \mathrm{Tr}(\Sigma)}{\rho}, \eta\right\} \tag{18}$$

with

$$\eta = \begin{cases} \min\left\{e^T|\Sigma^{-1}|e, (n - \rho\sqrt{n}\alpha)\|\Sigma^{-1}\| - (n-1)\alpha\right\} & \text{if } \Sigma \text{ is invertible;} \\ 2e^T|(\Sigma + \frac{\rho}{2}I)^{-1}|e - \mathrm{Tr}((\Sigma + \frac{\rho}{2}I)^{-1}) & \text{otherwise.} \end{cases}$$

Proof. Let

$$\mathcal{U} := \{U \in \mathcal{S}^n : \ |U_{ij}| \leq 1 \quad \forall ij\} \tag{19}$$

and

$$L(X, U) = \log\det X - \langle \Sigma + \rho U, X\rangle \quad \forall (X, U) \in \mathcal{S}^n_{++} \times \mathcal{U}. \tag{20}$$

Note that $X^* \in \mathcal{S}^n_{++}$ is the optimal solution of problem (16). It can be easily shown that there exist some $U^* \in \mathcal{U}$ such that (X^*, U^*) is a saddle point of $L(\cdot,\cdot)$ on $\mathcal{S}^n_{++} \times \mathcal{U}$, that is,

$$X^* = \arg\min_{X \in \mathcal{S}^n_{++}} L(X, U^*), \quad U^* \in \operatorname{Arg}\min_{U \in \mathcal{U}} L(X^*, U).$$

The above relations along with (19) and (20) immediately yield

$$X^*(\Sigma + \rho U^*) = I, \quad \langle X^*, U^* \rangle = e^T |X^*| e. \tag{21}$$

Hence, we have

$$X^* = (\Sigma + \rho U^*)^{-1} \succeq \frac{1}{\|\Sigma\| + \rho \|U^*\|} I,$$

which together with (19) and the fact $U^* \in \mathcal{U}$ implies that $X^* \succeq \frac{1}{\|\Sigma\| + n\rho} I$. Thus, as desired, $X^* \succeq \alpha I$, where α is given in (18).

We next bound X^* from above. In view of (21), we have

$$\langle X^*, \Sigma \rangle + \rho e^T |X^*| e = n, \tag{22}$$

which together with the relation $X^* \succeq \alpha I$ implies that

$$e^T |X^*| e \le \frac{n - \alpha \mathrm{Tr}(\Sigma)}{\rho}. \tag{23}$$

Now let $X(t) := (\Sigma + t\rho I)^{-1}$ for $t \in (0,1)$. By concavity of $\log\det(\cdot)$, one can easily see that $X(t)$ maximizes the function $\log\det(\cdot) - \langle \Sigma + t\rho I, \cdot \rangle$ over $\mathcal{S}^n_{++}$. Using this observation and the definition of X^*, we can have

$$\log\det X^* - \langle \Sigma + t\rho I, X^* \rangle \le \log\det X(t) - \langle \Sigma + t\rho I, X(t) \rangle,$$

$$\log\det X(t) - \langle \Sigma, X(t) \rangle - \rho e^T |X(t)| e \le \log\det X^* - \langle \Sigma, X^* \rangle - \rho e^T |X^*| e.$$

Adding the above two inequalities upon some algebraic simplification, we obtain that

$$e^T |X^*| e - t\mathrm{Tr}(X^*) \le e^T |X(t)| e - t\mathrm{Tr}(X(t)),$$

and hence

$$e^T |X^*| e \le \frac{e^T |X(t)| e - t\mathrm{Tr}(X(t))}{1 - t} \quad \forall t \in (0,1). \tag{24}$$

If Σ is invertible, upon letting $t \downarrow 0$ on both sides of (24), we have

$$e^T |X^*| e \le e^T |\Sigma^{-1}| e.$$

Otherwise, letting $t = 1/2$ in (24), we obtain

$$e^T |X^*| e \le 2e^T \left| \left(\Sigma + \frac{\rho}{2} I \right)^{-1} \right| e - \mathrm{Tr}\left(\left(\Sigma + \frac{\rho}{2} I \right)^{-1} \right).$$

Combining the above two inequalities and (23), we have

$$\|X^*\| \le \|X^*\|_F \le e^T |X^*| e \le \min\left\{ \frac{n - \alpha \mathrm{Tr}(\Sigma)}{\rho}, \gamma \right\}, \tag{25}$$

where

$$\gamma = \begin{cases} e^T|\Sigma^{-1}|e & \text{if } \Sigma \text{ is invertible;} \\ 2e^T|(\Sigma + \frac{\rho}{2}I)^{-1}|e - \mathrm{Tr}((\Sigma + \frac{\rho}{2}I)^{-1}) & \text{otherwise.} \end{cases}$$

Further, using the relation $X^* \succeq \alpha I$, we obtain that

$$e^T|X^*|e \geq \|X^*\|_F \geq \sqrt{n}\alpha,$$

which together with (22) implies that

$$\mathrm{Tr}(X^*\Sigma) \leq n - \rho\sqrt{n}\alpha.$$

This inequality, along with the relation $X^* \succeq \alpha I$, yields

$$\lambda_{\min}(\Sigma)((n-1)\alpha + \|X^*\|) \leq \mathrm{Tr}(X^*\Sigma) \leq n - \rho\sqrt{n}\alpha.$$

Hence if Σ is invertible, we further have

$$\|X^*\| \leq (n - \rho\sqrt{n}\alpha)\|\Sigma^{-1}\| - (n-1)\alpha.$$

This together with (25) implies that $X^* \preceq \beta I$, where β is given in (18). □

Remark. Some bounds on X^* were also derived in d'Aspremont, Banerjee, and El Ghaoui [10]. In contrast with their bounds, our bounds given in (18) are tighter. Moreover, our approach for deriving the above bounds can be generalized to handle the case where $\tilde{\alpha} > 0$ and $\tilde{\beta} = \infty$, but their approach cannot. Indeed, if $\tilde{\alpha} > 0$ and $\tilde{\beta} = \infty$, we can set $\alpha = \tilde{\alpha}$ and replace the above $X(t)$ by the optimal solution of

$$\begin{array}{ll} \max_X & \log\det X - \langle \Sigma + \rho I, X\rangle \\ \text{s.t.} & \tilde{\alpha} I \preceq X, \end{array}$$

which has a closed-form expression. By following a similar derivation as above, one can obtain a positive scalar β such that $X^* \preceq \beta I$. In addition, for the case where $\tilde{\alpha} = 0$ and $0 < \tilde{\beta} < \infty$, one can set $\beta = \tilde{\beta}$ and easily show that $X^* \geq \alpha I$, where $\alpha = \beta e^{-\beta(\mathrm{Tr}(\Sigma)+n\rho)}$.

From the above discussion, we conclude that problem (16) is equivalent to the following problem:

$$\begin{array}{ll} \max_X & \log\det X - \langle \Sigma, X\rangle - \rho e^T|X|e \\ \text{s.t.} & \alpha I \preceq X \preceq \beta I \end{array} \tag{26}$$

for some $0 < \alpha < \beta < \infty$.

We further observe that problem (26) can be rewritten as

$$\max_{X\in\mathcal{X}} \min_{U\in\mathcal{U}} \ \log\det X - \langle \Sigma + \rho U, X\rangle, \tag{27}$$

where $\mathcal{U}$ is defined in (19), and $\mathcal{X}$ is defined as follows:

$$\mathcal{X} := \{X \in \mathcal{S}^n : \ \alpha I \preceq X \preceq \beta I\}. \tag{28}$$

Therefore, we conclude that problem (16) is equivalent to (27). For the remainder of the paper, we will focus on problem (27) only.

3.3. Smooth optimization method for sparse covariance selection. In this subsection, we describe the implementation details of the smooth minimization algorithm proposed in section 2 for solving problem (27). We also compare the complexity of this algorithm with interior point methods, and two other first-order methods studied in d'Aspremont et al. [10], that is, Nesterov's smooth approximation scheme and block-coordinate descent method.

We first observe that the sets $\mathcal{X}$ and $\mathcal{U}$ both lie in the space $\mathcal{S}^n$, where $\mathcal{X}$ and $\mathcal{U}$ are defined in (28) and (19), respectively. Let $\mathcal{S}^n$ be endowed with the Frobenius norm, and let $\tilde{d}(X) = \log\det X$ for $X \in \mathcal{X}$. Then for any $X \in \mathcal{X}$, we have

$$\nabla^2 \tilde{d}(X)[H,H] = -\mathrm{Tr}(X^{-1}HX^{-1}H) \le -\beta^{-2}\|H\|_F^2$$

for all $H \in \mathcal{S}^n$, and hence, $\tilde{d}(X)$ is strongly concave on $\mathcal{X}$ with modulus β^{-2}. Using this result and Theorem 1 of [21], we immediately conclude that $\nabla f(U)$ is Lipschitz continuous with constant $L = \rho^2\beta^2$ on $\mathcal{U}$, where

$$f(U) := \max_{X\in\mathcal{X}} \log\det X - \langle \Sigma + \rho U, X\rangle \quad \forall U \in \mathcal{U}. \tag{29}$$

Denote the unique optimal solution of problem (29) by $X(U)$. For any $U \in \mathcal{U}$, we can compute $X(U)$, $f(U)$, and $\nabla f(U)$ as follows.

Let $\Sigma + \rho U = Q\mathrm{diag}(\gamma)Q^T$ be an eigenvalue decomposition of $\Sigma + \rho U$ such that $QQ^T = I$. For $i = 1,\dots,n$, let

$$\lambda_i = \begin{cases} \min\{\max\{1/\gamma_i, \alpha\}, \beta\} & \text{if } \gamma_i > 0; \\ \beta & \text{otherwise.} \end{cases}$$

It is not hard to show that

$$X(U) = Q\mathrm{diag}(\lambda)Q^T, \quad f(U) = -\gamma^T\lambda + \sum_{i=1}^n \log\lambda_i, \quad \nabla f(U) = -\rho X(U). \tag{30}$$

From the above discussion, we see that problem (27) has exactly the same form as (2) and also satisfies all assumptions imposed on problem (2). Therefore, it can be suitably solved by the smooth minimization algorithm proposed in section 2. The implementation details of this algorithm for problem (27) are described as follows.

Given $U_0 \in \mathcal{U}$, let $d(U) = \|U - U_0\|_F^2/2$ be the proximal function on $\mathcal{U}$, which is strongly convex function with modulus $\sigma = 1$. For our specific choice of the norm and $d(U)$, we clearly see that steps 2 and 3 of the smooth minimization algorithm can be solved as a problem of the form

$$V = \arg\min_{U\in\mathcal{U}} \langle G, U\rangle + \|U\|_F^2/2$$

for some $G \in \mathcal{S}^n$. In view of (19), we see that

$$V_{ij} = \max\{\min\{-G_{ij}, 1\}, -1\}, \quad i,j = 1,\dots,n.$$

In addition, for any $X \in \mathcal{X}$, we define

$$g(X) := \log\det X - \langle \Sigma, X\rangle - \rho e^T|X|e. \tag{31}$$

For the ease of comparison with its latter variant, we now present a complete version of the aforementioned smooth minimization algorithm for solving problem (27) and its dual.

SMOOTH MINIMIZATION ALGORITHM FOR COVARIANCE SELECTION (SMACS).
Let $\epsilon > 0$ and $U_0 \in \mathcal{U}$ be given. Set $X_{-1} = 0$, $L = \rho^2\beta^2$, $\sigma = 1$, and $k = 0$.
(1) Compute $\nabla f(U_k)$ and $X(U_k)$. Set $X_k = \frac{k}{k+2}X_{k-1} + \frac{2}{k+2}X(U_k)$.
(2) Find $U_k^{sd} = \mathrm{argmin}\left\{\langle \nabla f(U_k), U - U_k\rangle + \frac{L}{2}\|U - U_k\|_F^2 : U \in \mathcal{U}\right\}$.
(3) Find $U_k^{ag} = \mathrm{argmin}\{\frac{L}{2\sigma}\|U - U_0\|_F^2 + \sum_{i=0}^{k}\frac{i+1}{2}[f(U_i) + \langle \nabla f(U_i), U - U_i\rangle] : U \in \mathcal{U}\}$.
(4) Set $U_{k+1} = \frac{2}{k+3}U_k^{ag} + \frac{k+1}{k+3}U_k^{sd}$.
(5) Set $k \leftarrow k + 1$. Go to step 1 until $f(U_k^{sd}) - g(X_k) \le \epsilon$.
end

The iteration complexity of the above algorithm for solving problem (27) is established in the following theorem.

THEOREM 3.2. *The iteration complexity performed by the algorithm SMACS for finding an ϵ-optimal solution to problem* (27) *and its dual does not exceed* $\sqrt{2}\rho\beta \max_{U\in\mathcal{U}}\|U - U_0\|_F/\sqrt{\epsilon}$, *and moreover, if* $U_0 = 0$, *it does not exceed* $\sqrt{2}\rho\beta n/\sqrt{\epsilon}$.

Proof. From the above discussion, we know that $L = \rho^2\beta^2$, $D = \max_{U\in\mathcal{U}}\|U - U_0\|_F^2/2$, and $\sigma = 1$, which together with Theorem 2.2 immediately implies that the first part of the statement holds. Further, if $U_0 = 0$, we easily obtain from (19) that $D = \max_{U\in\mathcal{U}}\|U\|_F^2/2 = n^2/2$. The second part of the statement directly follows from this result and Theorem 2.2. ∎

Remark. By the definition of $\mathcal{U}$ (see (19)), we can easily show that $\min_{U_0\in\mathcal{U}}\max_{U\in\mathcal{U}}\|U - U_0\|_F$ has a unique minimizer $U_0 = 0$. This result together with Theorem 3.2 implies that the initial point $U_0 = 0$ gives the optimal worst-case iteration complexity for the algorithm SMACS.

Alternatively, d'Aspremont et al. [10] applied Nesterov's smooth approximation scheme [21] to solve problem (27). More specifically, let $\epsilon > 0$ be the desired accuracy, and let

$$\hat{d}(U) = \|U\|_F^2/2, \quad \hat{D} = \max_{U\in\mathcal{U}}\hat{d}(U) = n^2/2.$$

As shown in [21], the nonsmooth function $g(X)$ defined in (31) is uniformly approximated by the smooth function

$$g_\epsilon(X) = \min_{U\in\mathcal{U}} \log\det X - \langle \Sigma + \rho U, X\rangle - \frac{\epsilon}{2\hat{D}}\hat{d}(U)$$

on $\mathcal{X}$ with the error at most by $\epsilon/2$, and, moreover, the function $g_\epsilon(X)$ has a Lipschitz continuous gradient on $\mathcal{X}$ with some constant $L(\epsilon) > 0$. Nesterov's smooth optimization technique [19, 21] is then applied to solve the perturbed problem $\max_{X\in\mathcal{X}} g_\epsilon(X)$, and problem (27) is accordingly solved. It was shown in [10] that the iteration complexity of this approach for finding an ϵ-optimal solution to problem (27) does not exceed

$$\frac{2\sqrt{2}\rho\beta n^{1.5}\log\kappa}{\epsilon} + \kappa\sqrt{\frac{n\log\kappa}{\epsilon}}, \tag{32}$$

where $\kappa := \beta/\alpha$.

In view of (32) and Theorem 3.2, we conclude that the smooth optimization approach improves upon Nesterov's smooth approximation scheme at least by a factor of $\mathcal{O}(\sqrt{n}\log\kappa/\sqrt{\epsilon})$ in terms of the iteration complexity for solving problem (27). Moreover, the computational cost per iteration of the former approach is at least as cheap as that of the latter one.

d'Aspremont et al. [10] also studied a block-coordinate descent method for solving problem (16) with $\tilde{\alpha} = 0$ and $\tilde{\beta} = \infty$. Each iterate of this method requires computing the inverse of an $(n-1) \times (n-1)$ matrix and solving a box constrained quadratic programming with $n-1$ variables. As mentioned in section 3 of [10], this method has a local linear convergence rate. However, its global iteration complexity for finding an ϵ-optimal solution is theoretically unknown. Moreover, this method is not suitable for solving problem (16) with $\tilde{\alpha} > 0$ or $\tilde{\beta} < \infty$.

In addition, we observe that problem (26) (also (16)) can be reformulated as a constrained smooth convex problem that has an explicit $\mathcal{O}(n^2)$-logarithmically homogeneous self-concordant barrier function. Thus, it can be suitably solved by interior point methods (see Nesterov and Nemirovskii [22] and Vandenberghe, Boyd, and Wu [24]). The worst-case iteration complexity of interior point methods for finding an ϵ-optimal solution to (26) is $\mathcal{O}(n \log(\epsilon_0/\epsilon))$, where ϵ_0 is an initial gap. Each iterate of interior point methods requires $\mathcal{O}(n^6)$ arithmetic cost for assembling and solving a typically dense Newton system with $\mathcal{O}(n^2)$ variables. Thus, the total worst-case arithmetic cost of interior point methods for finding an ϵ-optimal solution to (26) is $\mathcal{O}(n^7 \log(\epsilon_0/\epsilon))$. In contrast to interior point methods, the algorithm SMACS requires $\mathcal{O}(n^3)$ arithmetic cost per iteration dominated by eigenvalue decomposition and matrix multiplication of $n \times n$ matrices. Based on this observation and Theorem 3.2, we conclude that the overall worst-case arithmetic cost of the algorithm SMACS for finding an ϵ-optimal solution to (26) is $\mathcal{O}(\rho\beta n^4/\sqrt{\epsilon})$, which is substantially superior to that of interior point methods, provided that $\rho\beta$ is not too large and ϵ is not too small.

3.4. Variant of the smooth minimization algorithm. As discussed in subsection 3.3, the algorithm SMACS has a nice theoretical complexity in contrast with interior point methods, Nesterov's smooth approximation scheme, and the block-coordinate descent method. However, its practical performance is still not very attractive (see section 4). To enhance the computational performance, we propose a variant of the algorithm SMACS for solving problem (27) in this subsection.

Our first concern of the algorithm SMACS is that the eigenvalue decomposition of two $n \times n$ matrices is required per iteration. Indeed, the eigenvalue decomposition of $\Sigma + \rho U_k$ and $\Sigma + \rho U_k^{sd}$ is needed at steps 1 and 5 to compute $\nabla f(U_k)$ and $f(U_k^{sd})$, respectively. We also know that the eigenvalue decomposition is one of major computations for the algorithm SMACS. To reduce the computational cost, we now propose a new termination criterion other than $f(U_k^{sd}) - g(X_k) \le \epsilon$ that is used in the algorithm SMACS. In view of Theorem 2.4, we know that

$$f(U_k) - g(X(U_k)) \to 0 \text{ as } k \to \infty.$$

Thus, $f(U_k) - g(X(U_k)) \le \epsilon$ can be used as an alternative termination criterion. Moreover, it follows from (30) that the quantity $f(U_k) - g(X(U_k))$ is readily available in step 1 of the algorithm SMACS with almost no additional cost. We easily see that the algorithm SMACS with this new termination criterion would require only one eigenvalue decomposition per iteration. Despite this clear advantage, we shall mention that the iteration complexity of the resulting algorithm is unfortunately unknown. Nevertheless, in practice we have found that the number of iterations performed by the algorithm SMACS with the above two different termination criteria are almost same. Thus, $f(u_k) - g(x(u_k)) \le \epsilon$ is a useful practical termination criterion.

For sparse covariance selection, the penalty parameter ρ is usually small, but the parameter β can be fairly large. In view of Theorem 3.2, we know that the iteration

complexity of the algorithm SMACS for solving problem (27) is proportional to β. Therefore, when β is too large, the complexity and practical performance of this algorithm become unattractive. To overcome this drawback, we will propose one strategy to dynamically update β.

Let X^* be the unique optimal solution of problem (27). For any $\hat{\beta} \in [\lambda_{\max}(X^*), \beta]$, we easily observe that X^* is also the unique optimal solution to the following problem:

$$(P_{\hat{\beta}}) \qquad \max_{X \in \mathcal{X}_{\hat{\beta}}} \min_{U \in \mathcal{U}} \log\det X - \langle \Sigma + \rho U, X\rangle, \tag{33}$$

where $\mathcal{U}$ is defined in (19) and $\mathcal{X}_{\hat{\beta}}$ is given by

$$\mathcal{X}_{\hat{\beta}} := \{X : \ \alpha I \preceq X \preceq \hat{\beta} I\}.$$

In view of Theorem 3.2, the iteration complexity of the algorithm SMACS for problem (33) is lower than that for problem (27), provided $\hat{\beta} \in [\lambda_{\max}(X^*), \beta)$. Hence, ideally we set $\hat{\beta} = \lambda_{\max}(X^*)$, which would give the lowest iteration complexity, but unfortunately $\lambda_{\max}(X^*)$ is unknown. However, we can generate a sequence $\{\hat{\beta}_k\}_{k=0}^{\infty}$ that asymptotically approaches $\lambda_{\max}(X^*)$ as the algorithm progresses. Indeed, in view of Theorem 2.4, we know that $X(U_k) \to X^*$ as $k \to \infty$, and we obtain that

$$\lambda_{\max}(X(U_k)) \to \lambda_{\max}(X^*) \text{ as } k \to \infty.$$

Therefore, we see that $\{\lambda_{\max}(X(U_k))\}_{k=0}^{\infty}$ can be used to generate a sequence $\{\hat{\beta}_k\}_{k=0}^{\infty}$ that asymptotically approaches $\lambda_{\max}(X^*)$. We next propose a strategy to generate such a sequence $\{\hat{\beta}_k\}_{k=0}^{\infty}$.

For convenience of presentation, we introduce some new notation. Given any $U \in \mathcal{U}$ and $\hat{\beta} \in [\alpha, \beta]$, we define

$$X_{\hat{\beta}}(U) := \arg\max_{X \in \mathcal{X}_{\hat{\beta}}} \log\det X - \langle \Sigma + \rho U, X\rangle, \tag{34}$$

$$f_{\hat{\beta}}(U) := \max_{X \in \mathcal{X}_{\hat{\beta}}} \log\det X - \langle \Sigma + \rho U, X\rangle. \tag{35}$$

Definition 1. *Given any $U \in \mathcal{U}$ and $\hat{\beta} \in [\alpha, \beta]$, $X_{\hat{\beta}}(U)$ is called "active" if $\lambda_{\max}(X_{\hat{\beta}}(U)) = \hat{\beta}$ and $\hat{\beta} < \beta$; otherwise it is called "inactive."*

Let ς_1, $\varsigma_2 > 1$, and $\varsigma_3 \in (0, 1)$ be given and fixed. Assume that $U_k \in \mathcal{U}$ and $\hat{\beta}_k \in [\alpha, \beta]$ are given at the beginning of the kth iteration for some $k \geq 0$. We now describe the strategy for generating the next iterate U_{k+1} and $\hat{\beta}_{k+1}$ by considering the following three different cases:

(1) If $X_{\hat{\beta}_k}(U_k)$ is active, find the smallest $s \in \mathcal{Z}_+$ such that $X_{\bar{\beta}}(U_k)$ is inactive, where $\bar{\beta} = \min\{\varsigma_1^s \hat{\beta}_k, \beta\}$. Set $\hat{\beta}_{k+1} = \bar{\beta}$, and apply the algorithm SMACS for problem $(P_{\hat{\beta}_{k+1}})$ starting with the point U_k and set its next iterate to be U_{k+1}.

(2) If $X_{\hat{\beta}_k}(U_k)$ is inactive and $\lambda_{\max}(X_{\hat{\beta}_k}(U_k)) \leq \varsigma_3 \hat{\beta}_k$, set $\hat{\beta}_{k+1} = \max\{\min\{\varsigma_2 \lambda_{\max}(X_{\hat{\beta}_k}(U_k)), \beta\}, \alpha\}$. Apply the algorithm SMACS for problem $(P_{\hat{\beta}_{k+1}})$ starting with the point U_k, and set its next iterate to be U_{k+1}.

(3) If $X_{\hat{\beta}_k}(U_k)$ is inactive and $\lambda_{\max}(X_{\hat{\beta}_k}(U_k)) > \varsigma_3 \hat{\beta}_k$, set $\hat{\beta}_{k+1} = \hat{\beta}_k$. Continue the algorithm SMACS for problem $(P_{\hat{\beta}_k})$, and set its next iterate to be U_{k+1}.

For the sequences $\{U_k\}_{k=0}^{\infty}$ and $\{\hat{\beta}_k\}_{k=0}^{\infty}$ recursively generated above, we observe that the sequence $\{X_{\hat{\beta}_{k+1}}(U_k)\}_{k=0}^{\infty}$ is always inactive. This together with (34), (35), (29), and the fact that $\hat{\beta}_k \le \beta$ for $k \ge 0$ implies that

$$f(U_k) = f_{\hat{\beta}_{k+1}}(U_k), \quad \nabla f(U_k) = \nabla f_{\hat{\beta}_{k+1}}(U_k) \ \ \forall k \ge 0. \tag{36}$$

Therefore, the new termination criterion $f(U_k) - g(X(U_k)) \le \epsilon$ can be replaced by

$$f_{\hat{\beta}_{k+1}}(U_k) - g(X_{\hat{\beta}_{k+1}}(U_k)) \le \epsilon \tag{37}$$

accordingly.

We now incorporate into the algorithm SMACS the new termination criterion (37) and the aforementioned strategy for generating a sequence $\{\hat{\beta}_k\}_{k=0}^{\infty}$ that asymptotically approaches $\lambda_{\max}(X^*)$, and we obtain a variant of the algorithm SMACS for solving problem (27). For convenience of presentation, we omit the subscript k from $\hat{\beta}_k$.

VARIANT OF THE SMOOTH MINIMIZATION ALGORITHM FOR COVARIANCE SELECTION (VSMACS).

Let $\epsilon > 0$, ς_1, $\varsigma_2 > 1$, and $\varsigma_3 \in (0,1)$ be given. Choose a $U_0 \in \mathcal{U}$. Set $\hat{\beta} = \beta$, $L = \rho^2\beta^2$, $\sigma = 1$, and $k = 0$.

(1) Compute $X_{\hat{\beta}}(U_k)$ according to (30).
  (1a) If $X_{\hat{\beta}}(U_k)$ is active, find the smallest $s \in \mathcal{Z}_+$ such that $X_{\bar{\beta}}(U_k)$ is inactive, where $\bar{\beta} = \min\{\varsigma_1^s\hat{\beta}, \beta\}$. Set $k = 0$, $U_0 = U_k$, $\hat{\beta} = \bar{\beta}$, $L = \rho^2\hat{\beta}^2$, and go to step 2.
  (1b) If $X_{\hat{\beta}}(U_k)$ is inactive and $\lambda_{\max}(X_{\hat{\beta}}(U_k)) \le \varsigma_3\hat{\beta}$, set $k = 0$, $U_0 = U_k$, $\hat{\beta} = \max\{\min\{\varsigma_2\lambda_{\max}(X_{\hat{\beta}}(U_k)), \beta\}, \alpha\}$, and $L = \rho^2\hat{\beta}^2$.
(2) If $f_{\hat{\beta}}(U_k) - g(X_{\hat{\beta}}(U_k)) \le \epsilon$, terminate. Otherwise, compute $\nabla f_{\hat{\beta}}(U_k)$ according to (30).
(3) Find $U_k^{sd} = \operatorname{argmin}\{\langle \nabla f_{\hat{\beta}}(U_k), U - U_k\rangle + \frac{L}{2}\|U - U_k\|_F^2 : \ U \in \mathcal{U}\}$.
(4) Find $U_k^{ag} = \operatorname{argmin}\{\frac{L}{2\sigma}\|U - U_0\|_F^2 + \sum_{i=0}^{k}\frac{i+1}{2}[f_{\hat{\beta}}(U_i) + \langle \nabla f_{\hat{\beta}}(U_i), U - U_i\rangle] : U \in \mathcal{U}\}$.
(5) Set $U_{k+1} = \frac{2}{k+3}U_k^{ag} + \frac{k+1}{k+3}U_k^{sd}$.
(6) Set $k \leftarrow k+1$, and go to step 1.

end

We next establish some preliminary convergence properties of the above algorithm.

PROPOSITION 3.3. *For the algorithm VSMACS, the following properties hold:*

(1) *Suppose that the algorithm VSMACS terminates at some iterate* $(X_{\hat{\beta}}(U_k), U_k)$. *Then* $(X_{\hat{\beta}}(U_k), U_k)$ *is an* ϵ*-optimal solution to problem* (27) *and its dual.*
(2) *Suppose that* $\hat{\beta}$ *is updated only for a finite number of times. Then the algorithm VSMACS terminates in a finite number of iterations and produces an* ϵ*-optimal solution to problem* (27) *and its dual.*

Proof. For the final iterate $(X_{\hat{\beta}}(U_k), U_k)$, we clearly know that $f_{\hat{\beta}}(U_k) - g(X_{\hat{\beta}}(U_k)) \le \epsilon$, and $X_{\hat{\beta}}(U_k)$ is inactive. As shown in (36), $f(U_k) = f_{\hat{\beta}}(U_k)$. Hence, we have $f(U_k) - g(X_{\hat{\beta}}(U_k)) \le \epsilon$. We also know that $U_k \in \mathcal{U}$, and $X_{\hat{\beta}}(U_k) \in \mathcal{X}$ due to $\hat{\beta} \in [\alpha, \beta]$. Thus, statement 1 immediately follows. After the last update of $\hat{\beta}$, the algorithm VSMACS behaves exactly like the algorithm SMACS as applied to solve problem $(P_{\hat{\beta}})$ except with the termination criterion $f(U_k) - g(X_{\hat{\beta}}(U_k)) \le \epsilon$. Thus, it

follows from statement 1 and Theorem 2.4 that statement 2 holds. Thus, it follows from statement 1 and Theorem 2.4 that statement 2 holds. □

4. Computational results. In this section, we compare the performance of the smooth minimization approach and its variant proposed in this paper with other first-order methods studied in [10, 16], that is, Nesterov's smooth approximation scheme and block-coordinate descent method for solving problem (16) (or, equivalently, (27)) on a set of randomly generated instances.

All instances used in this section were randomly generated in the same manner as described in d'Aspremont et al. [10]. First, we generate a sparse invertible matrix $A \in \mathcal{S}^n$ with positive diagonal entries and a density prescribed by ϱ. We then generate the matrix $B \in \mathcal{S}^n$ by

$$B = A^{-1} + \tau V,$$

where $V \in \mathcal{S}^n$ is an independent and identically distributed uniform random matrix, and τ is a small positive number. Finally, we obtain the following randomly generated sample covariance matrix:

$$\Sigma = B - \min\{\lambda_{\min}(B) - \vartheta, 0\}I,$$

where ϑ is a small positive number. In particular, we set $\varrho = 0.01$, $\tau = 0.15$, and $\vartheta = 1.0e-4$ for generating all instances.

As discussed in section 3.3, our smooth minimization approach has much better worst-case iteration complexity than Nesterov's smooth approximation scheme studied in d'Aspremont et al. [10] for problem (27). However, it is unknown how their practical performance differs from each other. In the first experiment, we compare the practical performance of our smooth minimization approach and its variant with Nesterov's smooth approximation scheme studied in d'Aspremont et al. [10] for problem (27) with $\alpha = 0.1$, $\beta = 10$, and $\rho = 0.5$. For convenience of presentation, we label these three first-order methods as SM, VSM, and NSA, respectively. The codes for them are written in MATLAB. More specifically, the code for NSA follows the algorithm presented in d'Aspremont et al. [10], and the codes for SM and VSM are written in accordance with the algorithms SMACS and VSMACS, respectively. Moreover, we set $\varsigma_1 = \varsigma_2 = 1.05$ and $\varsigma_3 = 0.95$ for the algorithm VSMACS. These three methods terminate once the duality gap is less than $\epsilon = 0.1$. All computations are performed on an Intel Xeon 2.66 GHz machine with Red Hat Linux version 8.

The performance of the methods NSA, SM, and VSM for the randomly generated instances are presented in Table 1. The row size n of each sample covariance matrix Σ is given in column one. The numbers of iterations of NSA, SM, and VSM are given in columns two to four, the objective function values are given in columns five to seven, and the CPU times (in seconds) are given in the last three columns, respectively. From Table 1, we conclude that (i) the method SM, namely, the smooth minimization approach, outperforms substantially the method NSA, that is, Nesterov's smooth approximation scheme; and (ii) the method VSM, namely, the variant of the smooth minimization approach, substantially outperforms the other two methods. In addition, we see from this experiment that Nesterov's smooth minimization approach [19] is generally more appealing than his smooth approximation scheme [21] whenever the problem can be solved as an equivalent smooth problem. Nevertheless, we shall mention that the latter approach has a much wider field of application (e.g., see [21]), where the former approach cannot be applied.

TABLE 1
Comparison of NSA, SM and VSM.

Problem	Iter			Obj			Time		
n	NSA	SM	VSM	NSA	SM	VSM	NSA	SM	VSM
50	3657	457	20	−76.399	−76.399	−76.393	49.0	2.7	0.1
100	7629	920	27	−186.717	−186.720	−186.714	900.4	38.4	0.4
150	20358	1455	49	−318.195	−318.194	−318.184	8165.7	188.8	2.0
200	27499	2294	102	−511.246	−511.245	−511.242	26172.5	698.8	9.2
250	45122	3060	128	−3793.255	−3793.256	−3793.257	87298.9	1767.9	19.8
300	54734	3881	161	−3187.163	−3187.171	−3187.172	184798.1	3994.0	45.5
350	64641	4634	182	−2756.717	−2756.734	−2756.734	351460.7	7613.9	83.6
400	74839	5308	176	−3490.640	−3490.667	−3490.667	614237.1	13536.7	116.9

From the above experiment, we have already seen that the method VSM outperforms substantially two other first-order methods, namely, SM and NSA for solving problem (27). In the second experiment, we compare the performance of the method VSM with the block-coordinate descent methods studied in d'Aspremont et al. [10] and Friedman, Hastie, and Tibshirani [16] on relatively large-scale instances. For convenience of presentation, we label these two methods BCD1 and BCD2, respectively. The method BCD2 was developed very recently and is a slight variant of the method BCD1. In particular, each iterate of BCD1 solves a box constrained quadratic programming by means of interior point methods, but each iterate of BCD2 applies a coordinate descent approach to solving a lasso (l_1-regularized) least-squares problem, which is the dual of the box constrained quadratic programming appearing in BCD1. It is worth mentioning that the methods BCD1 and BCD2 are only applicable for solving problem (16) with $\tilde{\alpha} = 0$ and $\tilde{\beta} = \infty$. Thus, we only compare their performance with our method VSM for problem (16) with such $\tilde{\alpha}$ and $\tilde{\beta}$. As shown in subsection 3.2, problem (16) with $\tilde{\alpha} = 0$ and $\tilde{\beta} = \infty$ is equivalent to problem (27) with α and β given in (18), and hence it can be solved by applying the method VSM to the latter problem instead.

The code for the method BCD1 was written in MATLAB by d'Aspremont and El Ghaoui [9], while the code for BCD2 was written in Fortran 90 by Friedman, Hastie, and Tibshirani [15]. The methods BCD1 and VSM terminate once the duality gap is less than $\epsilon = 0.1$. The original code [15] for BCD2 uses the average absolute change in the approximate solution as the termination criterion. In particular, the average absolute change in the approximate solution is evaluated at the end of each cycle consisting of n block-coordinate descent iterations, and their code terminates once it is below a given accuracy (see [16, p. 6] for details). According to our computational experience, we found that with such a criterion BCD2 is extremely hard to terminate for relatively large-scale instances (say $n = 300$) unless a maximum number of iterations is set. Obviously, it is not easy to choose a suitable maximum number of iterations for BCD2. Thus, to be as fair as possible to BCD1 and VSM, we simply replace their termination criterion detailed in [15] for BCD2 by the one with the duality gap less than $\epsilon = 0.1$. In other words, the duality gap is computed at the end of each cycle consisting of n block-coordinate descent iterations, and BCD2 terminates once it is below $\epsilon = 0.1$. It is worth remarking that the cost for computing a duality gap is $\mathcal{O}(n^3)$ since the inverse of an $n \times n$ symmetric matrix is needed. Thus, it is reasonable to compute the duality gap once every n iterations rather than each iteration.

TABLE 2
Comparison of BCD1, BCD2, and VSM.

Problem	Iter			Obj			Time		
n	BCD1	BCD2	VSM	BCD1	BCD2	VSM	BCD1	BCD2	VSM
100	124	200	33	−186.522	−186.433	−186.522	22.3	0.1	0.5
200	531	600	109	−449.210	−449.179	−449.209	300.0	1.3	9.5
300	1530	1500	146	−767.615	−767.608	−767.614	2428.2	80.9	48.5
400	2259	2400	154	−1082.679	−1082.651	−1082.677	8402.4	298.7	112.3
500	3050	3500	154	−1402.503	−1402.457	−1402.502	22537.1	640.2	211.5
600	3705	4200	165	−1728.628	−1728.587	−1728.627	48950.4	1215.0	397.6
700	4492	4900	163	−2057.894	−2057.862	−2057.892	92052.7	1972.5	611.1
800	4958	5600	169	−2392.713	−2392.671	−2392.712	147778.9	2872.3	943.2
900	5697	6300	161	−2711.874	−2711.827	−2711.874	219644.3	3593.7	1268.5
1000	6536	7000	161	−3045.808	−3045.768	−3045.808	344687.8	6098.7	1710.0

TABLE 3
Comparison of BCD2 and VSM.

Problem	Iter		Obj		Time	
n	BCD2	VSM	BCD2	VSM	BCD2	VSM
100	200	54	−186.433	−186.435	0.1	0.77
200	1200	239	−449.119	−449.122	2.1	21.6
300	3000	310	−767.525	−767.525	32.1	104.2
400	11778400	321	−1082.592	−1082.589	72000.0	223.3
500	6997000	309	−1402.420	−1402.413	72001.0	395.5
600	4637400	318	−1728.553	−1728.538	72004.0	765.2
700	3215100	310	−2057.823	−2057.804	72005.0	1330.0
800	2307200	309	−2392.644	−2392.623	72003.0	1789.2
900	1846800	289	−2711.806	−2711.784	72024.0	2394.0
1000	1257000	283	−3045.749	−3045.718	72051.0	3115.8

All computations are performed on an Intel Xeon 2.66 GHz machine with Red Hat Linux version 8. The performance of the methods BCD1, BCD2, and VSM for the randomly generated instances are presented in Table 2. The row size n of each sample covariance matrix Σ is given in column one. The numbers of iterations of BCD1, BCD2, and VSM are given in columns two to four, the objective function values are given in columns five to eight, and the CPU times (in seconds) are given in the last three columns, respectively. From Table 2, we conclude that both BCD2 and VSM substantially outperform BCD1. We also observe that our method VSM outperforms BCD2 for almost all instances except two relatively small-scale instances.

In the above experimentation, we compared the performance of BCD2 and VSM for $\epsilon = 0.1$. We next compare their performance on the same instances as above and apply the same termination criterion as above except that we set $\epsilon = 0.01$ and an upper bound of 20 hours computation time (or 72,000 seconds) per instance for both codes. The performance of the methods BCD2 and VSM is presented in Table 3. The row size n of each sample covariance matrix Σ is given in column one. The numbers of iterations of BCD2 and VSM are given in columns two to three, the objective function

values are given in columns four to five, and the CPU times (in seconds) are given in the last two columns, respectively. It shall be mentioned that BCD2 and VSM are both feasible methods, and, moreover, (16) and (27) are maximization problems. Therefore for these two methods, the larger the objective function value, the better. From Table 3, we observe that up to accuracy $\epsilon = 0.01$, the method BCD2 cannot solve almost all instances within 20 hours except the first three relatively small-scale ones, but our method VSM does solve each of these instances in less than one hour and produces better objective function values for almost all instances except the first three relatively small-scale ones. Also, it is interesting to observe that the number of iterations for VSM nearly doubles as the accuracy parameter ϵ increases by one digit, which is even better than the theoretical estimate, which is $\sqrt{10}$ according to Theorem 3.2.

5. Concluding remarks. In this paper, we proposed a smooth optimization approach for solving a class of nonsmooth strictly concave maximization problems. We also discussed the application of this approach to sparse covariance selection and proposed a variant of this approach. The computational results showed that the variant of the smooth optimization approach substantially outperforms the latter one, as well as two other first-order methods studied in d'Aspremont et al. [10] and Friedman et al. [16].

As discussed in subsection 3.3, problem (27) has the same form as (2) and satisfies all assumptions imposed on problem (2). Moreover, its associated objective function $\phi(X, U) = \log\det X - \langle \Sigma + \rho U, X\rangle$ is affine with respect to U for every fixed $X \in \mathcal{S}^n_{++}$. In view of these facts along with the remarks made in section 2, one can observe that problem (27) can be suitably solved by Nesterov's excessive gap technique [20]. Since the iterate complexity and the computational cost per iterate of this technique is the same as those of the algorithm SMACS, we expect that the computational performance of these two methods for solving (27) is similar. It would be interesting to implement Nesterov's excessive gap technique [20] and its variant (that is, the one in a similar fashion to the algorithm VSMACS) and compare their computational performance with SMACS and VSMACS, respectively.

Though the variant of the smooth optimization approach outperforms substantially the smooth optimization approach, we are currently only able to establish some preliminary convergence properties for it. A possible direction leading to a thorough proof of its convergence would be to show that the updates on $\hat{\beta}$ in the algorithm VSMACS can occur only for a finite number of times. Given that VSMACS is a nonmonotone algorithm, it is, however, highly challenging to analyze the behavior of the sequences $\{U_k\}$ and $\{X_{\hat{\beta}}(U_k)\}$ and hence the total number of updates on $\hat{\beta}$. Interestingly, we observed in our implementation that when $\hat{\beta} > \lambda_{\max}(X^*)$, the sequence $\{X_{\hat{\beta}}(U_k)\}$ generated by the algorithm VSMACS satisfies $\lambda_{\max}(X_{\hat{\beta}}(U_k)) \in [\lambda_{\max}(X^*), \hat{\beta})$, where X^* is the optimal solution of problem (27). Nevertheless, it remains completely open whether or not this holds in general. In addition, the ideas used in the variant of the smooth optimization approach are interesting in their own right even when viewed as some heuristics. They could also be used to enhance the practical performance of Nesterov's first-order methods [19, 21] for solving some general min-max problems.

The codes for the variant of the smooth minimization approach are written in MATLAB and C, which are available online at www.math.sfu.ca/∼zhaosong. The C code for this method can solve large-scale problems more efficiently, provided the LAPACK package is suitably installed. We will plan to extend these codes for solving

more general problems of the form

$$\begin{aligned} \max_X \quad & \log\det X - \langle \Sigma, X\rangle - \sum_{ij} \omega_{ij}|X_{ij}| \\ \text{s.t.} \quad & \tilde{\alpha} I \preceq X \preceq \tilde{\beta} I, \\ & X_{ij} = 0 \quad \forall (i,j) \in \Omega \end{aligned}$$

for some set Ω, where $\omega_{ij} = \omega_{ji} \geq 0$ for all $i, j = 1, \ldots, n$, and $0 \leq \tilde{\alpha} < \tilde{\beta} \leq \infty$ are some fixed bounds on the eigenvalues of the solution.

Acknowledgments. The author would like to thank Prof. Alexandre d'Aspremont for a careful discussion on the iteration complexity of Nesterov's smooth approximation scheme for sparse covariance selection given in [10]. Also, the author is in debt to two anonymous referees for numerous insightful comments and suggestions, which have greatly improved the paper.

REFERENCES

[1] J. Akaike, *Information theory and an extension of the maximum likelihood principle*, in Proceedings of the Second International Symposium on Information Theory, B. N. Petrov and F. Csaki, eds., Akedemiai Kiado, Budapest, 1973, pp. 267–281.

[2] A. Auslender and M. Teboulle, *Interior gradient and proximal methods for convex and conic optimization*, SIAM J. Optim., 16 (2006), pp. 697–725.

[3] O. Banerjee, L. El Ghaoui, A. d'Aspremont, and G. Natsoulis, *Convex optimization techniques for fitting sparse Gaussian graphical models*, in ICML '06: Proceedings of the 23rd International Conference on Machine Learning, ACM Press, New York, 2006, pp. 89–96.

[4] J. A. Bilmes, *Factored sparse inverse covariance matrices*, in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2 (2000), pp. 1009–1012.

[5] K. P. Burnham and R. D. Anderson, *Multimodel inference. Understanding AIC or BIC in model selection*, Sociol. Methods Res., 33 (2004), pp. 261–304.

[6] S. S. Chen, D. L. Donoho, and M. A. Saunders, *Atomic decomposition by basis pursuit*, SIAM J. Sci. Comput., 20 (1998), pp. 33–61.

[7] J. Dahl, V. Roychowdhury, and L. Vandenberghe, *Maximum Likelihood Estimation of Gaussian Graphical Models: Numerical Implementation and Topology Selection*, manuscript, University of California, Los Angeles, 2004.

[8] J. Dahl, L. Vandenberghe, and V. Roychowdhury, *Covariance selection for nonchordal graphs via chordal embedding*, Optim. Methods Softw., 23 (2008), pp. 501–520.

[9] A. d'Aspremont and L. El Ghaoui, *Covsel: First order methods for sparse covariance selection*, ORFE Department, Princeton University, Princeton, NJ, 2006.

[10] A. d'Aspremont, O. Banerjee, and L. El Ghaoui, *First-order methods for sparse covariance selection*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 56–66.

[11] A. Dempster, *Covariance selection*, Biometrics, 28 (1972), pp. 157–175.

[12] A. Dobra and M. West, *Bayesian Covariance Selection*, ISDS working paper, Duke University, Durham, NC, 2004.

[13] A. Dobra, C. Hans, B. Jones, J. R. Nevins, G. Yao, and M. West, *Sparse graphical models for exploring gene expression data*, J. Multivariate Anal., 90 (2004), pp. 196–212.

[14] D. L. Donoho and J. Tanner, *Sparse nonnegative solutions of underdetermined linear equations by linear programming*, Proc. Natl. Acad. Sci., 102 (2005), pp. 9446–9451.

[15] J. Friedman, T. Hastie, and R. Tibshirani, *Glasso: Graphical lasso for R*, Department of Statistics, Stanford University, Stanford, CA, 2007.

[16] J. Friedman, T. Hastie, and R. Tibshirani, *Sparse inverse covariance estimation with the graphical lasso*, Biostatistics, 9 (2008), pp. 432–441.

[17] J. Z. Huang, N. Liu, and M. Pourahmadi, *Covariance matrix selection and estimation via penalised normal likelihood*, Biometrika, 93 (2006), pp. 85–98.

[18] B. Jones, C. Carvalho, C. Dobra, A. Hans, C. Carter, and M. West, *Experiments in stochastic computation for high-dimensional graphical models*, Statist. Sci., 20 (2005), pp. 388–400.

[19] Y. E. Nesterov, *A method for unconstrained convex minimization problem with the rate of convergence* $O(1/k^2)$, Dokl. Akad. Nauk SSSR, 269 (1983), pp. 543–547 (in Russian).
[20] Y. Nesterov, *Excessive gap technique in nonsmooth convex minimization*, SIAM J. Optim., 16 (2005), pp. 235–249.
[21] Y. E. Nesterov, *Smooth minimization of nonsmooth functions*, Math. Programming, 103 (2005), pp. 127–152.
[22] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia, 1994.
[23] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Roy. Statist. Soc. Ser. B, 58 (1996), pp. 267–288.
[24] L. Vandenberghe, S. Boyd, and S.-P. Wu, *Determinant maximization with linear matrix inequality constraints*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 499–533.
[25] M. Yuan and Y. Lin, *Model selection and estimation in the Gaussian graphical model*, Biometrika, 94 (2007), pp. 19–35.

SIAM J. OPTIM.
Vol. 19, No. 4, pp. 1828–1845

CONVERGENCE ANALYSIS OF GENERALIZED ITERATIVELY REWEIGHTED LEAST SQUARES ALGORITHMS ON CONVEX FUNCTION SPACES*

NICOLAI BISSANTZ†, LUTZ DÜMBGEN‡, AXEL MUNK§, AND BERND STRATMANN†

Yehuda Vardi passed away unexpectedly on Jan. 13th, 2005, and we would like to dedicate this work to his memory.

Abstract. The computation of robust regression estimates often relies on minimization of a convex functional on a convex set. In this paper we discuss a general technique for a large class of convex functionals to compute the minimizers iteratively, which is closely related to majorization-minimization algorithms. Our approach is based on a quadratic approximation of the functional to be minimized and includes the iteratively reweighted least squares algorithm as a special case. We prove convergence on convex function spaces for general coercive and convex functionals F and derive geometric convergence in certain unconstrained settings. The algorithm is applied to total variation (TV) penalized quantile regression and is compared with a step size corrected Newton–Raphson algorithm. It is found that typically in the first steps the iteratively reweighted least squares algorithm performs significantly better, whereas the Newton type method outpaces the former only after many iterations. Finally, in the setting of bivariate regression with unimodality constraints we illustrate how this algorithm allows one to utilize highly efficient algorithms for special quadratic programs in more complex settings.

Key words. regression analysis, monotone regression, quantile regression, shape constraints, L^1 regression, nonparametric regression, total variation semi-norm, reweighted least squares, Fermat's problem, convex approximation, quadratic approximation, pool adjacent violators algorithm

AMS subject classifications. 62G07, 62J05, 65K05, 85-08

DOI. 10.1137/050639132

1. Introduction.

The computation of robust parametric and nonparametric regression estimators often requires the minimization of (convex) functionals on a set $\mathcal{C}$ which is determined by a priori information on the model underlying the data. For example, $\mathcal{C}$ can be a linear finite-dimensional space (linear model) or the set of isotonic vectors $m = (m_1, \ldots, m_d) \in \mathbb{R}^d$, $m_1 \le \ldots \le m_d$, with $d \le n$. To this end the functional

$$F^{(\rho)}(m) = \sum_{i=1}^{n} \rho(r_i(m)) \tag{1}$$

has to be minimized over $\mathcal{C} \subset \mathbb{R}^d$. Here $r_1, \ldots, r_n$ denote the (model-dependent) residuals of n data pairs (X_i, Y_i), $1 \le i \le n$, and ρ a given loss function [17]. Taking

*Received by the editors August 29, 2005; accepted for publication (in revised form) November 4, 2008; published electronically February 20, 2009. Financial support of the Swiss National Science Foundation, of the DFG by the grants Graduiertenkolleg 1023, SFB 475 and FOR 916, and by the DAAD is gratefully acknowledged.

http://www.siam.org/journals/siopt/19-4/63913.html

†Fakultät für Mathematik, Universitätsstraße 150, Mathematik III, NA 3/70, D-44780 Bochum, Germany (nicolai.bissantz@rub.de).

‡Institute of Mathematical Statistics and Actuarial Science, University of Bern, 3012 Bern, Switzerland (dumbgen@stat.unibe.ch).

§Institute for Mathematical Stochastics, Georg-August-University Göttingen, 37073 Göttingen, Germany (munk@math.uni.goettingen.de, bstrat@gmx.net).

$\rho(z) = z^2/2$ gives the ordinary least squares problem, while

$$\rho(z) = 2|z| \cdot \begin{cases} p & z \geq 0 \\ 1-p & z < 0 \end{cases} \tag{2}$$

with $0 < p < 1$ yields quantile regression [21, 30]. Other functions are Huber's [16] loss function

$$\rho(z) = \begin{cases} z^2/2 & |z| \leq \gamma \\ \gamma|z| - \gamma^2/2 & |z| > \gamma \end{cases}$$

or the logistic loss function $\rho(z) = \gamma^z \log(\cosh(z/\gamma))$ [7] for some $\gamma > 0$. An important extension of (1) are functionals

$$F(m) = F^{(\rho)}(m) + \lambda P(m)\ , \quad \lambda \geq 0, \tag{3}$$

where $P(m)$ denotes a penalizing term such as, for instance, the discrete total variation semi-norm of $m \in \mathbb{R}^d$,

$$P(m) = \sum_{j=1}^{d-1} |m_j - m_{j+1}|\ ; \tag{4}$$

see [23, 22] or [27]. In this paper a generalization of the iteratively reweighted least squares (IRLS) algorithm—therefore named GIRLS—is considered for minimization of a functional F as in (3) over any convex subset $\mathcal{C}$ of $\mathbb{R}^d$. This allows us to extend the IRLS algorithm, for example, to situations where $\mathcal{C}$ is defined as the space of monotone (or k-modal) vectors or to the problem of nonparametric regression estimates with total variation semi-norm penalization of its discrete derivative.

The general idea of the IRLS algorithm (and variants of it) is to approximate the functional F in a first step by smooth functionals F_δ such that $F_\delta \to F$ pointwise as $\delta \searrow 0$. The collection $(F_\delta)_{\delta>0}$ will be called a regularization of F (cf. Def. 1). In a second step, for each given base point $f \in \mathcal{C}$ the functional F_δ will be approximated by $G_\delta(f, .)$ (cf. Def. 2). Here $G_\delta : \mathcal{C} \times \mathbb{R}^d \to \mathbb{R}$ is a functional which is chosen such that a quick and numerically stable minimization can be performed. The resulting minimizer will serve as an approximation for the minimizer m_δ^* of F_δ and hence for a minimizer m^* of F. In particular, if it is possible to choose G_δ as a polynomial of degree two, the well-known IRLS algorithm may result [25, 28, 10].

The **GIRLS algorithm** can be summarized schematically as follows:

Preliminary Step: Determine a regularization $(F_\delta)_{\delta>0}$ of F and corresponding smooth approximations G_δ, $\delta > 0$.

Step 1: Initialize $\delta > 0$ and $m_\delta^{(0)} \in \mathcal{C}$.

Step 2: Repeat the following procedure until δ is sufficiently small: Compute

$$m_\delta^{(k)} := \operatorname*{argmin}_{m \in \mathcal{C}} G_\delta(m_\delta^{(k-1)}, m) \qquad \text{for } k = 1, 2, 3, \ldots \tag{5}$$

and terminate this iteration for a proper $k = k(\delta)$. Then replace $(\delta, m_\delta^{(0)})$ with $(\delta/2, m_\delta^{(k(\delta))})$.

Output: The final $m_\delta^{(k(\delta))}$ is our approximate minimizer of F over $\mathcal{C}$.

A more detailed description of this algorithm, including pseudocode and an explicit rule for $k(\delta)$ is provided in section 3.2.

The IRLS and related algorithms are based on the idea of majorizing functionals by a sequence of quadratic approximations and subsequent minimization. These have been treated extensively in the literature, e.g., [24, 19, 42, 29, 9, 18, 36, 37], and the

references therein. However, in most cases convergence is only shown for $\mathcal{C} = \mathbb{R}^d$. This simplifies proofs notably, since the minimizers can be represented as zeros of the derivatives of the functional. For arbitrary convex $\mathcal{C}$, however, the minimizers are no longer represented solely by such equality constraints, instead inequalities occur. Notable exceptions for general convex $\mathcal{C}$ are in [13], where however, the convergence results are restricted to a special class of functionals, requiring, e.g., $F(m) = O(\|m\|)$, or [39], who show convergence on convex polyhedral sets under the assumption that F is two times differentiable. Our findings generalize these results to the case of $\mathcal{C}$ being an arbitrary convex closed set as well as to more general functionals which are only required to be coercive and convex. This appears to be close to the weakest possible set of assumptions required for a general proof of convergence. Our proof adopts various arguments from convex analysis.

It is interesting to note that in the numerical literature the IRLS algorithm is denoted as the Weiszfeld algorithm [40, 41] who suggested this algorithm to solve the Fermat-Steiner-Weber problem [40, 41, 24, 19], which is known to the statistical community as the computation of the spatial median (as mentioned in [5, 6, 11]).

The remainder of this paper is organized as follows. First, we motivate the GIRLS algorithm for the special case of L^1-regression in section 2. Then we present the GIRLS algorithm in a general framework and prove various results about its convergence in section 3. Its convergence to the minimizer m^*_δ and hence to m^* as $\delta \searrow 0$ will be shown under very general assumptions (Theorem 2). Furthermore, in Theorem 5 we prove geometric, or, more precisely, at least Q-linear convergence of the sequence $(m_k)_k$ to m^*_δ under slightly stronger conditions (cf. [39] and [3]), and guidance is provided on the choice of the number of iterates in (5) and the regularization parameter δ. Finally, we show in Theorem 3 that any convex and coercive functional F can be regularized by a sequence F_δ s.t. each F_δ admits a quadratic approximation G_δ from above.

We stress that an advantage of the GIRLS approach is flexibility in the choice of F_δ and G_δ. This choice can be driven by various aspects, such as computational efficiency or rate of convergence (cf. Theorem 3). In this paper we emphasize the possibility to make use of efficient algorithms already available for the minimization of G_δ, such as the pool adjacent violators algorithm (PAVA) for isotonic weighted least squares approximation (see [33] for a comprehensive treatment). This is illustrated in section 4, where we describe the construction of F_δ and G_δ in some specific cases explicitly. In section 5 we discuss two numerical examples. In the first example we investigate in detail numerical performance of the GIRLS algorithm for the case of total variation (TV) penalized quantile regression. To this end the GIRLS algorithm is compared with a step size corrected Newton–Raphson algorithm. It is found that typically it outperforms the latter one in the first iteration steps significantly, in particular when the initial value is far from the optimum. This finding coincides with other numerical experiments, e.g., when applying the algorithms to L^1-penalized Poisson regression.

In the second example we apply the GIRLS algorithm to a two-dimensional TV minimization where we impose an additional unimodality constraint in one direction. We show that the GIRLS algorithm allows us to include the PAVA for the univariate unimodal subproblem, which is in general not possible for regression with two- or higher dimensional predictor. Note also that PAVA type methods are not available in general if an additional penalization term as in (3) is added. Again, the GIRLS algorithm offers a possibility to include them in each updating step.

In summary, the main advantage of the GIRLS algorithm is twofold. First, it is simple to perform and offers great flexibility for the choice of the approximating functionals G_δ. Second, it allows us to combine various restrictions and minimization

criteria (such as monotonicity constraints and roughness penalties). For such complex minimization problems, simple and quick algorithms such as PAVA or Newton type algorithms are not available in general, and more complicated and time consuming algorithms such as quadratic programming or interior point methods become necessary. Here the GIRLS algorithm represents a feasible alternative because it typically requires in each updating step the computation of minimizers (e.g., a weighted L^2 solution), which can be obtained easily. Further, our numerical experiments have shown that a rather small number of updating steps give already satisfactory results and the GIRLS algorithm outperfoms competitors in the first iterations, which is in accordance with previous numerical findings (see, e.g., [39]). Hence, as a practical rule of thumb, we find that the GIRLS algorithm is very simple to implement and provides a quick improvement of an initial value by a few iterations. It can be improved additionally by performing subsequent iterations by other, more sophisticated, optimization algorithms.

2. L^1-regression with the GIRLS algorithm. As a motivating example consider the L^1 linear regression problem for observations $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ in $\mathbb{R}^d \times \mathbb{R}$. Assuming that Y_i equals $X_i^\top m$ plus a random error, the goal is to compute

$$m := \operatorname*{argmin}_{m \in \mathbb{R}^d} \sum_{i=1}^{n} |Y_i - X_i^\top m| = \operatorname*{argmin}_{m \in \mathbb{R}^d} F(m), \tag{6}$$

an estimator of the unknown parameter vector $m \in \mathbb{R}^d$. Iteratively reweighted least squares is based on the idea that, in a first step, the L^1-norm F, being a convex functional, will be approximated (regularized) by a family of smooth convex functionals F_δ, $\delta > 0$, e.g.,

$$F_\delta(m) = \sum_{i=1}^{n} h_\delta \left(Y_i - X_i^\top m \right),$$

where

$$h_\delta(z) = [z^2 + \delta]^{1/2}. \tag{7}$$

The regularization of a nonsmooth functional as in (6) by (7) is well known, of course (see, e.g., [38]). It is supposed that minimization of F_δ is numerically better tractable than minimization of the original functional F in in (6). Then $m_\delta := \operatorname{argmin}_{m \in \mathbb{R}^d} F_\delta(m)$ will be an approximation of m (cf. Theorem 1). In order to compute m_δ the following recursion formula is iterated:

$$m_\delta^{(k+1)} = \operatorname*{argmin}_{m \in \mathbb{R}^d} \sum_{i=1}^{n} \frac{(Y_i - X_i^\top m)^2}{h_\delta(Y_i - X_i^\top m_\delta^{(k)})}. \tag{8}$$

Note that in each updating step the computation of $m_\delta^{(k+1)}$ means solving a simple diagonally reweighted least squares minimization problem, which can easily be done by using standard methods such as, e.g., Householder QR decomposition. As a starting value $m_\delta^{(0)}$ any (reasonable) choice, e.g., the least squares estimator, may serve.

It is instructive to indicate a proof for this simple case. The basic idea is to approximate $h_\delta(z)$ from above for any given real number r by a quadratic function $g_\delta(r, z) = c(r) + a(r)z^2/2$ of z such that $g_\delta(r, \cdot) \geq h_\delta$ and $g_\delta(r, r) = h_\delta(r)$. This can

be achieved indeed with

$$g_\delta(r,z) \;=\; h_\delta(r) + h_\delta(r)^{-1}(z^2 - r^2)/2; \tag{9}$$

see also Lemma 1 in section 4. The intrinsic reason is that h_δ is an even convex function whose second derivative h_δ'' is nonincreasing on $[0,\infty)$. Thus $m_\delta^{(k+1)}$ in (8) is the minimizer of

$$G_\delta(m_\delta^{(k)}, m) \;:=\; \sum_{i=1}^n g_\delta\left(Y_i - X_i^\top m_\delta^{(k)}, Y_i - X_i^\top m\right)$$

over all $m \in \mathbb{R}^d$. Note that F_δ as well as $G_\delta(m_\delta^{(k)}, \cdot)$ are convex functions such that $F_\delta(m) \le G_\delta(m_\delta^{(k)}, m)$ with equality for $m = m_\delta^{(k)}$, and their gradients satisfy

$$\nabla F_\delta(m_\delta^{(k)}) = \nabla G_\delta(m_\delta^{(k)}, m_\delta^{(k)}).$$

Here and in the following the gradient of G_δ is defined with respect to the second argument. Thus

$$F_\delta(m_\delta^{(k+1)}) \le G_\delta(m_\delta^{(k)}, m_\delta^{(k+1)}) \le G_\delta(m_\delta^{(k)}, m_\delta^{(k)}) = F_\delta(m_\delta^{(k)}),$$

and the second inequality in the latter display is strict if, and only if, $m_\delta^{(k)}$ differs from the solution m_δ. Consequently, $F_\delta(m_\delta^{(k)})$ is either strictly decreasing in k, or $m_\delta^{(k)} = m_\delta$ for sufficiently large k. This fact was established by [25] for the particular problem (6). Convergence of $m_\delta^{(k)}$ to m_δ as $k \to \infty$ follows from our general Theorem 2 below.

3. The GIRLS algorithm.

3.1. Main theorem and convergence analysis. Returning to the general setting, we always assume that our target functional $F : \mathbb{R}^d \to \mathbb{R}$ is convex and coercive, i.e., $F(x) \to \infty$ as $\|x\| \to \infty$. Moreover, let $\mathcal{C} \subset \mathbb{R}^d$ be closed and convex. This entails that the set

$$M^* := \operatorname*{argmin}_{m \in \mathcal{C}} F(m)$$

is a nonvoid, compact, and convex subset of $\mathcal{C}$. Now the first step is to approximate F by a family of strictly convex and smooth functionals F_δ, $\delta > 0$, converging pointwise to F as $\delta \searrow 0$. This is summarized in the following definition.

DEFINITION 1. *A functional $F_\delta : \mathbb{R}^d \to \mathbb{R}$ is called* regular*, if F_δ is strictly convex, continuously differentiable, and coercive. A* regularization of F *consists of regular functionals F_δ, $\delta > 0$, such that F_δ converges pointwise to F as $\delta \searrow 0$.*

Theorem 4 below shows that there exists always a regularization $(F_\delta)_{\delta>0}$ for F. It follows from strict convexity and coercivity of F_δ that it has a unique minimizer

$$m_\delta^* := \operatorname*{argmin}_{m \in \mathcal{C}} F_\delta(m),$$

which serves as an approximation to M^*. The next theorem provides an exact formulation of this fact.

THEOREM 1 (approximation of M^*). *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a convex and coercive functional, and let $(F_\delta)_{\delta>0}$ be a regularization of F. Then, as $\delta \searrow 0$,*

$$\left.\begin{matrix} F_\delta(m_\delta^*) \\ F(m_\delta^*) \end{matrix}\right\} \to \min_{x\in\mathcal{C}} F(x) \quad \text{and} \quad d(m_\delta^*, M^*) := \inf_{y\in M^*} \|m_\delta^* - y\| \to 0.$$

Before proving Theorem 1 we summarize some well-known facts about convex functionals (see [34]), which we utilize in the subsequent proofs. A convex functional on $\mathbb{R}^d$ is automatically continuous. If a sequence of convex functionals on $\mathbb{R}^d$ converges pointwise, then the convergence is uniform on arbitrary bounded sets. Finally, if $H : \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable, and if $\mathcal{C} \subset \mathbb{R}^d$ is closed and convex, then $f \in \mathcal{C}$ minimizes H over $\mathcal{C}$ if, and only if,

$$\nabla H(f)^\top (m - f) \geq 0 \quad \text{for all } m \in \mathcal{C}. \tag{10}$$

Proof of Theorem 1. For any set $S \subset \mathbb{R}^d$ let $\|F - F_\delta\|_S$ be the supremum norm of $F - F_\delta$ over S. Since M^* is compact, for any fixed $\epsilon > 0$, the set $B_\epsilon := \{m \in \mathcal{C} : d(m, M^*) \leq \epsilon\}$ is compact, too. Thus $\|F - F_\delta\|_{B_\epsilon}$ tends to zero as $\delta \searrow 0$. In particular, for sufficiently small $\delta > 0$,

$$\min_{m\in\mathcal{C}\,:\, d(m,M^*)=\epsilon} F_\delta(m) > \max_{m\in M^*} F_\delta(m). \tag{11}$$

To verify (11), first note that it holds with F in place of F_δ, by definition of M^*. Since $F_\delta \to F$ uniformly on B_ϵ, (11) holds for sufficiently small $\delta > 0$. But (11) implies that $F_\delta(m_o) > \min_{m\in M^*} F_\delta(m)$ for any $m_o \in \mathcal{C} \setminus B_\epsilon$, and hence $m_\delta^* \in B_\epsilon$. Let m_* be the metric projection of m_o onto M^* and write $m_o = m_* + tv$ for some unit vector $v \in \mathbb{R}^d$ and a scalar $t > \epsilon$. Then it follows from convexity of the function $t \mapsto F_\delta(m_* + tv)$ that

$$\begin{aligned} F_\delta(m_o) - \min_{m\in M^*} F_\delta(m) &\geq F_\delta(m_* + tv) - F_\delta(m_*) \\ &\geq (t/\epsilon)(F_\delta(m_* + \epsilon v) - F_\delta(m_*)) \\ &> 0 \end{aligned}$$

in case of (11). These considerations show already that $d(m_\delta^*, M^*) \to 0$ as $\delta \searrow 0$. Note also that in case of $m_\delta^* \in B_\epsilon$,

$$|F_\delta(m_\delta^*) - F(m_\delta^*)| \leq \|F - F_\delta\|_{B_\epsilon}, \tag{12}$$

and

$$F(m_\delta^*) - \min_{x\in\mathcal{C}} F(x) \leq \max_{x,y\in B_\epsilon\,:\,\|x-y\|\leq\epsilon} \big|F(y) - F(x)\big|.$$

Finally, the r.h.s. of the latter inequality tends to 0 as $\epsilon \searrow 0$, by compactness of M^* and continuity of F. These findings show that both $F(m_\delta^*)$ and $F_\delta(m_\delta^*)$ tend to $\min_{x\in\mathcal{C}} F(x)$ as $\delta \searrow 0$. □

The second step is to determine m_δ^* via approximations $G_\delta(f, \cdot)$ of F_δ for various $f \in \mathcal{C}$ as in (5). The following definition summarizes our assumptions on G_δ.

DEFINITION 2. *Let $F_\delta : \mathbb{R}^d \to \mathbb{R}$ be a regular functional. Another functional $G_\delta : \mathcal{C} \times \mathbb{R}^d \to \mathbb{R}$ is called a* smooth approximation of F_δ from above, *if it is continuous in both arguments and satisfies the following additional properties for arbitrary $f \in \mathcal{C}$:*

(i) $G_\delta(f,\cdot)$ *is strictly convex and continuously differentiable,*
(ii) $G_\delta(f,m) \geq F_\delta(m)$ *for all* $m \in \mathbb{R}^d$ *with equality for* $m = f$.

The functional G_δ *is called a* quadratic approximation of F_δ from above *if, in addition,* $G_\delta(f,\cdot)$ *is always a polynomial of order two, i.e.,*

$$G_\delta(f,m) = F_\delta(f) + \nabla F_\delta(f)^\top (m-f) + 2^{-1}(m-f)^\top B(f)(m-f) \tag{13}$$

for some symmetric, positive definite matrix $B(f) \in \mathbb{R}^{d\times d}$.

The next theorem is the main result of this paper.

THEOREM 2 (convergence of the GIRLS algorithm). *Let* $\mathcal{C} \subset \mathbb{R}^d$ *be a closed convex set and* $F_\delta : \mathbb{R}^d \to \mathbb{R}$ *be a regular functional which can be smoothly approximated from above by* G_δ. *Then the GIRLS algorithm, defined by* (5) *with an arbitrary starting point* $m_\delta^{(0)} \in \mathcal{C}$, *yields a sequence* $(m_\delta^{(k)})_{k=0}^\infty$ *converging to* m_δ^*.

Proof. At first we prove that $F_\delta(m_\delta^{(k)})$ is decreasing in k. It follows from Property (ii) in Definition 2 that the gradients $\nabla F_\delta(m)$ and $\nabla G_\delta(m_\delta^{(k)}, m)$ (defined with respect to the second argument) coincide for $m = m_\delta^{(k)}$. Thus it follows from the characterization (10), applied to $H = G_\delta(m_\delta^{(k)},\cdot)$ and $H = F_\delta$, respectively, that $m_\delta^{(k+1)} = m_\delta^{(k)}$ if, and only if, $m_\delta^{(k)} = m_\delta^*$. Otherwise the minimizer $m_\delta^{(k+1)}$ of $G_\delta(m_\delta^{(k)},\cdot)$ over $\mathcal{C}$ differs from $m_\delta^{(k)}$, whence Property (ii) in Definition 2 entails

$$F_\delta(m_\delta^{(k+1)}) \leq G_\delta(m_\delta^{(k)}, m_\delta^{(k+1)}) < G_\delta(m_\delta^{(k)}, m_\delta^{(k)}) = F_\delta(m_\delta^{(k)}).$$

By monotonicity of $(F_\delta(m_\delta^{(k)}))_k$, all points $m_\delta^{(k)}$ lie in the set $\{m \in \mathcal{C} : F_\delta(m) \leq F_\delta(m_\delta^{(0)})\}$, which is compact by continuity and coercivity of F_δ. Hence it is sufficient to show that any limit point m_o equals m_δ^*. Now, take an arbitrary convergent subsequence $(m_\delta^{(k_\ell)})_\ell$ with limit m_o. For any $v \in \mathcal{C}$,

$$\begin{aligned} F_\delta\left(m_\delta^{(k_\ell+1)}\right) &\leq G_\delta\left(m_\delta^{(k_\ell)}, m_\delta^{(k_\ell+1)}\right) \\ &\leq G_\delta\left(m_\delta^{(k_\ell)}, v\right) \\ &\to G_\delta(m_o, v) \quad \text{as } \ell \to \infty, \end{aligned}$$

by continuity of G_δ. But

$$\lim_{\ell\to\infty} F_\delta\left(m_\delta^{(k_\ell+1)}\right) \geq \lim_{\ell\to\infty} F_\delta\left(m_\delta^{(k_{\ell+1})}\right) = F_\delta(m_o) = G_\delta(m_o, m_o).$$

Thus $G_\delta(m_o, m_o) \leq G_\delta(m_o, v)$ for all $v \in \mathcal{C}$; i.e., m_o is the unique minimizer of $G_\delta(m_o,\cdot)$. As argued above, this entails that $m_o = m_\delta^*$. □

The next theorem states that convex and coercive functionals F can always be regularized and approximated quadratically from above. Hence GIRLS is, in principle, always applicable.

THEOREM 3 (regularization and approximation of F). *Let* $F : \mathbb{R}^d \to \mathbb{R}$ *be a convex and coercive functional. Then there exists a regularization* $(F_\delta)_{\delta>0}$ *of* F *such that each* F_δ *admits a quadratic approximation* G_δ *from above.*

In order to prove Theorem 3, we require the following result.

THEOREM 4. *Let* F *be a nonnegative, coercive, convex functional on* $\mathbb{R}^d$. *Then there are strictly convex and infinitely often differentiable functionals* $F_\delta \geq F$, $\delta > 0$, *such that* $F_\delta \to F$ *pointwise as* $\delta \searrow 0$.

Proof. Let $K(x) := 1\{\|x\| < 1\}C \exp\bigl(-(1-\|x\|^2)^{-1}\bigr)$, where C is chosen such that K integrates to one. This is a well-known example of an infinitely differentiable, nonnegative, even kernel function with compact support $\{x : \|x\| \le 1\}$. For $\delta > 0$ we define $K_\delta(x) := \delta^{-1}K(\delta^{-1}x)$ and

$$F_\delta(x) := \int F(y)K_\delta(x-y)\,\mathrm{d}y = \int F(x+\delta z)K(z)\,\mathrm{d}z.$$

It is well known that F_δ is infinitely often differentiable with limit F pointwise (cf. [35]. It is also inherits convexity from F, because for $x, y \in \mathbb{R}^d$ and $\lambda \in (0,1)$,

$$\begin{aligned} F_\delta((1-\lambda)x+\lambda y) &= \int F_\delta\bigl((1-\lambda)(x+\delta z)+\lambda(y+\delta z)\bigr)K(z)\,\mathrm{d}z \\ &\le \int \bigl((1-\lambda)F(x+\delta z)+\lambda F(y+\delta z)\bigr)K(z)\,\mathrm{d}z \\ &= (1-\lambda)F_\delta(x)+\lambda F_\delta(y). \end{aligned}$$

Moreover, since K is even,

$$F_\delta(x) = \int \frac{F(x+\delta z)+F(x-\delta z)}{2}K(z)\,\mathrm{d}z \ge \int F(x)K(z)\,\mathrm{d}z = F(x),$$

again by convexity of F. Finally, if F_δ fails to be strictly convex, we may add to F_δ the strictly convex function $x \mapsto \delta\|x\|^2$. □

We mention that the construction of F_δ given here is mainly for theoretical purposes, and may in practice be difficult to evaluate numerically due to the high dimensionality of the integral.

Proof of Theorem 3. Let $(F_\delta)_{\delta>0}$ be a regularization of F such that D^2F_δ is positive definite everywhere; cf. Theorem 4 and its proof. It may happen that $\limsup_{\|m\|\to\infty} F_\delta(m)/\|m\|^2 = \infty$, rendering quadratic approximation of F_δ from above impossible. Thus we modify the functions F_δ as follows: Let

$$c_\delta := \max_{\|m\|\le\delta^{-1}} \lambda_{\max}(D^2F_\delta(m))$$

with $\lambda_{\max}(A)$ denoting the largest eigenvalue of a symmetric matrix $A \in \mathbb{R}^{d\times d}$. Starting from the representation

$$F_\delta(m) = F_\delta(0) + \nabla F_\delta(0)^\top m + \int_0^1 m^\top D^2F_\delta(tm)m\,(1-t)\,dt,$$

we define

$$\tilde{F}_\delta(m) := F_\delta(0) + \nabla F_\delta(0)^\top m + \int_0^1 m^\top \min(D^2F_\delta(tm), c_\delta I)m\,(1-t)\,dt.$$

Here $\min(A, c_\delta I) \in \mathbb{R}^{d\times d}$ is obtained from the spectral representation of A by replacing each eigenvalue $\lambda_i(A)$ with $\min(\lambda_i(A), c_\delta)$. Note that $\tilde{F}_\delta$ is twice continuously differentiable with $\tilde{F}_\delta(0) = F_\delta(0)$, $\nabla\tilde{F}_\delta(0) = \nabla F_\delta(0)$, and $D^2\tilde{F}_\delta = \min(D^2F_\delta, c_\delta I)$. The Hessian matrix is positive definite with largest eigenvalue never exceeding c_δ. In addition, $\tilde{F}_\delta = F_\delta$ on $\{m : \|m\| \le \delta^{-1}\}$. Thus for sufficiently small $\delta > 0$, $\tilde{F}_\delta$ is regular, and a quadratic approximation of $\tilde{F}_\delta$ from above is given by

$$G_\delta(f,m) := \tilde{F}_\delta(f) + \nabla\tilde{F}_\delta(f)^\top(m-f) + c_\delta\|m-f\|^2/2. \qquad \square$$

Remark 1. In Definition 1 we assume that F_δ is strictly convex. This property is only required for notational convenience, because it guarantees uniqueness of the minimizer m_δ^*. A careful inspection of the proof of Theorem 1 shows, however, that convergence continues to hold if strict convexity is replaced with convexity. Only the assertion $d(m_\delta^*, M^*) \to 0$ has to be replaced by

$$\sup_{x \in M_\delta^*} \inf_{y \in M^*} \|x - y\| \to 0,$$

where $M_\delta^* := \operatorname{argmin}_{m \in \mathcal{C}} F_\delta(m)$. An analogous modification holds for Theorem 2.

We close the section with the following result, which shows under additional regularity conditions on F_δ and $\mathcal{C}$ geometric, or, more precisely, at least Q-linear convergence of the GIRLS algorithm (cf. [3, Theorem 4.1], for a related result).

THEOREM 5 (geometric convergence of the GIRLS algorithm). *Let $F_\delta : \mathbb{R}^d \to \mathbb{R}$ be coercive and twice continuously differentiable with positive definite Hessian matrix $D^2F(m_\delta^*) =: A$. Further let $G_\delta : \mathbb{R}^d \times \mathbb{R}^d$ be a quadratic approximation of F_δ from above with Hessian matrix $B(m_\delta^*) =: B$ as in* (13). *Then the GIRLS algorithm yields a sequence $(m_\delta^{(k)})_{k=0}^\infty$ converging to $m_\delta^* = \operatorname{argmin}_{\mathcal{C}} F_\delta$ such that*

$$\limsup_{k \to \infty} \frac{\|m_\delta^{(k+1)} - m_\delta^*\|_A}{\|m_\delta^{(k)} - m_\delta^*\|_A} \;\le\; 1 - \lambda_{\min}\left(B^{-1}A\right) \in [0, 1).$$

Here $\|v\|_A := (v^\top A v)^{1/2}$, and $\lambda_{\min}(B^{-1}A) \in (0, 1]$ denotes the smallest eigenvalue of $B^{-1}A$.

Proof. According to Theorem 2, $\lim_{k\to\infty} m_\delta^{(k)} = m_\delta^*$. Since $\mathcal{C} = \mathbb{R}^d$, $\nabla F_\delta(m_\delta^*) = 0$ and

$$\begin{aligned} m_\delta^{(k+1)} &= m_\delta^{(k)} - B(m_\delta^{(k)})^{-1} \nabla F_\delta(m_\delta^{(k)}) \\ &= m_\delta^{(k)} - B(m_\delta^{(k)})^{-1} \int_0^1 D^2 F_\delta\left((1-t)m_\delta^* + t m_\delta^{(k)}\right)(m_\delta^{(k)} - m_\delta^*)\,\mathrm{d}t \\ &= m_\delta^{(k)} - B^{-1}A(m_\delta^{(k)} - m_\delta^*) + o\Big(\|m_\delta^{(k)} - m_\delta^*\|\Big). \end{aligned}$$

Thus

$$\frac{\|m_\delta^{(k+1)} - m_\delta^*\|_A}{\|m_\delta^{(k)} - m_\delta^*\|_A} = \frac{\|(I - B^{-1}A)(m_\delta^{(k)} - m_\delta^*)\|_A}{\|m_\delta^{(k)} - m_\delta^*\|_A} + o(1),$$

and for any vector $v \in \mathbb{R}^d$,

$$\begin{aligned} \frac{\|(I - B^{-1}A)v\|_A^2}{\|v\|_A^2} &= \frac{v^\top (I - AB^{-1})A(I - B^{-1}A)v}{v^\top A v} \\ &= \frac{w^\top A^{-1/2}(I - AB^{-1})A(I - B^{-1}A)A^{-1/2}w}{\|w\|^2} \quad (\text{with } w := A^{1/2}v) \\ &= \frac{w^\top C^2 w}{\|w\|^2} \quad (\text{with } C := I - A^{1/2}B^{-1}A^{1/2}) \\ &\le \lambda_{\max}(C^2). \end{aligned}$$

It follows from property (ii) of G_δ in Definition 2 that $B - A$ is nonnegative definite, which implies that $\lambda_i(B^{-1}A) = \lambda_i(A^{1/2}B^{-1}A^{1/2}) \in (0, 1]$. This entails that C is nonnegative definite with $\lambda_{\max}(C^2) = \lambda_{\max}(C)^2 = (1 - \lambda_{\min}(B^{-1}A))^2$. ∎

TABLE 1
Generalized iteratively reweighted least squares algorithm (GIRLS).

Algorithm $m_\star \leftarrow$ **GIRLS**$(F, (F_\delta, G_\delta)_{\delta>0}, \delta_o, \delta_{\min}, m_o, \epsilon, k_{\max})$
$\delta \leftarrow \delta_o$
$m_\star \leftarrow m_o$
while $\delta \geq \delta_{\min}$ **do**
  $m_{\text{new}} \leftarrow \operatorname{argmin}_{m\in\mathcal{C}} G_\delta(m_\star, m)$
  $k \leftarrow 0$
  while $F(m_{\text{new}})/F(m_\star) < 1-\epsilon$ **and** $k < k_{\max}$ **do**
    $m_\star \leftarrow m_{\text{new}}$
    $m_{\text{new}} \leftarrow \operatorname{argmin}_{m\in\mathcal{C}} G_\delta(m_\star, m)$
    $k \leftarrow k+1$
  end while
  $\delta \leftarrow \delta/2$
end while.

3.2. Pseudocode, proper choice of δ, and the number of iterations. In practical applications the points m_δ^* are never calculated exactly. Instead after finitely many, say $k(\delta)$, iterations of (5) the iteration is terminated and the regularization parameter δ is decreased, e.g., replaced with $\delta/2$. An obvious question is how to choose these iteration numbers $k(\delta)$. We found empirically in most cases that for a fixed parameter $\delta > 0$, the values $F(m_\delta^{(k)})$ are decreasing for $k \leq k_o(\delta)$ and increasing in $k \geq k_o(\delta)$ for some fixed $k_o(\delta) \in \mathbb{N}$. Hence in case of a strictly positive target function F we may take

$$k(\delta) := \min\left(\left\{k \in \mathbb{N}_0 : F(m_\delta^{(k+1)})/F(m_\delta^{(k)}) \geq 1-\epsilon\right\} \cup \{k_{\max}\}\right) \tag{14}$$

for a small constant $\epsilon > 0$ and a large maximal number $k_{\max}$. In the examples discussed subsequently, we found that for $\epsilon = 10^{-5}$ and $k_{\max} = 100$, the number $k(\delta)$ was never larger than 30, which seems to compensate for the fact that the sequence $m_\delta^{(k)}$ converges only geometrically. This is similar to numerical findings with an implementation of an algorithm by [25, section 5] for the median and various parametric regression models.

Having determined $k(\delta)$ and $m_\delta^{(k(\delta))}$ for one particular $\delta > 0$, we define $m_{\delta/2}^{(0)} := m_\delta^{(k(\delta))}$ and repeat the same procedure with $\delta/2$ in place of δ, provided that $k(\delta) > 0$. We proceed until $\delta/2$ would be smaller than a certain threshold $\delta_{\min}$. Pseudocode for this algorithm is displayed in Table 1. Input parameters are F, its regularization $(F_\delta, G_\delta)_{\delta>0}$ augmented with smooth aproximations from above, a starting value $\delta_o > 0$ and a lower threshold $\delta_{\min} \in (0, \delta_o)$ for δ, a starting point $m_o \in \mathcal{C}$, and a threshold $\epsilon > 0$, as well as a maximal iteration number $k_{\max}$ for the inner while-loop.

4. Regularization and quadratic approximation for different types of regression problems. In the subsequent data examples the target functional $F(m)$ is always of type (1) or (3), i.e.,

$$F(m) = \sum_{i=1}^{n} \rho(r_i(m)) + \lambda P(m) \tag{15}$$

with $\lambda \geq 0$, where each residual $r_i(m)$ is an affine linear functional of $m \in \mathbb{R}^d$. Here each summand of F is regularized and approximated separately. We will start with an auxiliary result justifying the quadratic approximation (9).

LEMMA 1. *Let $h : \mathbb{R} \to \mathbb{R}$ be even and twice differentiable such that h'' is nonnegative and nonincreasing on $[0, \infty)$. For $r, z \in \mathbb{R}$ define*

$$g(r, z) := h(r) + (h'(r)/r)(z^2 - r^2)/2,$$

where $h'(0)/0 := h''(0)$. Then $g(r, z) \geq h(z)$ with equality if $z = \pm r$.

Proof. One verifies easily that $g(r, z)$ is even in both arguments with $g(r, r) = h(r)$. Thus it suffices to show that $g(r, z) \geq h(z)$ for any $r, z \geq 0$. Now,

$$\begin{aligned} g(r, z) - h(z) &= g(r, z) - h(r) - (h(z) - h(r)) \\ &= (h'(r)/r)(z^2 - r^2)/2 - h'(r)(z - r) - \int_r^z (h'(t) - h'(r))\, dt \\ &= (h'(r)/r)(z - r)^2/2 - \int_r^z (h'(t) - h'(r))\, dt \\ &= \int_r^z \left(\tilde{h}(r, 0) - \tilde{h}(r, t)\right)(t - r)\, dt \\ &= \int_{\min(r,z)}^{\max(r,z)} \left(\tilde{h}(r, 0) - \tilde{h}(r, t)\right)|t - r|\, dt, \end{aligned} \tag{16}$$

where $\tilde{h}(r, t) := (h'(t) - h'(r))/(t - r)$ for $t \neq r$, and $\tilde{h}(r, r) := h''(r)$. One can deduce easily from h'' being nonincreasing on $[0, \infty)$ that $\tilde{h}(r, \cdot)$ has the same property. Thus the integrand of (16) is nonnegative. ∎

Let us first describe how to approximate ρ itself in three special cases. After this we will discuss several penalizations P in (15). Finally we comment on isotonic regression, an example with $\mathcal{C} \neq \mathbb{R}^d$.

Quantile regression. Let $\rho(z)$ be given by (2). This may be rewritten as

$$\rho(z) = |z| + (2p - 1)z.$$

Hence we utilize the functions h_δ and g_δ from (7) and (9), which yields the regularization

$$z \mapsto h_\delta(z) + (2p - 1)z,$$

and by means of Lemma 1 the quadratic approximation

$$z \mapsto g_\delta(r, z) + (2p - 1)z = c_\delta(r) + h_\delta(r)^{-1} z^2/2 + (2p - 1)z$$

of $z \mapsto \rho(z)$, where $c_\delta(r)$ is an irrelevant constant.

***L*^{q}–regression.** Let $\rho(z) := |z|^q$ for some $q \in [1, \infty)$. If $1 \leq q < 2$, one may generalize definitions (7) and (9) immediately as follows:

$$\begin{aligned} h_\delta(z) &:= (z^2 + \delta)^{q/2}, \\ g_\delta(r, z) &:= h_\delta(r) + q(r^2 + \delta)^{1-q}(z^2 - r^2)/2 \\ &= c_\delta(r) + q(r^2 + \delta)^{1-q} z^2/2. \end{aligned}$$

Again it follows from Lemma 1 that $g_\delta(r, z) \geq h_\delta(z)$ with equality for $z = \pm r$.

In case of $q > 2$, the second derivative of $z \mapsto |z|^q$ is increasing in $|z|$ and unbounded, and hence Lemma 1 cannot be applied directly. To circumvent this problem, one could redefine

$$h_\delta(z) := \begin{cases} |z|^q & \text{if } |z| \leq \delta^{-1} \\ a_\delta + b_\delta |z| + q(q - 1)\delta^{2-q} z^2/2 & \text{otherwise} \end{cases}$$

with constants a_δ, b_δ such that h_δ is twice continuously differentiable, and then use the quadratic approximation

$$g_\delta(r,z) \;:=\; h_\delta(r) + h'_\delta(r)(z-r) + q(q-1)\delta^{2-q}(z-r)^2/2.$$

Logistic regression. For data sets with a covariable X and a dichotomous response $Y \in \{0,1\}$, maximum likelihood estimation of $M(X) := \log[P(Y=1\,|\,X)/P(Y=0\,|\,X)]$ involves "residuals" $z = (1/2 - Y)M(X)$ and

$$\rho(z) := h(z) + z \quad \text{with} \quad h(z) := \log[e^z + e^{-z}].$$

Note that h satisfies the conditions of Lemma 1 with $h'(r) = \tanh(r)$ and $h''(r) = 1 - \tanh(r)^2$. Thus regularization is superfluous, while quadratic approximation is straightforward. In this case, the well-known IRLS algorithm results ([28]).

Roughness penalties. Let us start with two particular examples for $P(m)$. For given real numbers $x_1 < x_2 < \cdots < x_d$ let M be a function on $[x_1, x_d]$ and $m := (M(x_j))_{j=1}^d$. Then let

$$\mathrm{TV}^{(0)}(m) := \sum_{j=1}^{d-1} |m_j - m_{j+1}|,$$
$$\mathrm{TV}^{(1)}(m) := \sum_{j=2}^{d-1} |\Delta_j m| \quad \text{with} \quad \Delta_j m := \frac{m_{j+1}-m_j}{x_{j+1}-x_j} - \frac{m_j - m_{j-1}}{x_j - x_{j-1}}.$$

If M is continuous and piecewise linear with knots in $\{x_1, \ldots, x_d\}$, then $\mathrm{TV}^{(0)}(m)$ and $\mathrm{TV}^{(1)}(m)$ are the total variation of M and its first derivative, respectively. One could also think about smoother functions M and approximate the total variation of its second or higher order derivative by suitable divided differences of m.

Generally, let $P(m)$ be a sum of several functionals of the form

$$m \mapsto |v^\top m|$$

with a given vector $v \in \mathbb{R}^d \setminus \{0\}$. For instance, $\mathrm{TV}^{(0)}(m)$ involves

$$v_i = v_i^{(j)} := \begin{cases} 1 & \text{if } i = j, \\ -1 & \text{if } i = j+1, \\ 0 & \text{else} \end{cases}$$

for $1 \le j < d$, while $\mathrm{TV}^{(1)}(m)$ involves

$$v_i = v_i^{(j)} := \begin{cases} (x_j - x_{j-1})^{-1} & \text{if } i = j-1, \\ -(x_j - x_{j-1})^{-1} - (x_{j+1} - x_j)^{-1} & \text{if } i = j, \\ (x_{j+1} - x_j)^{-1} & \text{if } i = j+1, \\ 0 & \text{else,} \end{cases}$$

for $1 < j < d$. Now an obvious strategy is to regularize $m \mapsto |v^\top m|$ by $m \mapsto h_\delta(v^\top m)$ and approximate this quadratically by

$$m \mapsto g_\delta(v^\top f, v^\top m) = c_\delta(v^\top f) + h_\delta(v^\top f)^{-1}\left(v^\top m\right)^2/2.$$

Often it is desirable to work with quadratic approximations $G(f,\cdot)$ whose Hessian matrix $B(f)$ is diagonal. For that purpose one can modify the quadratic term $Q(m) := \left(v^\top m\right)^2$ as follows:

$$\begin{aligned} Q(m) &= \left(v^\top f\right)^2 + 2f^\top vv^\top(m-f) + \left(v^\top(m-f)\right)^2 \\ &\le \left(v^\top f\right)^2 + 2f^\top vv^\top(m-f) + \|v\|^2 \sum_{i:v_i\neq 0} (m_i - f_i)^2 \\ &= c(v,f) - 2w(v,f)^\top m + \|v\|^2 \sum_{i:v_i\neq 0} m_i^2 \end{aligned}$$

for some irrelevant constant $c(v,f)$ and $w(v,f)_i := 1_{v_i\neq 0}\|v\|^2 f_i - v^\top f v_i$.

Isotonic regression. In some applications one seeks to minimize a functional such as (15) over all vectors in $\mathcal{C}_\nearrow := \{m \in \mathbb{R}^d : m_1 \le \cdots \le m_d\}$. In the simplest case, $d = n$ and $\rho(r_i(m)) = (Y_i - m_i)^q$ for some $q \in [1,\infty]$, where $q = \infty$ corresponds to supremum norm of $Y - m$. For this special case it is well known (see [2]) that an explicit solution exists only for $q = 1, 2, \infty$. In general, via regularization and suitable quadratic approximation from above, each updating step of the GIRLS algorithm involves minimization of

$$G_\delta(f,m) \;=\; C(f) + \sum_{i=1}^d w_i(f)(m_i - b_i(f))^2$$

over all vectors $m \in \mathcal{C}_\nearrow$ with certain weights $w_i(f) > 0$, an irrelevant constant $C(f)$, and certain numbers $b_i(f)$. This minimization problem can be solved explicitly by means of the PAVA ([33]).

5. Numerical examples. In this section we discuss the numerical performance of the GIRLS algorithm in practical applications. Our first example is about quantile regression and shows that GIRLS outperforms a Newton–Raphson type algorithm, at least in the first iterations. In the second example we consider a unimodal regression problem. For ordinary isotonic regression in one-dimensional settings, PAVA is known to be highly efficient (cf., e.g., [33]) in the sense of producing an exact solution in $O(d)$ steps. Replacing the isotonicity constraint on m with a penalty term $\mathrm{TV}^{(0)}(m)$ leads to minimization problems which can be solved efficiently with the closely related "taut string algorithm" (cf. [8]). However, analogous problems in multidimensional regression settings, or using $\mathrm{TV}^{(k)}(m)$ with $k \ge 1$, are computationally more involved. We mention, e.g., [15] for a reformulation as a bilateral contact problem and a rather involved solution by a two-level iterative method requiring a semi-smooth Newton method and a primal dual active set algorithm. GIRLS offers a different strategy: PAVA can be applied after approximating the target functional suitably by a quadratic program; hence, we can also understand GIRLS as a linking device to apply efficient special purpose algorithms for quadratic programs such as PAVA to corresponding nonquadratic and/or constrained problems.

Example 1 (estimation of smooth quantile functions). We applied the GIRLS algorithm and a competitor to a dataset containing the income (X) and the expenditure for food (Y) in the year 1973 for 7125 households in Great Britain (Family Expenditure Survey 1968–1983). This dataset has also been analyzed by [14]. In order to enhance the visual quality, we reduced these data to a random subset of size $n = 2000$. Moreover, since the empirical distributions of both variables are

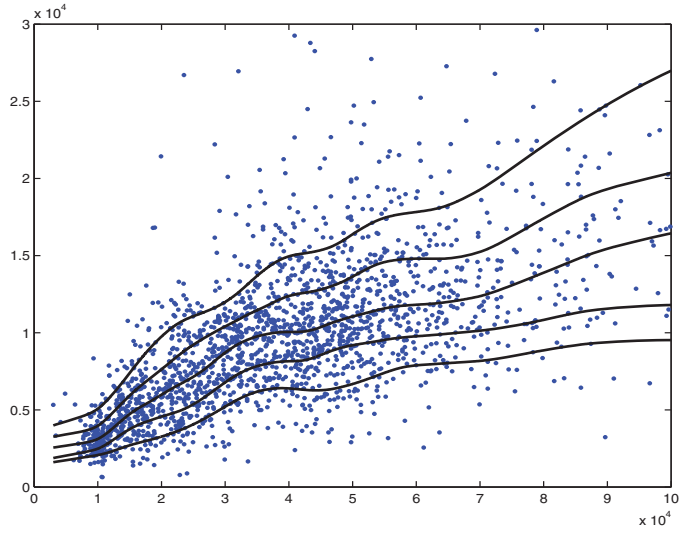


FIG. 1. *Quantile curves for food expenditure data.*

strongly skewed to the right, we plotted only the 1936 pairs (X_i, Y_i) in the range $[0, 10^5] \times [0, 3 \cdot 10^4]$. Figure 1 shows the data together with estimated p–quantile curves M_p for $p = 0.1, 0.25, 0.5, 0.75, 0.9$. Here we used the functional

$$F(m) \;=\; \sum_{i=1}^{n} \rho(Y_i - m_{C(i)}) + \lambda \, \mathrm{TV}^{(1)}(m)$$

on $\mathbb{R}^d$ with ρ given by (2), where $x_1 < x_2 < \cdots < x_d$ are the $d = 1945$ different X–values in the sample, $m_j = M_p(x_j)$, $X_i = x_{C(i)}$, and F was regularized as discussed in the last section. The tuning parameter λ was chosen to be $2 \cdot 10^6$ by visual inspection.

We now turn to the numerical performance of the GIRLS algorithm and compare it to a Newton–Raphson algorithm (robustified with a standard step-size correction as in [12]). In the first step we compare the performance of the two algorithms in the particular setting of our data example; the second step will consist of a small simulation study with artificial data. All computations were performed with Matlab on a 1.86Ghz Pentium-processor with 1GB Ram. As starting value of the iterations we used the polynomial regression p-quantile of order 1, and the tuning parameter δ was selected in a data-driven way from this starting value as the smaller of the median of its absolute residuals, and the median of its first order differences, in both cases divided by 1000. In the first step, we compared the computational efficiency of the GIRLS and the Newton–Raphson algorithm applied to quantile regression with the family expenditure data. To this end we recorded the computing times and number of iterations required for determination of the 25%-quantile curve by GIRLS and the Newton–Raphson algorithm, where we stopped the iterations as soon as the relative improvement of the function $F(m)$ between two subsequent iterations fell below a threshold parameter $\varepsilon = 10^{-12}$. Whereas GIRLS required $5.0s$ CPU time and 59 iterations to find the solution, the Newton–Raphson algorithm turned out to be significantly more expensive with $46.4s$ CPU time and 425 iterations. We also performed the same computations for a wide range of threshold parameters $\varepsilon = 10^{-6} \ldots 10^{-24}$, without significant change in the relative computational expense of the two methods.

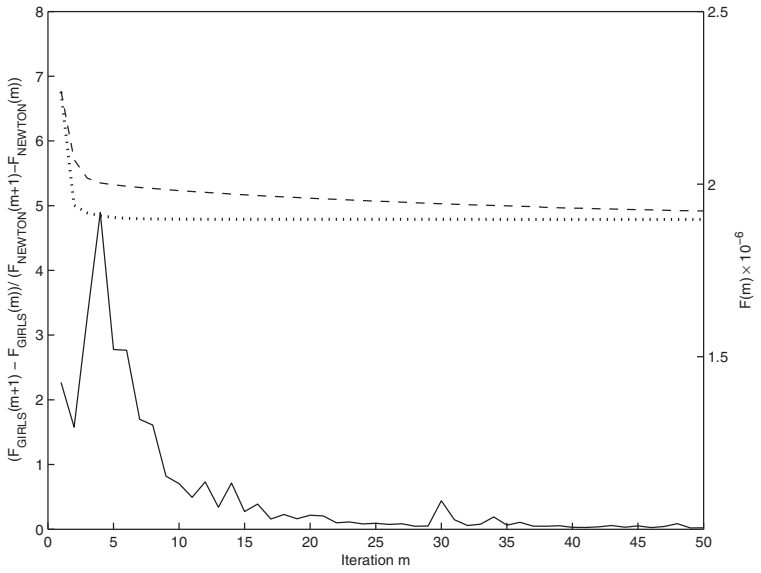


FIG. 2. *Mean performance of GIRLS and a Newton–Raphson-algorithm. The dashed and dotted curves show the mean value of the function $F(m)$ in dependence of the iteration number m for the Newton–Raphson algorithm and GIRLS, respectively, and the solid curve the stepwise rate of improvement $(F_{\mathrm{GIRLS}}(m+1) - F_{\mathrm{GIRLS}}(m)) / (F_{\mathrm{NEWTON}}(m+1) - F_{\mathrm{NEWTON}}(m))$ of $F(m)$ by the two methods. Here, a value larger than 1 indicates that in the mth iteration $F(m)$ decreased more for GIRLS than for the Newton–Raphson-algorithm.*

In the second step of our analysis we performed 100 simulations of the quantile regression problem with (artificial) regression data from the model

$$Y_i = \sin\left(X_i \cdot \frac{\pi}{2}\right) + \varepsilon_i,\ i = 1, \ldots, 1000.$$

Here, $X_i \sim U[0,1]$ are uniformly distributed, independent design variables, and $\varepsilon_i \sim N(0, 0.01)$ i.i.d. noise terms. We used both GIRLS and the Newton–Raphson method to estimate the 25%-quantile curve of the data by determining the minimizer of the function $F(m)$. From visual inspection of a pilot simulation, we chose the tuning parameter $\lambda = 2 \cdot 10^4$ for the subsequent simulations.

Figure 2 compares the mean performance of the GIRLS algorithm with the performance of the Newton–Raphson-method with step-size correction. The two curves in the top show the value of the function $F(m)$ in dependence of the iteration number m, and the bottom curve represents the mean improvement (in the simulations) of the solution, measured by the decrease of the target function $F(m)$. From the figure, we conclude that the first steps of GIRLS are significantly more efficient than for the Newton–Raphson method, whereas the latter catches up after approximately the 8^{th} iteration.

In summary, in the computations with the penalized quantile regression problem, GIRLS outperformed the Newton–Raphson method for the family expenditure data. Moreover, in our simulation we found GIRLS to improve the solution much faster than the Newton–Raphson method in the first ≈ 8 iterations. Only if the initial value of the Newton algorithm has been chosen very close to the true minimizer M^*, we found the performance of the Newton algorithm to be superior. Hence, for practical purposes it seems be advisable to combine both algorithms such that GIRLS will be

used at least as an initial algorithm which efficiently provides a good initial value for a subsequently performed Newton-type algorithm. Moreover, if both GIRLS and the Newton–Raphson method are available, it may be useful to compute both a GIRLS and a Newton step, with subsequent selection of the better one.

Example 2 (GIRLS as a device to utilize efficient algorithms for special quadratic programs in more complex settings). In our second example we will briefly illustrate the flexibility of the GIRLS algorithm to combine several constraints. Precisely, we combine unimodality constraints with TV penalization. Ordinary isotonic regression and hence unimodal regression involves the solution of a weighted least squares problem, and efficient algorithms, the PAVA in particular, are available. If we add a TV penalty, the problem is no longer a quadratic program, and PAVA is not applicable directly. By replacing the problem to be solved by a sequence of quadratic programs, GIRLS makes it possible again to apply PAVA for unimodal regression with TV penalization. To this end consider a two-dimensional regression problem where Y_{ij} are observations to be fitted by $m = (m_{ij})_{i,j}$, $i = 1, \dots, m$, $j = 1, \dots, n$. We want to minimize the sum $F(m)$ of the quadratic cost functional

$$\|Y - m\|^2 = \sum_{ij} (Y_{ij} - m_{ij})^2,$$

and the total variation penalty $\lambda \operatorname{TV}(m)$, where $\lambda > 0$ and

$$\operatorname{TV}(m) := \sum_{i=1}^{m} \sum_{j=1}^{n} |m_{i+1,j} - m_{ij}| + \sum_{i=1}^{m} \sum_{j=1}^{n} |m_{i,j+1} - m_{ij}|.$$

For the regularization F_δ and quadratic approximation $G_\delta(f, \cdot)$, the least squares term is kept unchanged, while each summand $|m_{(a)} - m_{(b)}|$ of $\operatorname{TV}(m)$ is treated as described in section 4: First we regularize it by $h_\delta(m_{(a)} - m_{(b)})$. Then as a first quadratic approximation we may use

$$g_\delta(f_{(a)} - f_{(b)}, m_{(a)} - m_{(b)}) = C_{\delta,(a),(b)}(f) + \frac{(m_{(a)} - m_{(b)})^2}{2h_\delta(f_{(a)} - f_{(b)})}.$$

For our purposes it turns out to be more suitable to replace the enumerator $(m_{(a)} - m_{(b)})^2$ with

$$2\Big(m_{(a)} - \frac{f_{(a)} + f_{(b)}}{2}\Big)^2 + 2\Big(m_{(b)} - \frac{f_{(a)} + f_{(b)}}{2}\Big)^2,$$

which is never less than $(m_{(a)} - m_{(b)})^2$ with equality for $m = f$. The advantages of this latter quadratic approximation are computational simplicity and feasibility of isotonic (and hence unimodal) least squares algorithms. This allows us, e.g., to impose in addition unimodality in vertical direction. Hence, simply in each step for each vertical line the unimodal regression is calculated by means of some standard variation of the PAVA.

Acknowledgments. The authors are thankful to the associate editor and three anonymous referees for their constructive criticism that helped to improve this paper significantly. Moreover, they would like to thank G. Jongbloed and O. Scherzer for helpful comments. The third author is indebted to Y. Vardi for interesting discussions on a previous version of this paper. Parts of this paper were written while

L. Dümbgen was visiting Göttingen University as lecturer within the PhD Program "Applied Statistics and Empirical Methods," and while N. Bissantz was visiting the University of Bern.

REFERENCES

[1] R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk, *Statistical Inference Under Order Restrictions. The Theory and Application of Isotonic Regression,* John Wiley & Sons, London, New York, Sydney, 1972.

[2] R. E. Barlow and V. Ubhaya, *Isotonic approximation,* in Optimisation Methods in Statistics, J. S. Rustagi, ed., Academic Press, 1971, New York, pp. 77–86.

[3] D. Böhning and B. Lindsay, *Monotonicity of quadratic-approximation algorithms,* Ann. Inst. Statist. Math., 40 (1988), pp. 641–663.

[4] A. W. Bowman and A. Azzalini, *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations,* Oxford University Press, Oxford, UK, 1997.

[5] B. M. Brown, *Statistical uses of the spatial median,* J. Roy. Statist. Soc. Ser. B, 45 (1983), pp. 25–30.

[6] B. M. Brown, P. Hall, and G. A. Young, *On the effect of inliers on the spatial median,* J. Multivariate Anal., 63 (1997), pp. 88–104.

[7] D. Coleman, P. Holland, N. Kaden, V. Klema, and S. C. Peters, *A system of subroutines for iteratively reweighted least squares computations,* ACM Trans. Math. Software, 6 (1980), pp. 327–336.

[8] P. L. Davies and A. Kovac, *Local extremes, runs, strings and multiresolution (With discussion and rejoinder by the authors),* Ann. Statist., 29 (2001), pp. 1–65.

[9] J. de Leeuw and G. Michailidis, *Discussion article on the paper by Lange, Hunter & Yang*, J. Comput. Graph. Statist., 9 (2000), pp. 26–31.

[10] Y. Dodge and J. Jurečková, *Adaptive Regression,* Springer-Verlag, New York, 2000.

[11] G. R. Ducharme and P. Milasevic, *Spatial median and directional data,* Biometrika, 74 (1987), pp. 212–215.

[12] L. Dümbgen, S. Freitag-Wolf, and G. Jongbloed, *Estimating a unimodal distribution from interval-censored data,* J. Amer. Statist. Assoc., 101 (2006), pp. 1094–1106.

[13] U. Eckhardt, *Weber's problem and Weiszfeld's algorithm in general spaces,* Math. Program., 18 (1980), pp. 186–196.

[14] W. Härdle and J. S. Marron, *Bootstrap simultaneous error bars for nonparametric regression,* Ann. Statist., 19 (1991), pp. 778–796.

[15] W. Hinterberger, M. Hintermüller, K. Kunisch, M. von Oehsen, and O. Scherzer, *Tube methods for BV regularisation*, J. Math. Imaging Vision, 19 (2003), pp. 219–235.

[16] P. J. Huber, *Robust estimation of a location parameter,* Ann. Math. Stat., 35 (1964), pp. 73–101.

[17] P. J. Huber, *Robust Statistics,* John Wiley & Sons Inc., New York, 1981.

[18] D. R. Hunter and K. Lange, *Quantile Regression via an MM Algorithm,* J. Comput. Graph. Statist., 9 (2000), pp. 60–77.

[19] I. N. Katz, *Local convergence in Fermat's problem,* Math. Program., 6 (1974), pp. 89–104.

[20] K. Lange, D. R. Hunter, and I. Yang, *Optimization Transfer Using Surrogate Objective Functions,* J. Comput. Graph. Statist., 9 (2000), pp. 1–20.

[21] R. Koenker and G. Bassett, *Regression quantiles,* Econometrica, 46 (1978), pp. 33–50.

[22] R. Koenker, P. Ng, and S. Portnoy, *Quantile smoothing splines,* Biometrika, 81 (1994), pp. 673–680.

[23] H. R. Künsch, *Robust priors for smoothing and image restoration,* Ann. Inst. Statist. Math., 46 (1994), pp. 1–19.

[24] H. W. Kuhn, *A note on Fermat's problem,* Math. Program., 4 (1973), pp. 98–107.

[25] M. G. Lejeune and P. Sarda, *Quantile regression: A nonparametric approach,* Comput. Statist. Data Anal., 6 (1988), pp. 229–239.

[26] E. Mammen, J. S. Marron, B. A. Turlach, and M. P. Wand, *A general projection framework for constrained smoothing,* Statist. Sci., 16 (2001), pp. 232–248.

[27] E. Mammen and S. van de Geer, *Locally adaptive regression splines,* Ann. Statist., 25 (1997), pp. 387–413.

[28] P. McCullagh and J. Nelder, *Generalized Linear Models,* 2*nd ed.,* Chapman & Hall, London, 1989.

[29] D. P. O'Leary, *Robust regression computation using iteratively reweighted least squares,* SIAM J. Matrix Anal. Appl., 11 (1990), pp. 466–480.

[30] S. Portnoy, *Local asymptotics for quantile smoothing splines,* Ann. Statist., 25 (1997), pp. 414–434.
[31] D. A. Ratkowsky, *Nonlinear Regression Modelling,* Dekker, New York, 1983.
[32] T. Robertson and P. Waltman, *On estimating monotone parameters,* Ann. Math. Statist., 39 (1968), pp. 1030–1039.
[33] T. Robertson, F. T. Wright, and R. L. Dykstra, *Order Restricted Statistical Inference*, John Wiley & Sons Ltd., Chichester, UK, 1988.
[34] R. T. Rockafellar, *Convex Analysis,* Princeton University Press, Princeton, NJ, 1970.
[35] E. Stein and R. Shakarchi, *Real Analysis*, Princeton Lectures in Analysis, 2005.
[36] Y. Vardi and C.-H. Zhang, *The multivariate L_1-median and associated data depth,* Proc. Natl. Acad. Sci. USA, 97 (2000), pp. 1423–1426.
[37] Y. Vardi and C-H. Zhang, *A modified Weiszfeld algorithm for the Fermat-Weber location problem,* Math. Program. Ser. A, 90 (2001), pp. 559–566.
[38] C. R. Vogel and M. E. Oman, *Iterative methods for total variation denoising*, SIAM J. Sci. Comput., 17 (1996), pp. 227–238.
[39] H. Voss and U. Eckhardt, *Linear Convergence of generalized Weiszfeld's method,* Computing, 25 (1980), pp. 243–251.
[40] E. Weiszfeld, *Sur un problème de minimum dans l'espace,* Tohoku Math. J., 42 (1936), pp. 274–280.
[41] E. Weiszfeld, *Sur le point pour lequel la somme des distances de n points donnés est minimum,* Tohoku Math. J., 43 (1937), pp. 355–386.
[42] R. Wolke and H. Schwetlick, *Iteratively reweighted least squares: Algorithms, convergence analysis, and numerical comparisons,* SIAM J. Sci. Statist. Comput., 9 (1988), pp. 907–921.

SIAM J. OPTIM.
Vol. 19, No. 4, pp. 1846–1880

ON THE IMPLEMENTATION OF INTERIOR POINT DECOMPOSITION ALGORITHMS FOR TWO-STAGE STOCHASTIC CONIC PROGRAMS*

SANJAY MEHROTRA† AND M. GÖKHAN ÖZEVIN‡

Abstract. In this paper we develop a practical primal interior decomposition algorithm for two-stage stochastic programming problems. The framework of this algorithm is similar to the framework in [S. Mehrotra and M. G. Özevin, "Decomposition based interior point methods for two-stage stochastic convex quadratic programs with recourse," to appear in *Oper. Res.*; S. Mehrotra and M. G. Özevin, *SIAM J. Optim.*, 18 (2007), pp. 206–222] and [G. Zhao, *Math. Program.*, 90 (2001), pp. 507–536], however, their algorithm is altered in a simple yet fundamental way to achieve practical performance. In particular, this new algorithm weighs the log-barrier terms in the second stage problems differently from the theoretical algorithms analyzed in [S. Mehrotra and M. G. Özevin, *Oper. Res.*, to appear], [S. Mehrotra and M. G. Özevin, *SIAM J. Optim.*, 18 (2007), pp. 206–222], and [G. Zhao, *Math. Program.*, 90 (2001), pp. 507–536]. We give a method for generating a suitable starting point; a method for selecting a good starting barrier parameter; a heuristic for first stage step-length calculation without performing line searches; and a method for adaptive addition of new scenarios over the course of the algorithm. The decomposition algorithm is implemented to solve two-stage stochastic conic programs with recourse whose underlying cones are Cartesian products of linear, second order, and semidefinite cones. The performance of primal decomposition method is studied on a set of randomly generated test problems as well as a two-stage stochastic programming extension of the Markowitz portfolio selection model. The computational results show that an efficient and stable implementation of the primal decomposition method is possible. These results also show that in problems with a large number of scenarios, the adaptive addition of scenarios can yield computational savings of up to 80%.

Key words. stochastic programming, conic programming, semidefinite programming, Benders decomposition, interior point methods, primal-dual methods

AMS subject classifications. 90C, 90C15, 90C22, 90C25, 90C51

DOI. 10.1137/050643805

1. Introduction.

We consider the two-stage stochastic conic programming (TSSCP) problem in the form

$$\begin{aligned} &\max b^T x - \frac{1}{2} x^T G x + \rho(x) \\ &\text{s.t. } Dx = d, \\ &\qquad Ax + s_1 = c, \\ &\qquad s_1 \in \mathcal{K}_1, \end{aligned} \tag{1.1}$$

where

$$\rho(x) := \sum_{i=1}^{K} \pi_i \rho_i(x) \tag{1.2}$$

*Received by the editors October 31, 2005; accepted for publication (in revised form) September 30, 2008; published electronically February 27, 2009.
http://www.siam.org/journals/siopt/19-4/64380.html

†Corresponding author. Department of Industrial Engineering & Management Sciences, Northwestern University, Evanston, IL 60208 (mehrotra@northwestern.edu).

‡ZS Associates, Evanston, IL 60201 (gokhan.ozevin@zsassociates.com). Research was supported in part by grants NSF-DMI-0200151, DMI-0522765, and ONR-N00014-01-1-0048 and was performed while the author was at Northwestern University.

and

$$\begin{aligned} \rho_i(x) := \max\ & f_i^T y_i - \frac{1}{2} y_i^T H_i y_i \\ \text{s.t.}\ & R_i y_i + U_i x = z_i, \\ & W_i y_i + T_i x + s_{2,i} = h_i, \\ & s_{2,i} \in \mathcal{K}_{2,i}. \end{aligned} \tag{1.3}$$

The symmetric cones $\mathcal{K}_1$ and $\mathcal{K}_{2,i}$ are Cartesian products of nonnegative orthant, second order cones and the cones of positive semidefinite matrices. By $\mathcal{K}_1^+$ and $\mathcal{K}_{2,i}^+$ we represent the interior of these cones. The size of these cones in the first stage and second stage problems may be different. The columns of submatrices of A, W_i, and T_i corresponding to the semidefinite cones are vectorizations of symmetric matrices. In other words, the cone of semidefinite matrices are considered as a set of vectors to comply with the above standard form of TSSCP. The matrices G and H_i are symmetric positive semidefinite.

Faybusovich [6, 7] presented a unified description of interior point algorithms for linear, second order cones and semidefinite programming using the theory of Euclidean Jordan algebra. We also describe our interior decomposition algorithms using the Jordan algebra operations. For a brief introduction of Jordan algebra operations, see [1, 27, 28].

Let us define the following feasibility sets:

$$\begin{aligned} &\mathcal{F}^0 := \{(x, s_1) \mid Dx = d, Ax + s_1 = b,\ s_1 \in \mathcal{K}_1\}, \\ &\mathcal{F}_{2,i}(x) := \{(y_i, s_{2,i}) \mid R_i y_i + U_i x = z_i, W_i y_i + s_{2,i} = h_i - T_i x,\ s_{2,i} \in \mathcal{K}_{2,i}\}, \\ &\mathcal{F}_i^1 := \{(x, s_1) \mid \mathcal{F}_{2,i}(x) \neq \emptyset\}, \\ &\mathcal{F}^1 := \{\cap_{i=1}^K \mathcal{F}_i^1\} \cap \mathcal{F}^0, \\ &\mathcal{F} := \{(x, s_1) \times (y_1, s_{2,1}, \ldots, y_K, s_{2,K}) \mid Dx = d, Ax + s_1 = b, \\ &\qquad s_1 \in \mathcal{K}_1;\ R_i y_i + U_i x = z_i, W_i y_i + s_{2,i} = h_i - T_i x,\ s_{2,i} \in \mathcal{K}_{2,i}\}. \end{aligned}$$

Here $\mathcal{F}_{2,i}(x)$ is the second stage feasibility set parameterized by the first stage decision vector x, $\mathcal{F}^1$ is the set of first stage feasible solutions for which all second stage problems are also feasible, and $\mathcal{F}$ is the set of feasible solutions for the extensive formulation of TSSCP.

Mehrotra and Özevin [18] presented a decomposition-based primal interior algorithm for the two-stage stochastic semidefinite programming problem with recourse. In their setting the first stage cone $\mathcal{K}_1$ and the second stage cones $\mathcal{K}_{2,i}$ are $p \times p$ and $r \times r$ symmetric positive semidefinite matrices, respectively. Mehrotra and Özevin [18] show that starting from a well-centered first stage solution (starting barrier parameter μ^0) a short-step path-following interior point algorithm working on the first stage variables requires $O(\sqrt{p + rK}) \ln(\mu^0/\mu^*)$ first stage interior point iterations to obtain a well-centered solution for barrier parameter μ^*, where K is the number of scenarios. Their analysis extends the earlier results of Zhao [31] for the two-stage linear stochastic programming problems and the results of Mehrotra and Özevin [17] for the quadratic case. The work of Zhao [31] and Mehrotra and Özevin [17, 18] provides a framework for implementing primal interior decomposition algorithms for two-stage stochastic programming problems. Several simplifying assumptions were made by Zhao [31] and Mehrotra and Özevin [17, 18] while analyzing their algorithms. The

purpose of this paper is to develop a more practical primal interior decomposition algorithm for two-stage stochastic conic programs.

The basic steps of a primal interior decomposition algorithm are as follows.

ALGORITHM 1. A BASIC PRIMAL INTERIOR DECOMPOSITION FRAMEWORK.

Initialization.

Let x^1 be a starting point, $\mu^1 > 0$ a starting barrier parameter, and μ^* the desired termination value of the barrier parameter. Also, let $\beta > 0, \zeta \in (0,1)$, and $\theta > 0$ be suitable scalar parameters.

Step 1.

Step 1.1. Solve all second stage centering problems for the current x and μ.

Step 1.2. Using the second stage solutions, compute the first stage Newton direction.

Step 1.3. Compute the local norm $\delta(\mu, x)$ of the Newton direction Δx as a measure of distance from the current point x to the first stage μ-center.

Step 1.4. Update first stage solution as $x := x + \theta \Delta x$. If $\delta(\mu, x) \leq \beta$, go to Step 2; otherwise go to Step 1.1.

Step 2. If $\mu \leq \mu^*$ (or some other termination criterion is satisfied) stop; otherwise reduce $\mu = \zeta\mu$ and go to Step 1.1.

Note that a variant of Algorithm 1 that reduces μ rapidly (long-step variant) may iterate several times in the loop Step 1.1–Step 1.4 after reducing the value of μ in Step 2. In the short-step method this loop is executed only once, but this method reduces μ slowly. The iterates in the loop Step 1.1–Step 1.4 are called the inner iterations of Algorithm 1. These iterations constitute most of the work in Algorithm 1. Theoretical values for parameters β, ζ, θ, and the choice of $\delta(\mu, x)$ are given in [31, 17, 18] in the context of proving polynomial time convergence of short-step and long-step methods.

This paper develops a different primal interior decomposition algorithm from the one analyzed in [31, 17, 18]. While the algorithm analyzed in [31, 17, 18] has a better worst-case theoretical complexity than the one proposed in this paper, computational experience suggests that the new proposed algorithm gives better performance in practice. The proposed algorithm differs from the algorithms in [31, 17, 18] in Step 1.1. In particular, it defines the second stage centering problems differently. We will present this algorithm and compare it with the algorithms in [31, 17, 18] in section 2. We have some probabilistic/heuristic justifications for the superior performance of the proposed algorithm, which we intend to document elsewhere.

In the primal decomposition framework (for algorithms in [31, 17, 18], as well as the new algorithm) we need to address several issues to develop practical efficient implementations. These issues are the following:

(i) (Initialization.) We need to find a suitable feasible first stage solution and ensure the feasibility of all second stage problems corresponding to this solution. If TSSCP satisfies a full recourse assumption, i.e., all second stage problems are feasible for all first stage solutions, then feasibility of second stage is not an issue. However, this assumption may not hold in general.

(ii) (Steps 1.1, 1.2.) The assumption of the availability of exact solutions of the second stage centering problems in [31, 17, 18] considerably simplifies the convergence analysis. However, solving second stage problems exactly is not possible. The practical implications of working with inexact solutions on first stage iterations and algorithmic robustness are not clear.

(iii) (Step 1.3.) It is not clear how closely the algorithm should follow the first stage central path (value of parameter β) for it to be practical and robust. Following the central path too closely would unnecessarily increase computational efforts, whereas if we ignore the central path the algorithm may not converge.

(iv) (Step 1.4.) The theoretical analysis assumes taking fixed steps along the Newton direction. This is conservative since taking larger steps give faster convergence. We need to find a good way to select the first stage step-length θ without performing line searches. Performing a line search over the barrier objective (defined in section 2) is not practical since an evaluation of this function for a given μ and x requires solutions of all second stage problems, which is equivalent to computing a new first stage Newton direction.

(v) (Step 2.) A practical value of ζ.

(vi) (Adaptive addition of scenarios.) The primal decomposition interior point methods for two-stage stochastic programming are appealing because the computation of the first stage Newton direction decomposes across second stage scenarios. This naturally allows the possibility of increasing the number of second stage scenarios as the algorithm progresses. The analysis in [31, 17, 18] is for a fixed number of scenarios. Note that adding a significant number of new scenarios modifies the recourse function and changes the first stage central path. A theoretical analysis of this change is challenging. However, adding scenarios adaptively as the algorithm progresses has the potential of saving computational efforts, especially when the number of scenarios is very large. How can adaptive addition of scenarios be implemented in practice?

(vii) (Warm-start.) Is it possible to use solutions from previously solved second stage problems to warm-start new related second stage problems? There are two situations where a warm-start might be possible. The first possibility is when x and/or μ changes in Steps 1.4 and 2 in a preexisting scenario. The second possibility is when scenarios are added adaptively, and a centered solution of a new scenario is desired for the current x and μ.

In stochastic programming we have the additional issue of determining the number of scenarios (when the uncertain parameters have continuous distribution, or have a huge finite support) required to ensure a desired quality of solution. This is achieved through generating lower and upper bounds, and a statistical analysis around these bounds. Methods for generating such bounds are discussed in Linderoth, Shapiro, and Wright [12] and are beyond the scope of this paper. In this paper we will assume that the problem has a given large number of scenarios.

Within the context of the proposed practical algorithm we provide resolution of (i)–(vii) to various extents. We give numerical results for the proposed implementation strategies. Computational results are presented in sections 4–7. These results are obtained on an extension of the classical Markowitz portfolio optimization problem and a set of randomly generated conic programming test problems that have linear, second order, and semidefinite cones. The test problems are described in section 3. Section 4 focuses on various algorithmic parameters and line search in the primal decomposition algorithm. Section 5 presents results on various characteristics of the algorithm exhibited by the problems we solved. Section 6 develops a methodology for adaptive addition of scenarios, and section 7 discusses results obtained using scenario addition and warm-start during scenario addition. The computational results were obtained using MATLAB version 6.5.1 on an IBM T42 using Pentium M 1.7 GHz with 1 GB of RAM. The entire code is written using MATLAB's internal libraries.

2. Primal interior decomposition algorithms. In this section we describe the algorithm we have found to work better in practice. We compare this algorithm with the algorithms presented and analyzed in Zhao [31] and Mehrotra and Özevin [17, 18]. The algorithm of [31, 17, 18] is described in section 2.1. In section 2.2 we describe the proposed practical algorithm, and in section 2.3 we compare the two algorithms computationally.

2.1. Zhao [31], Mehrotra and Özevin [17, 18] primal decomposition algorithm. In [31, 17, 18] the authors redefine the recourse function in (1.2) as

$$\rho(x) := \sum_{i=1}^{K} \bar{\rho}_i(x),$$

where

$$\begin{aligned} \bar{\rho}_i(x) := \max\ & \bar{f}_i^T y_i - \frac{1}{2} y_i^T \bar{H}_i y_i \\ \text{s.t. } & (y_i, s_{2,i}) \in \mathcal{F}_{2,i}(x), \end{aligned} \tag{2.1}$$

$\bar{f}_i = \pi_i f_i$, and $\bar{H}_i = \pi_i H_i$. They consider the log-barrier problem associated with (2.1):

$$\begin{aligned} \bar{\rho}_i(\mu, x) := \max\ & \bar{f}_i^T y_i - \frac{1}{2} y_i^T \bar{H}_i y_i + \mu \ln\det(s_{2,i}) \\ \text{s.t.} & (y_i, s_{2,i}) \in \mathcal{F}_{2,i}(x). \end{aligned} \tag{2.2}$$

The first stage log-barrier problem in [31, 17, 18] is defined as

$$\begin{aligned} \max\ & \bar{\eta}(\mu, x) := b^T x - \frac{1}{2} x^T G x + \bar{\rho}(\mu, x) + \mu \ln\det(s_1) \\ \text{s.t. } & (x, s_1) \in \mathcal{F}^1, \end{aligned} \tag{2.3}$$

where

$$\bar{\rho}(\mu, x) := \sum_{i=1}^{K} \bar{\rho}_i(\mu, x).$$

For a given x and μ, the optimality conditions for (2.2) are

$$\begin{aligned} & R_i y_i + U_i x = z_i, \\ & W_i y_i + T_i x + s_{2,i} = h_i, \\ & W_i^T \bar{\lambda}_i + R_i^T \bar{\gamma}_i + \bar{H}_i y_i = \bar{f}_i, \\ & P(\bar{\lambda}_i^{1/2}) s_{2,i} = \mu \iota_{2,i}, \\ & \bar{\lambda}_i \in \mathcal{K}_{2,i}^+,\ s_{2,i} \in \mathcal{K}_{2,i}^+. \end{aligned} \tag{2.4}$$

Let $x(\mu) := \text{argmax}\{\bar{\eta}(\mu, x)\}$ for a given $\mu > 0$. We refer to $x(\mu)$ as the first stage μ-center and to the trajectory $\{x(\mu),\ \mu > 0\}$ as the first stage central path. We denote the unique solution of (2.4) for any given $x \in \mathcal{F}_i^1$ and $\mu > 0$ by $(y_i(\mu, x),\ s_{2,i}(\mu, x),\ \bar{\lambda}_i(\mu, x),\ \bar{\gamma}_i(\mu, x)),\ i = 1, \ldots, K$, and call this solution the second

stage μ-center. The μ-centers for different values of μ form the second stage central path for a given x as μ decreases from ∞ to zero.

A brief summary of the key Jordan algebra operations is as follows. In (2.4) the matrix $P(\lambda_i^{1/2})$ is the *Jordan quadratic presentation* of the vector $\lambda_i^{1/2}$ and $\iota_{2,i}$ is the *identity solution* of the cone $\mathcal{K}_{2,i}$ [6, 7]. The vector $\lambda_i^{1/2}$ is the *square root* (in the Jordan algebra sense) of the dual multiplier λ_i, that is, $\lambda_i^{1/2}$ is a unique vector such that $\lambda_i = P(\lambda_i^{1/2})\iota$. For any given symmetric cone $\mathcal{K}$ and a given $\lambda \in \mathcal{K}$, the *identity solution* ι of $\mathcal{K}$ is a unique element of $\mathcal{K}$ such that $L(\lambda)\iota = \lambda$ for all $\lambda \in \mathcal{K}$, where $L(\lambda)$ is the multiplication by λ linear operator of the cone $\mathcal{K}$. Furthermore, the vector λ^{-1} is the unique inverse of $\mathcal{K}^+$ such that $P(\lambda)\lambda^{-1} = \lambda$. The vector $\lambda^{1/2}$ is the *square root* of $\lambda \in \mathcal{K}$ such that $L(\lambda^{1/2})\lambda^{1/2} = \lambda$. The gradient and Hessian of the function $\ln\det(\lambda)$ has the form

$$\nabla \ln\det(\lambda) = \lambda^{-1},\ \lambda \in \mathcal{K}^+,$$
$$\nabla^2 \ln\det(\lambda) = -P(\lambda^{-1}),\ \lambda \in \mathcal{K}^+.$$

The form of $P(\lambda)$ in case of primitive symmetric cones as well as Cartesian products of primitives are given in [27]. The gradient and Hessian of the barrier objective $\bar{\eta}(\mu, x)$ are calculated as follows:

$$\nabla\bar{\eta}(\mu,x) = b - Gx + \sum_{i=1}^{K}(T_i^T\bar{\lambda}_i(\mu,x) + U_i^T\bar{\gamma}_i(\mu,x)) - \mu A^T s_1^{-1}, \tag{2.5}$$

$$\nabla^2\bar{\eta}(\mu,x) = -G + \sum_{i=1}^{K}(T_i^T\nabla\bar{\lambda}_i(\mu,x) + U_i^T\nabla\bar{\gamma}_i(\mu,x)) - \mu A^T P(s_1^{-1})A. \tag{2.6}$$

The quantities $\nabla\bar{\lambda}(\cdot)$, and $\nabla\bar{\gamma}_i(\cdot)$ are calculated as follows. Differentiating the optimality conditions (2.4) with respect to x, we obtain

$$\begin{aligned} &R_i\nabla y_i = -U_i,\\ &W_i\nabla y_i + \nabla s_{2,i} = -T_i,\\ &W_i^T\nabla\bar{\lambda}_i + R_i^T\nabla\bar{\gamma}_i + \bar{H}_i\nabla y_i = 0,\\ &P(\bar{\lambda}_i^{1/2})\nabla s_{2,i} + P(s_{2,i}^{1/2})\nabla\bar{\lambda}_i = 0. \end{aligned} \tag{2.7}$$

Solving (2.7), we get

$$\begin{aligned} &\nabla\bar{\gamma}_i = (R_iZ_i^{-1}R_i^T)^{-1}(U_i - R_iZ_i^{-1}W_i^TQ_iT_i),\\ &\nabla y_i = -Z_i^{-1}(R_i^T\nabla\bar{\gamma}_i + W_i^TQ_iT_i),\\ &\nabla s_{2,i} = -(T_i + W_i\nabla y_i),\\ &\nabla\bar{\lambda}_i = -Q_i\nabla s_{2,i}, \end{aligned} \tag{2.8}$$

where

$$Q_i := (P(s_{2,i}^{1/2}))^{-1}P(\bar{\lambda}_i^{1/2}) \text{ and } Z_i := W_i^TQ_iW_i + \bar{H}_i.$$

For a fixed μ, the solution $(x(\mu), s_1(\mu); y_i(\mu,\ x(\mu)), s_{2,i}(\mu,\ x(\mu)),\ i = 1,\dots,K)$ of (2.2)–(2.3) is the same as the maximizer of the log-barrier function

$$b^Tx - \frac{1}{2}x^TGx + \sum_{i=1}^{K}\left(\bar{f}_iy_i - \frac{1}{2}y_i^T\bar{H}_iy_i\right) + \mu\ln\det(s_1) + \mu\sum_{i=1}^{K}\ln\det(s_{2,i}) \tag{2.9}$$

over the set $\mathcal{F}$.

In the basic primal decomposition framework (Algorithm 1) the first stage Newton direction Δx (Step 1.2) is the optimal solution of

$$\begin{aligned} &\max \nabla\bar{\eta}(x)^T\Delta x + \frac{1}{2}\Delta x^T\nabla^2\bar{\eta}(x)\Delta x \\ &\text{s.t.} \quad D\Delta x = 0. \end{aligned} \tag{2.10}$$

In Step 1.3, the local norm of the Newton step calculated at x is given by

$$\delta(\mu, x) = \sqrt{-\frac{1}{\mu}\Delta x^T\nabla^2\bar{\eta}(\mu, x)\Delta x}.$$

For two-stage stochastic semidefinite programs Mehrotra and Özevin [18] show that the log-barrier recourse function $\bar{\eta}(\mu, x)$ is a strongly μ-self-concordant function and forms a strongly self-concordant family. This is used to establish convergence for short- and long-step path-following algorithms as discussed in the introduction.

2.2. A practical primal decomposition algorithm. We now describe a more practical primal decomposition algorithm within the framework of Algorithm 1. In this algorithm we work directly with the form of the second stage problem (1.3) without scaling the second stage objectives as is done in (2.1). Let us consider the log-barrier problem associated with (1.3):

$$\begin{aligned} \rho_i(\mu, x) := \ &\max f_i^T y_i - \frac{1}{2}y_i^T H_i y_i + \mu \ln\det(s_{2,i}) \\ &\text{s.t.} \ \ (y_i, s_{2,i}) \in \mathcal{F}_{2,i}(x). \end{aligned} \tag{2.11}$$

The first stage log-barrier problem is now defined as

$$\begin{aligned} &\max \eta(\mu, x) := b^T x - \frac{1}{2}x^T G x + \mu \ln\det(s_1) + \rho(\mu, x) \\ &\text{s.t.} \quad x \in \mathcal{F}^1, \end{aligned} \tag{2.12}$$

where

$$\rho(\mu, x) := \sum_{i=1}^{K} \pi_i \rho_i(\mu, x). \tag{2.13}$$

We can show that for a fixed μ, the solutions $(x(\mu), s_1(\mu); y_i(\mu, x(\mu)), s_{2,i}(\mu, x(\mu)),$ $i = 1, \ldots, K)$ of (2.11)–(2.12) are the same as the maximizer of the log-barrier function

$$b^T x - \frac{1}{2}x^T G x + \sum_{i=1}^{K} \pi_i \left(f_i y_i - \frac{1}{2}y_i^T H_i y_i \right) + \mu \ln\det(s_1) + \mu \sum_{i=1}^{K} \pi_i \ln\det(s_{2,i}) \tag{2.14}$$

over $\mathcal{F}$.

Note that in contrast to (2.9), in (2.14) the second stage log-barrier terms are scaled with the corresponding probabilities. Hence, when solving the second stage problems to compute the Newton direction for a given x, solutions are computed for a smaller value of μ (scaled by π_i). We do not have an obvious analogue of the result that $\bar{\eta}(\mu, x)$ is a strongly μ-self-concordant function for the function $\eta(x, \mu)$. An inferior worst-case theoretical bound is proved in Mehrotra and Özevin [19]. However, a probabilistic analysis in Mehrotra [16] shows that a better convergence result

(possibly independent of the number of scenarios) is possible under a probabilistic self-concordance assumption.

The algebraic development of the computation of Newton direction for (2.12) is similar to the development for (2.3). For a given x and μ, the optimality conditions of (2.11) are

$$\begin{aligned} & R_i y_i + U_i x = z_i, \\ & W_i y_i + T_i x + s_{2,i} = h_i, \\ & W_i^T \lambda_i + R_i^T \gamma_i + H_i y_i = f_i, \\ & P(\lambda_i^{1/2}) s_{2,i} = \mu \iota_{2,i}, \\ & \lambda_i \in \mathcal{K}_{2,i}^+, \ s_{2,i} \in \mathcal{K}_{2,i}^+, \end{aligned} \tag{2.15}$$

where $\iota_{2,i}$ is defined as before. The gradient and Hessian of the barrier objective $\eta(\mu, x)$ are calculated as follows:

$$\nabla \eta(\mu, x) = b - Gx - \sum_{i=1}^{K} \pi_i (T_i^T \lambda_i(\mu, x) + U_i^T \gamma_i(\mu, x)) - \mu A^T s_1^{-1}, \tag{2.16}$$

$$\nabla^2 \eta(\mu, x) = -G - \sum_{i=1}^{K} \pi_i (T_i^T \nabla \lambda_i(\mu, x) + U_i^T \nabla \gamma_i(\mu, x)) - \mu A^T P(s_1^{-1}) A. \tag{2.17}$$

In the basic primal decomposition framework (Algorithm 1) the first stage Newton direction (Step 1.2) is the optimal solution of

$$\begin{aligned} \max \ & \nabla \eta(x)^T \Delta x + \frac{1}{2} \Delta x^T \nabla^2 \eta(x) \Delta x \\ \text{s.t.} \quad & D \Delta x = 0, \end{aligned} \tag{2.18}$$

which can be calculated by solving the KKT system associated with (2.18):

$$\begin{aligned} & \nabla^2 \eta(x) \Delta x - D^T \tau = -\nabla \eta(x) \\ & D \Delta x = 0. \end{aligned} \tag{2.19}$$

In Step 1.3, the local norm of the Newton step calculated at x becomes

$$\delta(\mu, x) = \sqrt{-\frac{1}{\mu} \Delta x^T \nabla^2 \eta(\mu, x) \Delta x}.$$

2.3. Performance comparison of basic primal decomposition algorithms. The computations of Newton direction in sections 2.1 and 2.2 were implemented using the same starting point, parameter settings, and subroutines. These settings are discussed in the next section in the context of the algorithm in section 2.2. We ensured that same relative precision is achieved when solving the second stage KKT systems (2.4) and (2.15) to compute the second stage solutions in their respective problems. Table 1 gives results on a set of 100 scenario test problems. These problems are described in section 3.2. Columns "Itr." give the number of inner iterations in both algorithms, and columns "Convergence" indicate if convergence to eight digits of accuracy was achieved for these problems when the algorithms terminate. A failure to converge to a solution with eight digits of accuracy is indicated by "No." Various reasons for this failure are explained in section 4.2. Superiority of the algorithm in section 2.2 is obvious.

TABLE 1
Comparison of theoretical and practical algorithms ($\beta = 1$, $\nu = 1$, $\zeta = 0.1$; using line search).

	Section 2.1 algorithm		Section 2.2 algorithm	
Problem	Itr.	Convergence	Itr.	Convergence
MCR1-100	318	No	22	Yes
MCR2-100	236	No	27	Yes
MCR3-100	290	Yes	22	Yes
MCR4-100	254	No	29	Yes
MCR5-100	215	No	38	Yes
MCR6-100	323	Yes	35	Yes

3. Test problems. We studied the performance of our decomposition algorithm on randomly generated two-stage stochastic conic programs and a specific stochastic programming problem arising from a two-stage extension of the classical Markowitz's mean-variance model. We now describe the generation of these test problems.

3.1. An application from finance: Two-stage extension of Markowitz's mean-variance model. The seminal mean-variance model of Markowitz [14] provides a quantitative framework for establishing a balance of the risk and return characteristics of various asset classes. In this model, return on the portfolio is measured by the expected value of the portfolio return, and the associated risk is quantified by the variance of the portfolio return.

Consider a portfolio that invests in n assets over a single period. Denote by $x \in \mathbb{R}^n$ the portfolio vector, and by $\tilde{r} \in \mathbb{R}^n$ the random vector of asset returns over a given horizon. The mean-variance setting assumes that $\tilde{r}$ has a multivariate normal distribution with mean $\bar{r}$ and covariance Q. The corresponding mean-variance problem is given by $\min\{w^T Q w : \bar{r}^T w \geq \rho, e^T w = 1\}$. By minimizing the portfolio variance by varying the level of expected return ρ we can derive the so-called mean-variance efficient frontier.

We formulate a two-stage extension of the static single-period Markowitz model to address these issues. Let us first introduce some notation in preparation to formulating this model. This model can be viewed as an alternative to the robust optimization model proposed by Goldfarb and Iyengar [9] in incorporating uncertainty in the covariance matrix. The current time is denoted by t_1. At time t_1 after having observed the return vector r_0 the investor forms a portfolio and at time $t_2 > t_1$ he can revise his portfolio. The revised portfolio is kept until t_3. The problem is to decide a portfolio at time t_1 in anticipation of its revision at time t_2. The evolution of the asset return vector and the covariance matrix is approximated with K scenarios, which we index by $i = 1, \ldots, K$.

We define the following decision variables and parameters:

- w_1 := vector of first stage portfolio positions,
- r_1 := vector of asset returns in the first period,
- $\bar{r}_1$:= expected asset returns in the first period,
- $\bar{r}_1^{min}$:= lower bound for the expected asset returns in the first period,
- Q_1 := covariance of asset returns in the first period,
- $w_{2,i}$:= vector of second stage portfolio positions under scenario i,
- $r_{2,i}$:= vector of asset returns in the second period under scenario i,
- $\bar{r}_{2,i}$:= expected asset returns in the second period under scenario i,
- $\bar{r}_{2,i}^{min}$:= minimum required return at the end of the second period under scenario i,
- $Q_{2,i}$:= covariance of asset returns in the second period under scenario i.

We have the following two-stage extension of the basic mean-variance model:

$$\begin{aligned}
&\min {w_1}^T Q_1 w_1 + \sum_{i=1}^{K} \frac{1}{K} w_{2,i}^T Q_{2,i} w_{2,i} \\
&\text{s.t. } \bar{r}_1^T w_1 \geq \bar{r}_1^{min}, \\
&\quad e^T w_1 = 1, \\
&\quad \bar{r}_{2,i}^T w_{2,i} \geq \bar{r}_{2,i}^{min},\ i = 1, \ldots, K, \\
&\quad e^T w_{2,i} = 1,\ i = 1, \ldots, K, \\
&\quad \|w_1 - w_{2,i}\| \leq \tau_{2,i},\ i = 1, \ldots, K. \qquad (3.1)
\end{aligned}$$

In the above model, the magnitude of the variation between first and second stage solutions is bounded by the constraints (3.1). In practice an investor would avoid large changes between the optimal first and second stage portfolio positions since such changes can lead to prohibitive amounts of transaction costs. Moreover, limiting the variation between the first and second period portfolio positions would make first period optimal portfolio composition less sensitive to perturbations in the problem parameters. Alternatively, the model may be viewed as a way of finding a "centroid" solution when multiple estimates of r and Q exist. In both respects, one may prefer our model to the single-period robust portfolio optimization model introduced in Goldfarb and Iyengar [9], since robust models tend to be pessimistic.

For the test problems generated for this paper we use a multivariate GARCH model to describe the return process:

$$\begin{aligned}
r_t &= \phi_0 + \phi_1 r_{t-1} + \epsilon_t,\ \epsilon_t \sim N(0, Q_t), \\
Q_t &= C + A * H_{t-1} + B * Q_{t-1}, \qquad (3.2)
\end{aligned}$$

where the symbol $*$ denotes the Hadamard product of two matrices and $H_t = [\epsilon_{i,t}\ \epsilon_{j,t}]_{i,j=1,\ldots,n}$.

We randomly generated the model parameters ϕ_0, ϕ_1, A, C, and B, as well as the starting asset returns vector r_1 and covariance Q_1. The matrices A, B, and C are generated as symmetric positive semidefinite matrices to ensure that the forecasted conditional covariance matrices Q_t are positive semidefinite [11]. We calibrated the probability distributions from which we sample A, B, and C such that the forecasted asset returns in general fall into the $[0.5, 1.5]$ range, volatilities vary around 10%, and correlations take values between -0.5 and 0.5. These parameter ranges allow us to generate problems that are realistic, but also have significant volatility to test the algorithm.

We generate scenarios $(\bar{r}_{2,i}, Q_{2,i})$ using an n-dimensional Sobol' sequence [5, 8], where n is the number assets in the portfolio. To obtain a set of K samples, $(\epsilon_{1,1}, \ldots, \epsilon_{1,K})$, of the residual vector ϵ_1, we start by generating the K n-dimensional Sobol' points $(S_1, \ldots, S_K)$. Then we set $\epsilon_{1,i} = Q_1^{1/2} \Phi^{-1}(S_i), i = 1, \ldots, K$, where Φ is the cumulative normal distribution function. Finally, we generate $(r_{2,i}, Q_{2,i})$ using the following relationships:

$$\begin{aligned}
r_{2,i} &= \phi_0 + \phi_1 (r_1 + \epsilon_{1,i}), \\
Q_{2,i} &= C + A * H_{1,i} + B * Q_1,\ i = 1, \ldots, K.
\end{aligned}$$

To generate the minimum required expected asset returns $\bar{r}_1^{min}$ and $\bar{r}_{2,i}^{min}$ we assume that the investor wants to beat a fixed-weight strategy $\bar{w}$. We randomly generate the vector $\bar{w}$ and then set $\bar{r}_1^{min} = r_1\bar{w}$ and $\bar{r}_{2,i}^{min} = r_{2,i}\bar{w},\ i = 1, \ldots, K$. In this paper, we report results on instances of this model for $n = 20$, 30, and 40 and for different values of $\tau_{2,i}$. The computational results in section 7 indicate that smaller values of $\tau_{2,i}$ tend to generate second stage problems with which warm-start is harder.

3.2. Generation of the random test problems. While two-stage conic extensions of Markowitz model provide an important application example, a more systematic study on the algorithmic behavior is possible on random test problems. We generate these test problems by taking discrete approximations of the following continuous stochastic program:

$$\begin{aligned} \max\ \eta(x) := b^T x + \rho(x) \\ \text{s.t.}\quad Dx = d, \\ Ax + s_1 = c, \\ \|x\|_2 \le \tau_1 \\ s_1 \in \mathcal{K}_1, \end{aligned} \tag{3.3}$$

where

$$\rho(x) := E(\rho(x, \tilde{\omega})), \tag{3.4}$$

and for each realization $\omega \in \Omega$ of the random variable $\tilde{\omega}$

$$\begin{aligned} \rho(x, \omega) := \max\ & f_\omega^T y_\omega \\ \text{s.t.}\ & R_\omega y_\omega + U_\omega x = z_\omega, \\ & W_\omega y_\omega + T_\omega x + s_{2,\omega} = h_\omega, \\ & \|y_\omega\|_2 \le \tau_2 \\ & s_{2,\omega} \in \mathcal{K}_{2,\omega}. \end{aligned} \tag{3.5}$$

In our test problems the first and second stage cones $\mathcal{K}_1$ and $\mathcal{K}_{2,i}$ both consist of a linear, a second order, and two semidefinite cones. The 2-norm inequality constraint in (3.3)–(3.5) gives an additional second order cone. The data D, A, b for the first stage problems uses a random variable with uniform distribution to generate entries in $[-L, L]$. We ensure that D has full row rank and that A has full column rank. We then set $d = D\tilde{x}$ and $c = A\tilde{x} + \iota_1$, where $\tilde{x}$ is generated using a random variable with uniform distribution over $[0, L]^t$, and ι_1 is the *identity element* of the cone $\mathcal{K}_1$. This ensures that the first stage has a feasible interior solution. The problem data $W(\omega)$ and $T(\omega)$ consist of 4 blocks corresponding to the 4 primitive blocks that make up $\mathcal{K}_{2,\omega}$. We assume that t entries in $f(\omega), R(\omega), U(\omega)$ and in each block of the second stage problem data $W(\omega)$ and $T(\omega)$ are random. We start by randomly generating the vector f^0 and the matrices R^0, U^0, W^0, and T^0 whose entries are generated from a uniform distribution over $[-L, L]$. The data f^0, R^0, U^0, W^0, and T^0 are used as a base, which is further randomized to get the second stage scenarios. This is done as follows.

In order to generate second stage scenarios we discretize the continuous probability distribution of $\tilde{\omega}$ using a quasi–Monte Carlo technique. $\tilde{\omega}$ is taken to be a uniform

t-dimensional random vector in $[-L, L]$. In particular, we use a t-dimensional Sobol' sequence to discretize $\tilde{\omega}$ and generate scenarios as follows. Let $(S_1, \ldots, S_N)$ be the Sobol' sequence approximating the t-dimensional uniform distribution. To construct N second stage problems we generate N samples, $\omega_1, \ldots, \omega_N$, of the random vector $\tilde{\omega}$ from this Sobol' sequence using N points. We set $\omega_i = p_{\%} * L * (1 - 2S_i),\ i = 1, \ldots, N$. To obtain the second stage scenario data f_i, R_i, U_i, W_i, and T_i, t elements are randomized in every block (each block corresponding to a cone) of f^0, R^0, U^0, W^0, and T^0 by adding entries of ω_i. Here $p_{\%}$ is the randomization (measured as %) of the base data. A larger value of $p_{\%}$ gives a greater randomization of the base data. Next, we generate a random vector $\tilde{y}$ of the same size as the second stage variable y_ω and set $z_i = R_i\tilde{y} + U_i\tilde{x}$ and $h_i = W_i\tilde{y} + T_i\tilde{x} + \iota_{2,i}$. We choose an appropriate value of τ_1 and $\tau_{2,i}$ to ensure that the feasible sets are bounded. This ensures the feasibility of TSSCP. In (3.3) and (3.5) 2-norm constraints on x and y_ω are given separately for clarity; however, in the rest of the paper they are considered as a second order cone block in A, W_i, T_i and their slacks will be included as a second order cone block in s_1 and $s_{2,i}$, respectively.

Table 2 gives problems generated using the above-described method. For all problems, $L = 2.5$ is used. This table gives information on the number of equality constraints (Equality(F,S)), the size of linear cones (LINC(F,S)), the size of second order cones in the first and second stage problems (SOC(F,S)), the size of semidefinite cones (SDC), and the dimension of variables x and y_i in each of the problems. The suffix (F,S) indicates information in a column for first and second stage problems. The cone information is given for each cone. For example, $[16, 25][16, 25]$ in the Problem MCR2 row indicates two semidefinite constraints in each of first and second stage problems. The size of the first stage semidefinite cone is 4×4 and the size of the second stage semidefinite cone is 5×5. Problems MCR1–MCR6 are generated by taking a reasonable amount of randomness in the second stage ($p_{\%} = 5\%$), Problems MCL1–MCL6 are generated by taking a large amount of randomness ($p_{\%} = 50\%$), and Problems MCH1–MCH6 are generated by taking a huge amount of randomness ($p_{\%} = 500\%$). For all of these problems the size of first and second stage cones are equal and the dimension of the Sobol' sequence is two ($t = 2$). In each second stage problems, two elements f_i, R_i, U_i and two elements in each block of W_i and T_i (totaling 22 elements) are perturbed using the two elements of ω_i. The 2-norm constraint is not randomized.

Problems MCL13–MCL63 have dimensions similar to those of problems MCL1–MCL6; however, these problems were generated using a 33-dimensional Sobol' sequence ($t = 33$) and setting $p_{\%} = 50\%$. Three-dimensional subvectors of the 33-dimensional Sobol' points (altogether 11 distinct subvectors) are used to perturb three elements of f_i, R_i, U_i and three elements in each block of W_i and T_i. Thus these problems are expected to be random in a different way than MCL1–MCL6.

Problems MCRS1–MCRS6 have increasing size of the second stage cone and decision variables y_i. Here Problem MCRS1 is identical to Problem MCR2. Problems MCRF1–MCRF6 have increasing size of the first stage cone and decision variable x. Here Problem MCRF1 is identical to Problem MCR2. These problems were generated using a 2-dimensional Sobol' sequence ($t = 2$) and setting $p_{\%} = 5\%$. When referring to the randomly generated problems we use `name`–N, where N indicates the number of second stage scenarios in the instance.

4. Implementing the primal interior decomposition algorithm. In this section we discuss a practical implementation of primal interior decomposition algo-

TABLE 2
Description of the random test problems.

Problem	Equality(F, S)	LINC(F, S)	SOC(F, S)	SDC(F, S)	Dim(x, y_i)
MCR1	3,3	2,2	[11, 5][11, 5]	[16, 16][16, 16]	10,10
MCR2	3,3	3,3	[21, 10][21, 10]	[16, 25][16, 25]	20,20
MCR3	3,3	5,5	[31, 15][31, 15]	[16, 36][16, 36]	30,30
MCR4	3,3	5,5	[51, 25][51, 25]	[36, 64][36, 64]	50,50
MCR5	3,3	10,10	[101, 50][101, 50]	[100, 81][100, 81]	100,100
MCR6	3,3	15,15	[151, 75][151, 75]	[100, 169][100, 169]	150,150
MCL1	3,3	2,2	[11, 5][11, 5]	[16, 16][16, 16]	10,10
MCL2	3,3	3,3	[21, 10][21, 10]	[16, 25][16, 25]	20,20
MCL3	3,3	5,5	[31, 15][31, 15]	[16, 36][16, 36]	30,30
MCL4	3,3	5,5	[51, 25][51, 25]	[36, 64][36, 64]	50,50
MCL5	3,3	10,10	[101, 50][101, 50]	[100, 81][100, 81]	100,100
MCL6	3,3	15,15	[151, 75][151, 75]	[100, 169][100, 169]	150,150
MCH1	3,3	2,2	[11, 5][11, 5]	[16, 16][16, 16]	10,10
MCH2	3,3	3,3	[21, 10][21, 10]	[16, 25][16, 25]	20,20
MCH3	3,3	5,5	[31, 15][31, 15]	[16, 36][16, 36]	30,30
MCH4	3,3	5,5	[51, 25][51, 25]	[36, 64][36, 64]	50,50
MCH5	3,3	10,10	[101, 50][101, 50]	[100, 81][100, 81]	100,100
MCH6	3,3	15,15	[151, 75][151, 75]	[100, 169][100, 169]	150,150
MCL13	3,3	2,2	[11, 5][11, 5]	[16, 16][16, 16]	10,10
MCL23	3,3	3,3	[21, 10][21, 10]	[16, 25][16, 25]	20,20
MCL33	3,3	5,5	[31, 15][31, 15]	[16, 36][16, 36]	30,30
MCL43	3,3	5,5	[51, 25][51, 25]	[36, 64][36, 64]	50,50
MCL53	3,3	10,10	[101, 50][101, 50]	[100, 81][100, 81]	100,100
MCL63	3,3	15,15	[151, 75][151, 75]	[100, 169][100, 169]	150,150
MCRS1	3,3	3,3	[21, 10][21, 10]	[16, 25][16, 25]	20,20
MCRS2	3,3	3,5	[21, 10][31, 15]	[16, 25][16, 36]	20,30
MCRS3	3,3	3,5	[21, 10][51, 25]	[16, 25][16, 36]	20,30
MCRS4	3,3	3,10	[21, 10][101, 50]	[16, 25][36, 64]	20,50
MCRS5	3,3	3,15	[21, 10][151, 75]	[16, 25][100, 81]	20,100
MCRS6	3,3	3,25	[21, 10][251, 200]	[16, 25][100, 169]	20,150
MCRF1	3,3	3,3	[21, 10][21, 10]	[16, 25][16, 25]	20,20
MCRF2	3,3	5,3	[31, 15][21, 10]	[16, 36][16, 25]	30,20
MCRF3	3,3	5,3	[51, 25][21, 10]	[16, 36][16, 25]	30,20
MCRF4	3,3	10,3	[101, 50][21, 10]	[36, 64][16, 25]	50,20
MCRF5	3,3	15,3	[151, 75][21, 10]	[100, 81][16, 25]	100,20
MCRF6	3,3	25,3	[251, 200][21, 10]	[100, 169][16, 25]	150,20

rithms. This implementation uses the knowledge that the underlying problem is a stochastic program. The central idea is to estimate the properties of a large-scale stochastic problem with the help of problems with few scenarios. This information is used to devise implementation heuristics. In the rest of this paper we assume that the probabilities π_i (weights) in the TSSCP (1.1)–(1.3) are $1/K$.

4.1. Initialization. We need an interior starting point that is feasible for the first stage, and for which all second stage problems have feasible interior solutions. For this purpose we consider the feasibility barrier centering problem

$$\max_{(x,s_1)\in\mathcal{F}^0, s_1\in\mathcal{K}_1^+} \ln\det(s_1)$$

and its KKT conditions

$$\begin{aligned} & Dx = d, \\ & Ax + s_1 = c, \\ & D^T\tau + A^T\upsilon = 0, \\ & P(s_1^{1/2})\upsilon = \iota_1, \\ & s_1, \upsilon \in \mathcal{K}_1^+. \end{aligned} \tag{4.1}$$

Equations (4.1) are solved to a desirable accuracy. An infeasible primal-dual interior point method based on the description in [27, 28] is used. If such a solution is not found within a desirable accuracy, then the problem is "infeasible." Otherwise, the iterate at which the primal-dual algorithm terminates is taken as a solution. Some details of this implementation are given in section 4.2. The properties of interior point methods ensure that if an interior feasible solution is available, then the infeasible primal-dual interior method will provide such a solution [23, 10]. The solution of (4.1) may not provide a starting point for which all second stage problems are feasible if the problem does not have full recourse. We use an artificial variable (y_i^a) for the second stage problem to ensure a full recourse. The second stage problems with the artificial variable take the form

$$\begin{aligned} \max\ & f_i^T y_i - \frac{1}{2} y_i^T H_i y_i - M y_i^a \\ \text{s.t. } & R_i y_i + U_i x = z_i, \\ & W_i y_i + T_i x + s_{2,i} - y_i^a \iota_{2,i} = h_i, \\ & y_i^a \geq 0, s_{2,i} \in \mathcal{K}_{2,i}, \end{aligned} \tag{4.2}$$

where M is a sufficiently large constant to ensure that $y_i^a \to 0$, as a solution of TSSCP is approached. It is easy to construct a feasible solution of (4.2), which is taken as a starting point for the centering problem (2.11) defined for (4.2).

We still have the problem of identifying a proper value of M. We achieve this in the preprocessing phase. We experiment with an approximation of the original large-scale problem with a small number of scenarios (e.g., $N = 100$) with different values of M, which is taken as a multiple of the infinite norm of the problem data. The smallest value of M that successfully reduces y_i^a below 10^{-10} for all sampled problems is multiplied by a constant for later use. This constant is 10 for the experiments reported in this paper. This strategy allows us to work with values of M that are not unnecessarily large. Taking very large values of M may cause numerical difficulties. Although this approach is heuristic it has worked successfully in all our computations. One may think about adjusting the value of M dynamically over the course of the algorithm.

4.1.1. Initial warm-start. When solving different small sampled problems in the beginning, the first stage solution from the first sampled problem is used to warm-start all subsequent problems.

4.1.2. Selection of the starting barrier parameter. Starting with a μ that is too small may take too many inner iterations to reach the initial centered point, whereas choosing an unnecessarily large starting μ increases the number of outer iterations. We use information from the preprocessing phase to identify a suitable

starting μ. When solving a sampled problem we identify the value of μ that gives only a single digit of accuracy in this problem. Solving this problem to optimality also gives us the magnitude of the optimal objective. Let us denote the objective identified in this step by $\hat{o}$. We let $\hat{\mu}^1 = \hat{o}/\hat{\nu}$, where $\hat{\nu}$ is the order of the first stage cone $\mathcal{K}_1$. This value of $\hat{\mu}^1$ is expected to correspond to a solution with zero to one digit of accuracy in the objective value, which can be verified by building a confidence interval around the barrier function value. Next we try different values of μ $(0.1\hat{\mu}^1, 10\hat{\mu}^1)$ as the starting μ, and record the number of inner iterations required to solve the problem. We select a value of μ that requires the least number of inner iterations to get an optimal solution starting from that value. The corresponding solution is used to start the main algorithm for TSSCP.

4.2. Solution of the second stage centering problems. Each inner iteration of our algorithm solves second stage centering problems defined after updating x and/or μ. It is neither necessary nor possible to exactly solve second stage centering problems. In the following, (4.3) ensures sufficient feasibility in the solution, and (4.4) ensures proximity (controlled by the parameter ν) to the central path:

$$\|R_i y_i + U_i x - z_i\| \leq \texttt{tol}_{feas},\ \|W_i y_i + T_i x + s_{2,i} - h_i\| \leq \texttt{tol}_{feas},$$

$$\|W_i^T \lambda_i + R_i^T \gamma_i + H_i y_i - f_i\| \leq \texttt{tol}_{feas}, \tag{4.3}$$

$$\|P(\lambda_i^{1/2}) s_{2,i} - \mu\iota\| \leq \nu\mu. \tag{4.4}$$

We start from the solution of the second stage problem available at the end of the previous iteration. These starting points are no longer feasible or centered. We employ an infeasible primal-dual interior point method to solve the second stage centering problems. The primal-dual method implements the Nesterov–Todd [22, 29] scaling to symmetrize the KKT system and take steps along the Newton direction for recentering. A second stage line search is performed for computing the step-length as follows.

If a full Newton step (length 1) is feasible, we give priority to restoring primal and dual feasibility and take this step. When a full Newton step is not feasible, we choose a step-length that minimizes $\|\xi\| := \|\xi_{feas}, \xi_{cent}\|$, where $(\xi_{feas} = R_i y_i + U_i x - z_i, W_i y_i + T_i x + s_{2,i} - h_i, W_i^T \lambda_i + R_i^T \gamma_i + H_i y_i - f_i)$ is the vector of primal and dual infeasibilities, and $\xi_{cent} = (P(\lambda_i^{1/2}) s_{2,i} - \mu\iota)$ is the residual vector in the complementarity condition (4.4). The derivative of $\|\xi\|^2$ is a third order polynomial in the step-length, which has one or three real roots. Let ϱ denote the maximum step-length that we can take without violating the conic constraints on the primal slack $s_{2,i}$ and on the dual multipliers (λ_i) associated with the primal constraints. If there are no real roots less than ϱ, ϱ is the optimal step-length. If there is only one real root less than ϱ, it is the optimal step-length. If all three roots are real and less than ϱ, either the smallest or the largest root is the optimal step-length. When ϱ is the optimal step-length we take a slightly shorter step $(0.9^*\varrho)$ to ensure that primal and dual iterates stay far away from the boundary. In the results reported in this paper no attempts are made to take different steps along primal and dual directions, and a predictor-corrector strategy is not used.

Table 3 gives computational results for Problem MCR3 with 100 second stage scenarios using $\zeta = 0.1$ and different combinations of β and ν. The first number in each cell of this table gives the number of outer iterations required in our implementation. The second number gives the average number of second stage Newton step calculations

TABLE 3

Number of inner iterations and average iterations for recentering second stage problems for Problem MCR3-100 using accurate (3 digits) line search.

$\zeta = 0.1$	ν			
β	0.01	0.1	1	10
$(2-\sqrt{(3)})/2$	31 1.09	31 1.01	30 1.00	31 0.99
0.5	24 1.11	26 1.01	24 1.00	25 1.00
1	21 1.10	21 1.01	22 1.00	22 1.00
3	20 1.10	19 1.01	20 1.00	19 1.00
5	- -	- -	- -	- -

(total number of second stage Newton steps over all the scenarios/total number of calls to the second stage problems) for each call to the second stage problem. We use $\texttt{tol}_{feas} = 10^{-9}$ in condition (4.3). Note that the average number of second stage Newton iterations is close to one for the values of ν in the range 0.01 to 10, and the number of outer iterations is also insensitive to the accuracy of the second stage problem solutions. The values of β used for results in this table are chosen to ensure closeness to the first stage central path. Larger values of β do not produce stable performance, as seen from the results reported in Tables 3 and 4. For these results the algorithm is terminated when the relative improvement in the objective value at two successive major iterations is less than eight digits. It usually corresponds to a value of μ between 10^{-7} to 10^{-8}. The number of inner iterations taken by the algorithm is given in Table 3 if a problem is successfully solved to eight digits of accuracy. For all runs reported in Table 3 we used a relatively accurate line search to avoid any misinterpretations. In all the runs performing a line search, the search is terminated when we have three digits of accuracy in the step-length parameter. The iteration counts are indicated by "–" when the implementation fails to achieve eight digits of precision. There are several reasons for this failure. For some problems the Cholesky factorization method in MATLAB becomes unstable. For other problems the solution does not have eight digits of accuracy, even though the change in the objective between two successive major iterations is less than eight digits. This may be due to the fact that for $\beta = 5$ the neighborhood of the central path may be too wide to correctly measure termination based on either relative improvement or the barrier parameter value criterion.

When warm-starting from the previous solution, typically only one (occasionally two) second stage iteration is required. Furthermore, the number of inner iterations of the primal decomposition algorithm remain unchanged for ν in a very large range. Computationally we observe that typically the termination conditions (4.4) are satisfied for much smaller values of $\texttt{tol}_{feas}$ since Newton steps are very effective from a warm-start solution. Furthermore, from Table 3 we observe that the average number of iterations required to recenter the second stage grows only slightly for smaller values of ν.

4.3. First stage centering and barrier reduction rate. Tables 4, 7, and 8 give computational results for Problem MCR3-100 for $\nu = 1$, and using different combinations of β and ζ with exact and inexact line search. The results in Table 4 show that the total number of inner iterations increase for large values of ζ (μ is

TABLE 4
Number of inner iterations and average inner iterations per outer iteration for Problem MCR3-100 using accurate (3 digits) line search.

$\nu = 1$	ζ			
β	0.05	0.1	0.25	0.5
$(2-\sqrt{(3)})/2$	29 4.33	30 3.86	36 2.83	56 2.35
0.5	26 3.83	24 3.14	31 2.42	50 2.09
1	24 3.50	22 2.71	27 2.08	49 2.04
3	- -	20 2.43	26 2.00	26 1.04
5	- -	- -	25 1.92	25 1.00

decreased slowly). Also when μ is reduced too aggressively the total number of inner iterations tends to increase, indicating that recentering becomes more difficult in this case. In our experiments the choices $\zeta = 0.1$ and $\beta = 1$ give good results for all the problems.

4.4. First stage step-length. The analysis of long-step primal interior point algorithms, such as the one given in [18, 21, 31] is based on taking fixed-length steps along the Newton direction. The choice of this step-length is given as $\theta = (1+\delta)^{-1}$, where δ is the local norm of the Newton direction, as explained in sections 2.1 and 2.2. This step-length is usually very conservative, and there is a need to develop a step-length selection strategy that is more efficient. The strategy of taking a constant step to the boundary, which works well for the primal-dual algorithms [13, 15], is not always stable in the current setting. Also, in practice we cannot perform an elaborate line search with backtracking, since each barrier function evaluation (or its derivative evaluation) requires solution of all second stage scenarios. This is almost equal to the cost of computing the first stage direction afresh. Hence, we need a heuristic that avoids unnecessary barrier function evaluations while not increasing the total number of inner iterations significantly.

The following step-length selection strategy has worked well in our computational experiments: $\theta = \min\{\alpha(1+\delta)^{-1}, 1, 0.9\varrho\}$, where ϱ is the maximum step to the boundary of the feasible set, and α is a scalar constant. We take this step and check the value of barrier function at the new point. This evaluation of the barrier function is combined with the computation of the new Newton direction if the step is accepted. Hence, there is no efficiency loss in this function evaluation. We ensure that the barrier function has reduced sufficiently at the new point, which is almost always the case. In a few rare occasions, when the barrier function is not reduced sufficiently, we can backtrack by aggressively reducing α. In our computations when this happened we simply set $\alpha = 1$; i.e., in this case we take $\theta = (1+\delta)^{-1}$. This has always reduced the barrier function to a desirable amount. The use of the parameter α is justified through empirical observations, which indicate that when an iterate is close to the boundary, the optimum step-length is a multiple of $(1+\delta)^{-1}$. Such results are shown in Table 5 for problem MCR4-100. We use $\alpha = 4$ in our actual computations. The results in Table 6 show that the choice of parameter α is important in reducing the total number of inner iterations. Note that the number of inner iterations are slightly larger for both small and large values of α. This is because for a small value of α the step is conservative, while for a large value of α the step is

TABLE 5

Behavior of feasible and optimal first stage step-lengths for problem MCR4-100 ($\beta = 1$, $\nu = 1$, $\zeta = 0.1$).

Iter. #	μ^k	$(1+\delta)^{-1}$	Max feasible step-length	Line search step-length	Line search step-length / $(1+\delta)^{-1}$
1	1.00E+00	0.02	0.07	0.07	3.50
2	1.00E+00	0.21	1.65	1.46	6.95
3	1.00E+00	0.28	1.95	1.48	5.29
4	1.00E+00	0.5	4.14	1.38	2.76
5	1.00E-01	0.08	0.22	0.22	2.75
6	1.00E-01	0.21	0.77	0.6	2.86
7	1.00E-01	0.41	2.3	1.35	3.29
8	1.00E-01	0.56	10.59	2.08	3.71
9	1.00E-02	0.04	0.1	0.09	2.25
10	1.00E-02	0.13	0.37	0.35	2.69
11	1.00E-02	0.38	2.34	1	2.63
12	1.00E-02	0.64	17.21	1.4	2.19
13	1.00E-03	0.03	0.09	0.09	3.00
14	1.00E-03	0.2	0.43	0.38	1.90
15	1.00E-03	0.47	2.54	0.9	1.91
16	1.00E-03	0.85	146.46	1.1	1.29
17	1.00E-04	0.03	0.1	0.1	3.33
18	1.00E-04	0.42	1.42	0.73	1.74
19	1.00E-04	0.82	20.69	1.04	1.27
20	1.00E-05	0.03	0.1	0.1	3.33
21	1.00E-05	0.62	3.38	0.86	1.39
22	1.00E-06	0.03	0.1	0.1	3.33
23	1.00E-06	0.61	2.55	0.81	1.33
24	1.00E-07	0.03	0.1	0.1	3.33
25	1.00E-07	0.57	1.73	0.73	1.28
26	1.00E-08	0.03	0.1	0.1	3.33
27	1.00E-08	0.51	1.13	0.6	1.18

usually so large that we take 0.9ϱ as a step-length instead of $\alpha(1+\delta)^{-1}$. Reasons for failure indicated in these tables are similar to those discussed before. While analyzing 100 scenario problems of different size and randomness we find that the heuristic step-length strategy performs well, taking within 10–20% of the number of inner iterations required by an implementation using "exact" line search. The "exact" line search is performed using bisection method, and it is terminated when the interval of uncertainty is reduced to 10^{-3}. The worst offenders (about 25% worse) were problems MCRF14–MCRF6 (see Table 10) suggesting that further refinements may provide improvements.

5. Basic behavior of the primal decomposition algorithm.

5.1. Comparison with a direct method. Since we can formulate TSSCP as a large conic programming problem (extensive formulation), one can use a primal-dual interior point method for solving the extensive formulation. SeDuMi [26] is a popular solver that can be used to directly solve the extensive formulations of the problem we generate. Table 9 gives a comparison of the total number of iterations taken by the primal decomposition algorithm (for $\alpha = 4, \beta = 1, \zeta = 0.1, \nu = 1$) and those taken by SeDuMi 1.05 [26] for an extensive formulation of the problem. In order to compare the performance of primal decomposition algorithm with that of SeDuMi, it is more appropriate to consider the number of inner iterations taken by the primal decomposition algorithm with the number of iterations taken by SeDuMi to achieve the same accuracy in the solution. SeDuMi implements Mehrotra's predictor-corrector method.

TABLE 6

Heuristic step-length selection: Number of inner iterations for problems MCR1–MCR6 for varying values of β, ($\zeta = 0.1$, $\nu = 1$).

		β				
Problem	α	$\frac{2-\sqrt{3}}{2}$	0.5	1	3	5
MCR1-100	2	37	36	35	25	-
	4	35	28	31	25	-
	∞	35	32	27	31	-
	line search	-	26	22	20	-
MCR2-100	2	34	31	31	-	-
	4	34	31	30	-	-
	∞	35	32	32	-	-
	line search	-	28	27	23	22
MCR3-100	2	-	31	25	-	-
	4	35	31	25	-	-
	∞	35	31	25	-	-
	line search	30	24	22	20	-
MCR4-100	2	48	40	39	29	-
	4	39	37	32	-	-
	∞	40	38	32	-	-
	line search	-	32	29	-	-
MCR5-100	2	62	56	54	46	-
	4	52	43	40	-	-
	∞	54	45	43	-	-
	line search	47	42	38	-	-
MCR6-100	2	71	55	51	43	-
	4	55	46	43	-	-
	∞	57	48	45	-	-
	line search	49	41	35	-	-

TABLE 7

Heuristic step-length selection: Number of inner iterations and average inner iterations per outer iteration for Problem MCR3-100.

$\nu = 1,\ \alpha = 4$	ζ			
β	0.05	0.1	0.25	0.5
$(2-\sqrt{3})/2$	35	33	58	99
	5.83	4.43	4.83	4.13
0.5	30	28	46	73
	5.00	3.50	3.83	3.04
1	21	24	34	51
	3.50	3.14	2.83	2.13
3	-	-	24	26
	-	-	2.00	1.08
5	-	-	-	-
	-	-	-	-

The work required to compute one predictor step in SeDuMi is roughly equivalent to computing a Newton direction in an inner iteration of the primal decomposition algorithm. A comparison of CPU times would be misleading for a variety of reasons. These codes are written using different linear algebra libraries (SeDuMi uses its C code, while we depend on MATLAB). Because MATLAB is an interpreted language, it must interpret every line in a for-loop every time it goes through it, and this makes MATLAB very slow in handling loops. Furthermore, when solving problems we generate scenarios on the fly, while SeDuMi reads an extensive formulation, generating and writing to a text file that is readable by SeDuMi which is memory intensive. We were unable to feed larger problems to SeDuMi for this reason. For the problems

TABLE 8
*Heuristic step-length selection: Number of inner iterations and average number of second stage Newton iterations per call for Problem MCR*3-100.

$\zeta = 0.1,\ \alpha = 4$	ν			
β	0.01	0.1	1	10
$(2-\sqrt{3})/2$	33 1.35	33 1.12	33 1.03	33 1.03
0.5	28 1.40	28 1.14	28 1.03	28 1.03
1	24 1.48	24 1.17	24 1.04	24 1.04
3	- -	- -	- -	- -
5	- -	- -	- -	- -

TABLE 9
Comparison of our implementation ($\zeta = 0.1$, $\beta = 1$, $\nu = 1$, and $\alpha = 4$) to SeDuMi.

	Our implementation		SeDuMi		
Problem	Objective	Itr.	Primal obj.	Dual obj.	Itr.
MCR1-100	1.631426527	31	1.631426587	1.631426582	22
MCR2-100	3.384065478	30	3.384065491	3.384065487	24
MCR3-100	204.2468748	25	204.2468761	204.2468760	22
MCR4-100	21.08956769	32	21.08956771	21.08956770	25

TABLE 10
Number of inner iterations with increasing problem size ($\zeta = 0.1$, $\beta = 1$, $\nu = 1$, and $\alpha = 4$).

Problem	Heur.	Line search	Problem	Heur.	Line search
MCRS1-100	30	27	MCRF1-100	30	27
MCRS2-100	27	25	MCRF2-100	30	28
MCRS3-100	23	22	MCRF3-100	31	30
MCRS4-100	22	21	MCRF4-100	46	37
MCRS5-100	24	21	MCRF5-100	51	40
MCRS6-100	32	27	MCRF6-100	46	36

that were solved by both SeDuMi and the primal decomposition algorithm, results indicate that both algorithms correctly achieve eight digits of accuracy. SeDuMi took 10–30% fewer iterations. Recall that the corrector step in the predictor-corrector method reduces the total number of iterations by 20–50%. Hence, it appears that for a "Newton" step–based algorithm the total number of iterations taken by the primal decomposition algorithm is comparable.

5.2. Performance with increasing problem size. Table 10 gives the number of iterations required by the primal decomposition method with increasing size of cones in the problem (called *problem size* for short). The problems with increasing second stage size are MCRS1–MCRS6, and those with increasing first stage size are MCRF1–MCRF6. These results show that the number of inner iterations are not significantly affected by the second stage problem size, whereas the number of inner iterations show an early upward trend but become stable as the first stage problem size grows. This indicates that the complexity of the algorithm depends on the first stage problem size; however, the iteration growth is not very rapid, which is generally the case with interior point methods.

TABLE 11

Number of inner iterations as problem randomness increases ($\zeta = 0.1$, $\beta = 1$, $\nu = 1$, and $\alpha = 4$). (Results for problems MCR1-100 to MCR6-100 are given in Table 6.)

Problem	Heur.	Line search	Problem	Heur.	Line search	Problem	Heur.	Line search
MCL1-100	27	23	MCH1-100	26	24	MCL13-100	29	24
MCL2-100	36	29	MCH2-100	31	26	MCL23-100	31	26
MCL3-100	24	22	MCH3-100	25	25	MCL33-100	26	24
MCL4-100	32	29	MCH4-100	31	29	MCL43-100	32	30
MCL5-100	40	39	MCH5-100	38	36	MCL53-100	43	38
MCL6-100	43	37	MCH6-100	39	38	MCL63-100	40	37

TABLE 12

Number of inner iterations as number of scenarios increases ($\zeta = 0.1$, $\beta = 1$, $\nu = 1$, and $\alpha = 4$).

	# scenarios							
	10		100		1,000		10,000	
Problem	Heur.	Line search	Heur.	Line search	Heur.	Line search	Heur.	Line search
MCR2	31	27	30	27	32	26	32	26
MCR3	25	21	25	22	25	22	25	22

5.3. Performance of the algorithm with increasing randomness. We generated problems with increasing amount of randomness in the problem data by varying the value of $p_{\%}$. Recall that for Problems MCR1–MCR6 $p_{\%} = 5\%$, for Problems MCL1–MCL6 $p_{\%} = 50\%$, and for Problems MCH1–MCH6 $p_{\%} = 500\%$. We also generated problems MCL13–MCL63 where the dimension of the underlying random variable is 33 and $p_{\%} = 50\%$. Results for 100-scenario instances of these problems are given in Table 11. When comparing the number of inner iterations for these problems with those for problems MCR1–MCR6 (see Table 6) no significant trend is observed with increasing randomness. From these results we infer that the number of inner iterations required by the algorithm is not dependent on the randomness in the problem data.

5.4. Performance of the algorithm with increasing number of scenarios. In Table 12 we give the number of inner iterations taken when solving Problems MCR3 and MCR4 with an increasing number scenarios. Implementation details for problems with larger than 100 scenarios are discussed in section 7. A comparison of these results shows no increasing trend in the required number of inner iterations with increasing number of scenarios.

6. Adaptive scenario addition heuristic. The primal decomposition algorithm naturally allows adaptive addition of scenarios as the algorithm progresses. It is important to explore this possibility since adaptive addition of scenario may lead to significant computational savings. Recall that when the number of scenarios is fixed (no adaptive scenario addition) we move along a fixed first stage central path as μ is reduced. However, in the adaptive algorithm the first stage central path changes when new scenarios are added. As the results in section 7 show, we need to be careful while adding scenarios adaptively. This is because if a large number of scenarios are added for a small value of the barrier parameter, the change in the primal central path caused by this scenario addition may result in a poorly conditioned problem. This, in turn, may require significantly more inner iterations to recenter and could even cause numerical breakdowns.

In this section we develop a heuristic that addresses this issue. The key idea behind our heuristic is to add the number of scenarios in such a way that adding scenarios and reducing the value of barrier parameter in the algorithm have a similar impact on displacement of the first stage center. For this purpose we devise methods for estimating the change in the barrier objective value caused by scenario addition, and also suggest ways to measure the change in the barrier objective resulting from reducing μ. The assumptions made while developing this heuristic are justified through empirical evidence.

We add new scenarios when μ is reduced in Step 2 of Algorithm 1. Our heuristic aims to maintain

$$|\eta_{N_k}(\mu^k, x_{N_k}(\mu^k)) - \eta_{N_{k-1}}(\mu^k, x_{N_{k-1}}(\mu^k))| \approx \Delta\eta_{N_k}(\mu^k),$$

where $\eta_{N_k}(\mu^k, x_{N_k}(\mu^k))$ is the optimum objective of the sample-average log-barrier problem with N_k samples for $\mu = \mu^k$ and

$$\Delta\eta_{N_k}(\mu^k) := \eta_{N_k}(\mu^k, x_{N_k}(\mu^k)) - \eta_{N_k}(\mu^k, x_{N_k}(\mu^{k-1})). \tag{6.1}$$

Note that the sample-average function is a random variable when scenarios are generated randomly (Monte Carlo). We represent the corresponding random variable by $\hat{\eta}_N(\cdot,\cdot)$. Every randomly generated batch of N-samples yields a different realization $\eta_N(\cdot,\cdot)$ of $\hat{\eta}_N(\cdot,\cdot)$.

A justification of the heuristic idea is as follows. In the adaptive implementation of our algorithm, at the beginning of the kth outer iteration we decrease the barrier parameter from μ^{k-1} to $\mu^k = \zeta\mu^{k-1}$, and we simultaneously increase the number of scenarios from N^{k-1} to N^k. In general, both of these actions may increase the distance of current point x to the central path. Decreasing μ moves away the μ-center, whereas increasing the number of scenarios may change the central path. Since experimentally we know that it takes only a few iterations to recenter after decreasing μ in the case where the number of scenarios is fixed, by keeping the impact of adding to the number of scenarios of the same order as the impact of decreasing μ, we expect that the number of iterations required to converge to the new center with added scenarios will be of the same order.

6.1. Confidence interval estimation. Various estimations required to build our heuristic are motivated by the following theorem from Shapiro [25].

THEOREM 6.1. *Consider the stochastic program* $\inf_{x\in X}\int_\Omega g(x,\omega)P(d\omega)$ *and let* z^* *be its optimal objective. Let* $\hat{z}_N$ *denote the optimal objective of the sample average problem* $\inf_{x\in X} N^{-1}\sum_{i=1}^N g(x,\omega^i)$*, where* ω^i *are independent. Suppose* X *satisfies the following conditions:*

(i) $g(x,\cdot)$ *is measurable for all* $x \in X$*;*

(ii) *there exists some function* $f:\ \Omega \to \mathbb{R}$ *such that* $\int_\Omega |f(\omega)|^2 P(d\omega) < \infty$*, and*

$$|g(x_1,\omega) - g(x_2,\omega)| \le f(\omega)|x_1 - x_2|$$

for all $x_1, x_2 \in X$*;*

(iii) *for some* $\bar{x} \in X$, $\int_\Omega g(\bar{x},\omega)dP(\omega) < \infty$*, and* $E\{g(x,\omega)\} = \int_\Omega g(x,\omega)dP(\omega)$ *has a unique minimizer* $x^* \in X$*. Then* $\sqrt{N}(\hat{z}_N - z^*)$ *converges in distribution to normal distribution* $\Phi(0, \mathrm{Var}[g(x^*,\xi)])$*, where* $\mathrm{Var}[g(x^*,\omega)] = \int_\Omega g(x^*,\omega)^2 dP(\omega) - E\{g(x^*,\omega)\}^2$.

Recall that since the log-barrier recourse function $\eta_N(\mu,x)$ is concave in x, it has a unique maximizer. Thus it satisfies the uniqueness condition in (iii) of Theorem 6.1. Furthermore, since concave functions are Lipschitz continuous in the interior

of their domain, condition (ii) of Theorem 6.1 is also satisfied. The assumption that first and second stage problems are feasible and bounded ensures the remaining regularity conditions in Theorem 6.1. Hence, when the random variable governing the stochastic program has a continuous distribution and the second stage problems are generated using Monte Carlo sampling, Theorem 6.1 applies. In this case, we have $\mathrm{Var}[\hat{\eta}_N(\mu, x_N(\mu))] = \sigma^2(\mu, x(\mu))/N$, where $\sigma^2(\mu, x(\mu)) = \mathrm{Var}[\rho(\mu, x(\mu), \tilde{\omega})]$ [8]. Here, $\rho(\cdot, \cdot, \tilde{\omega})$ is the log-barrier counterpart of $\rho(\cdot, \tilde{\omega})$ in (3.4). Note also that we used the observation that the randomness of $\hat{\eta}_N(\mu, x_N(\mu))$ is due to the log-barrier recourse function. Theorem 6.1 shows that $\sqrt{N}(\hat{\eta}_N(\mu, x_N(\mu)) - \eta(\mu, x(\mu)))$ converges in distribution to a normal random variable with a mean of zero and variance of $\sigma^2(\mu, x(\mu))$.

For any batch of N samples, this result provides a statistical bound on $|\eta_N(\mu, x_N(\mu)) - \eta(\mu, x(\mu))|$ that holds with a certain confidence:

$$|\eta_N(\mu, x_N(\mu)) - \eta(\mu, x(\mu))| \mathrel{\widetilde{\leq}} z_\alpha \frac{s_N(\mu, x_N(\mu))}{\sqrt{N}}.$$

Here $s_N^2(\mu, x_N(\mu))$ is an N-sample approximation of the sample variance estimator of $\sigma^2(\mu, x(\mu))$, and the value of z_α provides the level of confidence we want in the bound. In particular, z_α satisfies $P\{\Phi(0,1) \leq z_\alpha\} = 1 - \alpha$. The statistical and approximate nature of this bound is indicated by using the symbol $\widetilde{\leq}$.

We use a quasi–Monte Carlo (QMC) (Sobol' sequence) technique instead of Monte Carlo to generate second stage samples since it is known that better variance convergence rates are possible by using QMC techniques. For example, the empirical results in Glasserman [8] and our own experiments suggest that a Sobol' sequence gives much faster convergence than a Monte Carlo sampling. Unfortunately, the convergence rate theory for these techniques is not yet fully developed in the stochastic programming setting. For example, we do not know the asymptotic distribution of $\hat{\eta}_N(\mu, x_N(\mu)) - \eta(\mu, x(\mu))$ for scenarios generated using a Sobol' sequence, and whether an analogue of Theorem 6.1 holds. Heuristically, we expect that when scenarios are generated using a Sobol' sequence,

$$|\eta_N(\mu, x_N(\mu)) - \eta(\mu, x_N(\mu))| \mathrel{\widetilde{\leq}} z_\alpha RMS_N(\mu, x_N(\mu)),$$

where $RMS_N(\mu, x_N(\mu))$ represents the "root-mean-square" error of $\hat{\eta}_N(\mu, x_N(\mu))$, and z_α provides a confidence interval as before. The root-mean-square error is calculated by randomizing the Sobol' sequence as follows.

Let S_i represent the Sobol' sequence used to generate the second stage scenarios. To generate the jth set of N scenarios we generate a random (Monte Carlo) vector U_j, uniformly distributed over the unit hypercube of appropriate dimension, and generate a new sequence: $\tilde{S_i}^j = (S_i + U_j) \bmod 1,\ i = 1, \ldots, N$. The problem corresponding to $\tilde{S_i}^j$ is generated as described in section 3.2. This process is repeated for $j = 1, \ldots, m$ to generate m random problems each having N scenarios. Let $\eta_{N(j)}(\mu, x_{N(j)}(\mu))$ denote the log-barrier objective of the jth problem for a given μ and $x_{N(j)}(\mu)$. Now

$$RMS_N(\mu, x_N(\mu)) := \sqrt{\frac{1}{m-1}\sum_{j=1}^{m}(\eta_{N(j)}(\mu, x_{N(j)}(\mu)) - \bar{\eta}_N)^2}, \quad \text{where} \tag{6.2}$$

$$\bar{\eta}_N := \frac{1}{m}\sum_{j=1}^{m}\eta_{N(j)}(\mu, x_{N(j)}(\mu)).$$

Calculation of $RMS_N(\mu, x_N(\mu))$ for large N requires significant work, which is not practical in an algorithmic setting. However, in QMC we expect RMS to decrease by $O(1/N^\kappa)$, where $1 \geq \kappa > 0.5$ [8, Chapter 5]. Note that for Monte Carlo $\kappa = 0.5$. The knowledge of κ can allow us to approximate $RMS_N(\mu, x_N(\mu))$ for large N by scaling $RMS_{N'}(\mu, x_{N'}(\mu))$ calculated for some small value N'. However, even this computation may be expensive, since computation of $RMS_{N'}(\mu, x)$ requires several (m) replications of problems with small sample size. With this in mind in our implementation we perform the computation of κ once off-line, while we further approximate the computation of $RMS_N(\mu, x_N(\mu))$ by taking the (possibly biased) estimation

$$RMS_N(\mu, x_N(\mu)) \approx \hat{\sigma}_N(\mu, x_N(\mu))/N^\kappa,$$

where

(6.3)

$$\hat{\sigma}_N^2(\mu, x_N(\mu)) := \frac{1}{N-1}\sum_{i=1}^{N}(\rho_i(\mu, x_N(\mu)) - \bar{\rho}_N)^2, \text{ where } \bar{\rho}_N := \frac{1}{N}\sum_{i=1}^{N}\rho_i(\mu, x_N(\mu)).$$

Note that as a consequence of this approximation we now have

$$|\eta_N(\mu, x_N(\mu)) - \eta(\mu, x(\mu))| \lesssim z_\alpha \hat{\sigma}_N(\mu, x_N(\mu))/N^\kappa,$$

which is similar to the approximation for the Monte Carlo case except that we now have a tighter inequality due to $1/N^\kappa$. With the above discussion the problem of estimating a bound on $|\hat{\eta}_N(\mu, x_N(\mu)) - \eta(\mu, x(\mu))|$ is now reduced to the problem of estimating $\hat{\sigma}_N^2(\mu, x(\mu))$ and κ. The discussion in section 6.2 shows that good approximations of $\hat{\sigma}_N(\mu, x_N(\mu))$ are available cheaply because of the empirically observed properties of the primal central path. Fortunately very accurate values of κ are not required since the adaptive scenario generation heuristic allows a lot of flexibility. The calculation of κ is discussed in section 6.3.

6.2. Empirical observations on the first stage central path and the variance of second stage barrier objectives. We empirically observe that $\hat{\sigma}_N(\mu, \cdot)$ does not change significantly with μ at solutions on (or near) the first stage central path. Furthermore, it also does not change significantly with increasing value of N. Hence, it is sufficient to compute $\hat{\sigma}_N(\mu, \cdot)$ for a large value of μ and a small sample size N' at $x_{N'}(\mu)$. These empirical observations are illustrated in Table 13. Results in this table are obtained from 100, 1,000, and 10,000 scenario approximations of problem MCR3. We solved each of these problems using the primal decomposition method. We calculated $\hat{\sigma}_N(\mu^k, \hat{x}_N(\mu^k))$ as defined in (6.3) at the approximate first stage μ-center $\hat{x}_N(\mu^k)$ in every outer iteration k of the primal decomposition algorithm.

In the above experiment for each outer iteration k, we also calculated the changes in the value of η_N to observe its behavior. From the results in Table 13 we observe that $\Delta\eta_N(\mu^k)$ scales proportionally to μ and does not change significantly as N increases. This suggests that it is sufficient to solve the sample average approximation to a low accuracy (zero or one digit) to estimate $\hat{\sigma}_{N'}(\mu, \cdot)$, and $\Delta\eta_N(\mu)$ for any μ. Note that this last property is due to the form of the objective in (2.12) (due to the weighting of the second stage barrier term by π_i), and it is not obviously shared by the objective function in (2.3), where the "contribution" of the barrier grows as more scenarios are added.

TABLE 13
Empirical properties of the first stage central path.

N	μ^k	$\hat{\sigma}_N(\mu^k, \hat{x}_N(\mu^k))$	$\Delta\eta_N(\mu^k)$
100	1E+0	4.998	-
	1E-1	5.167	9.17E+00
	1E-2	5.210	1.08E+00
	1E-3	5.200	1.13E-01
	1E-4	5.198	1.14E-02
	1E-5	5.198	1.13E-03
	1E-6	5.198	1.14E-04
	1E-7	5.198	1.42E-05
1,000	1E+0	5.470	-
	1E-1	5.639	9.15E+00
	1E-2	5.666	1.08E+00
	1E-3	5.650	1.13E-01
	1E-4	5.647	1.16E-02
	1E-5	5.646	1.17E-03
	1E-6	5.646	1.14E-04
	1E-7	5.646	1.94E-05
10,000	1E+0	5.319	-
	1E-1	5.505	9.16E+00
	1E-2	5.530	1.07E+00
	1E-3	5.506	1.13E-01
	1E-4	5.496	1.15E-02
	1E-5	5.495	1.14E-03
	1E-6	5.495	1.13E-04
	1E-7	5.495	1.66E-05

6.3. Estimating convergence rate of the root-mean-square error. For problem MCR3 we experimented with $N = 10$, 50, 250, 1,000, and 5,000 scenario approximations of the TSSCP. We generated $m = 100$ batches of 5,000 samples. We identified $\hat{x}^1_{10}(0.001)$, an approximate first stage μ-center of the first 10 scenario problem for $\mu = 0.001$, and then solved all the second stage problems in all batches setting $x = \hat{x}^1_{10}(0.001)$ and $\mu = 0.001$. Let $i(j)$ be the index of the ith sample in the jth set with N samples. For $j = 1, \ldots, 100$ and N= 10, 50, 250, 1,000, and 5,000 we calculated

$$\eta_{N(j)}(0.001, \hat{x}_{10}(0.001)) = b^T x - \frac{1}{2}x^T G x + \mu \ln\det(s_1) + \frac{1}{N}\sum_{i(j)=1}^{N} \rho_{i(j)}(\mu, x)\Bigg|_{\mu=0.001,\ x=\hat{x}^1_{10}(0.001)}.$$

Using these values we obtained the root-mean-square error $RMS_N(0.001, \hat{x}^1_{100}(0.001))$ for $N = 10$, 50, 250, 1,000, and 5,000. Table 14 gives the root-mean-square errors for Problem MCR3 calculated in this way. An average convergence rate of $\kappa = 0.84$ is observed. Note that a different choice of x may give a slightly different rate of convergence, and also one may think about building a confidence interval for the rate of convergence. We have not done this in our experiments, since a crude estimate is sufficient for us. In our numerical experiments on adaptive scenario addition, we used $\kappa = 0.85$ for all randomly generated problems, and $\kappa = 0.7$ for all Markowitz problems. The dimension of the Sobol' sequence used to generate Markowitz problems is greater (20 to 40 vs. 2 or 33 in the randomly generated problems), potentially resulting in a slower convergence rate.

TABLE 14
Number of scenarios vs. RMS.

N	$RMS_N(0.001, \hat{x}^1_{10}(0.001))$
10	2.254958E-01
50	5.581944E-02
250	1.150173E-02
1,000	4.230485E-03
5,000	1.309438E-03

6.4. A heuristic for adaptive scenario addition. Let μ^3 be the value of barrier parameter after three major iterations in the preprocessing phase and let $\hat{x}(\mu^3)$ be a sufficiently well-centered first stage solution for $\mu = \mu^3$. From the discussion in section 6.1 it follows that

$$\begin{aligned}
&|\eta_{N_k}(\mu^k, \hat{x}_{N_k}(\mu^k)) - \eta_{N_{k-1}}(\mu^k, \hat{x}_{N_{k-1}}(\mu^k))| \\
&= |\eta_{N_k}(\mu^k, \hat{x}_{N_k}(\mu^k)) - \eta(\mu^k, x(\mu^k)) + \eta(\mu^k, x(\mu^k)) - \eta_{N_{k-1}}(\mu^k, \hat{x}_{N_{k-1}}(\mu^k))| \\
&\lessapprox z_\alpha \left(\frac{\hat{\sigma}_{N_k}(\mu^k, \hat{x}_{N_k})}{N_k^\kappa} + \frac{\hat{\sigma}_{N_{k-1}}(\mu^k, \hat{x}_{N_{k-1}})}{N_{k-1}^\kappa} \right) \\
&\lessapprox z_\alpha \left(1 + \left(\frac{N^{k-1}}{N^k} \right)^\kappa \right) \frac{\hat{\sigma}_{N'}(\mu^3, \hat{x}(\mu^3))}{N_{k-1}^\kappa} \\
&\lessapprox 2 z_\alpha \frac{\hat{\sigma}_{N'}(\mu^3, \hat{x}(\mu^3))}{N_{k-1}^\kappa}.
\end{aligned}$$

We may now choose N_k such that

$$2 z_\alpha \frac{\hat{\sigma}_{N'}(\mu, \cdot)}{N_{k-1}^\kappa} = \Delta\eta_{N_k}(\mu^k). \tag{6.4}$$

Note that the heuristic calculation in (6.4) gives an estimation of the number of scenarios at iteration $(k-1)$ in anticipation of the impact of reduction in the barrier parameter. Hence, the heuristic anticipates the future and loads sufficiently many scenarios in the current outer iteration so that the central path does not change dramatically when we load new scenarios in the next outer iteration. In our computations we set $N' = 100$ and used $\Delta\eta_{N_k}(\mu^k) \approx \Delta\eta_{100}(\mu^3)\frac{\mu^k}{\mu^3}$.

7. Computational results on adaptive scenario addition and warm-start. After adding new scenarios we need to solve the corresponding second stage centering problem for the current value of x and μ. In order to exploit the solutions of the scenarios that are already in the problem, we may look for a scenario that is close to the scenario to be added. One possibility is to look for a scenario that is close in terms of problem data to the new scenario. A solution from the neighboring scenario may provide a good starting point. The Sobol' sequences that we used to generate our scenarios do not provide this neighborhood information a priori. Therefore, for each Sobol' point S_k we heuristically identified a neighboring point among the points $S_1, \ldots, S_{k-1}$. We then used the final iterate from this neighboring scenario for the current value of x and μ as a starting solution for the new scenario. Let us denote this warm-start strategy as Warm-Start-1.

Our experience with Warm-Start-1 has been mixed. For easy problems, e.g., when the second stage problems have only linear constraints and linear objectives, we

observed that the strategy of using a solution from a nearby scenario works. However, for more difficult conic problems Warm-Start-1 exacerbates the performance. For these problems starting from a well-centered point of a near by scenario, especially at small values of μ, Newton iterations were not able to absorb the infeasibility within a practical number of iterations. In general, our experience is that such warm-starts with conic and semidefinite problems are harder for the same amount of perturbation in the problem data.

We also experimented with an alternative approach (Warm-Start-2), which was more stable in the current setting. In this approach, once we complete the final inner iteration of the first outer iteration, we also solve all the second stage problems that are yet not included in the problem, for the starting $\mu = \mu^1$ and the latest first stage solution $x = x(\mu^1)$. For $\mu = \mu^1$, the warm-start is very efficient and each second stage problem can be solved within a few iterations. We store all these second stage solutions for later use. When a new scenario is added in a subsequent outer iteration k ($k > 1$), we warm-start its solution from the iterate stored for this scenario in the first outer iteration. We proceed in two steps: In the first step, we solve the new problem for $\mu = \mu^1$ and $x = x(\mu^k)$ and in the second step we reduce μ from μ^1 to μ^k.

7.1. Discussion of randomly generated problems. Tables 15–20 present computational results for problems MCR3-10,000, MCL3-10,000, MCH3-10,000, and MCL33-10,000. Run 1 in these tables corresponds to an implementation where all scenarios are loaded from the beginning, Run 2 uses the scenario addition heuristic of section 6 with $z_\alpha = 1.96$, Run 3 delays addition of these scenarios one major iteration forward, and Run 4 delays addition of scenarios two major iterations forward. Run 5 and Run 6 on MCR3-10,000 are identical to Run 1 and Run 2, except that the algorithm is terminated with a larger value of μ. These runs are included because for MCR3-10,000, using the information from $\hat{\sigma}_{N'}(\mu, \cdot)$, we estimate that a 10,000-scenario problem can produce a "statistical accuracy" (i.e., have a confidence interval that ensures that the objective is accurate to that degree) of about four digits. This accuracy is achieved for $\mu = 10^{-4}$. Although the additional major iterations solve the optimization problem more accurately, they do not improve the quality of the stochastic programming solution. To get an additional digit of "statistical" accuracy in the stochastic programs, we will require between 100,000 to one million scenarios. Solving problems of such sizes is not possible in the MATLAB environment. In Tables 15–26 row "# inner itr" gives the number of first stage Newton direction calculations"; row "# new sce add itr" gives the number of second stage centering iterations needed to add new scenarios while adaptively adding scenarios; row "# 2nd stage recenter itr" gives the total number of second stage centering iterations; row "# 2nd stage rectr calls" gives the total number of second stage calls for centering; row "avg new sce add itr" gives the average number of second stage iterations required to add a new scenario while adaptively adding scenarios; and row "avg recenter itrs" gives the average number of second stage Newton iterations per call.

We make the following observations based on results in Tables 15–16 for Problem MCR3-10,000. Table 16 gives relevant statistics on algorithm performance. First, we observe that a well-designed heuristic is necessary for gaining efficiency for adaptive scenario addition. Note that for Problem MCR3-10,000 Run 3 required significantly more iterations and time than Run 2. Run 4, which was more aggressive in delaying scenario addition, could not even solve the problems. In this run at $\mu = 10^{-4}$, after additional scenarios were loaded, the algorithm failed to recenter within 50 inner iterations and terminated. There are only four digits of accuracy in the solution at

TABLE 15
*Number of scenarios in different scenario addition strategies for MCR*3-10,000 ($p_{\%} = 5\%$, $\hat{\sigma}_{100}(10^{-3}, \hat{x}(10^{-3})) = 5.20$).

Outer itr.	μ	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7
1	1E+01	10,000	100	100	100	10,000	100	100
2	1E-01	10,000	100	100	100	10,000	100	100
3	1E-02	10,000	225	100	100	10,000	225	100
4	1E-03	10,000	3,300	225	100	10,000	3,300	225
5	1E-04	10,000	10,000	3,300	225	10,000	10,000	10,000
6	1E-05	10,000	10,000	10,000	3,300			
7	1E-06	10,000	10,000	10,000	10,000			
8	1E-07	10,000	10,000	10,000	10,000			

TABLE 16
*Computational performance: MCR*3-10,000.

	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7
# inner itr	25	27	41	70	16	18	21
# new sce add itr	10,016	62,926	75,923	-	10,016	62,926	68,866
# 2nd stage recenter itr	260,000	154,675	184,971	168,016	166,400	58,388	72,045
# 2nd stage rectr calls	250,000	154,575	184,800	167,875	160,000	58,350	71,900
final objective	204.622445	204.622445	204.622444	204.671617	204.621714	204.621671	204.620192
run time	3,238	2,043	2,526	1,971	2,141	970	1,323
avg new sce add itrs	1.00	6.29	7.59	-	1.00	6.29	6.89
avg recenter iters	1.04	1.00	1.00	1.00	1.04	1.00	1.00
infeasibility	5.54E-12	5.16E-12	2.42E-10	-	-	-	3.37E-10

this value of μ. Second, the warm-start for the existing scenario (after an update of x and/or μ) was effective. However, the addition of a new scenario took relatively more iterations. The computational savings in Run 2 over Run 1 are about 40%. This is because about half of the iterations in Run 2 are done with all the scenarios. When comparing the CPU times of Run 5 and Run 6 we observe that the CPU time savings are about 55%. Run 7, which aggressively delays scenario addition, performs worse than Run 6.

When analyzing the results in Tables 17–22 for Problems MCL3-10,000 and MCH3-10,000, we find a repeat of the above observations. Although Run 4 managed to optimize Problem MCL3-10,000, it took about 60% more time compared to Run 1. Run 4 for Problem MCH3-10,000 was just as efficient as Run 2. Note that more scenarios are added sooner for Problems MCL3-1,000 and MCH3-10,000 since they have larger values of $\hat{\sigma}_{N'}(\mu, \cdot)$. Interestingly, for these problems delaying scenario addition by one or two major iterations did not cause the algorithm to fail although it increased the number of inner iterations. We think this may be because for a larger value of μ it is easier to recover proximity to the first stage central path. However, for a smaller value of μ the recovery to the central path is harder since in this case the iterates are closer to the boundary of the first stage feasible set.

We observe that the average number of inner iterations taken when adding new scenarios in Run 1 increase slightly from MCR3-10,000 to MCL3-10,000 to MCH3-10,000. This shows that the ability to warm-start for an existing scenario depends on the value of $\sigma_{\hat{N}'}(\mu, \cdot)$. As $\sigma_{\hat{N}'}(\mu, \cdot)$ decreases, for a fixed number of scenarios the distance between a given scenario and its closest neighbor decreases. The same is true if the number of scenarios were to increase. In both cases, the quality of final iterates of neighboring scenarios as starting point increases.

7.2. Discussion on two-stage stochastic Markowitz problems. We solved the two-stage Markowitz problem with 20, 30, and 40 securities setting $\tau_{2,i} = 0.05$ and

TABLE 17
Number of scenarios in different scenario addition strategies for MCL3-10,000 ($p_{\%} = 50\%$, $\hat{\sigma}_{100}(10^{-3}, \hat{x}(10^{-3})) = 16.57$).

Outer itr.	μ	Run 1	Run 2	Run 3	Run 4
1	1E+01	10,000	100	100	100
2	1E-01	10,000	100	100	100
3	1E-02	10,000	850	100	100
4	1E-03	10,000	10,000	850	100
5	1E-04	10,000	10,000	10,000	850
6	1E-05	10,000	10,000	10,000	10,000
7	1E-06	10,000	10,000	10,000	10,000
8	1E-07	10,000	10,000	10,000	10,000

TABLE 18
Computational performance: MCL3-10,000.

	Run 1	Run 2	Run 3	Run 4
# inner itr	25	25	26	60
# new sce add itr	10,041	53,314	65,392	76,466
# 2nd stage recenter itr	260,744	153,385	124,517	370,784
# 2nd stage rectr calls	250,000	153,250	124,400	370,650
final objective	205.620644	205.620643	205.620643	205.620643
run time	2,699	1,983	1,776	4,349
avg new sce add itrs	1.00	5.33	6.54	7.65
avg recenter iters	1.04	1.00	1.00	1.00
infeasibility	5.52E-12	5.73E-12	5.92E-12	5.52E-12

TABLE 19
Number of scenarios in different scenario addition strategies for MCH3-10,000 ($p_{\%} = 500\%$, $\hat{\sigma}_{100}(10^{-3}, \hat{x}(10^{-3})) = 156.20$).

Outer itr.	μ	Run 1	Run 2	Run 3	Run 4
1	1E+01	10,000	100	100	100
2	1E-01	10,000	900	100	100
3	1E-02	10,000	10,000	900	100
4	1E-03	10,000	10,000	10,000	900
5	1E-04	10,000	10,000	10,000	10,000
6	1E-05	10,000	10,000	10,000	10,000
7	1E-06	10,000	10,000	10,000	10,000
8	1E-07	10,000	10,000	10,000	10,000

TABLE 20
Computational performance: MCH3-10,000.

	Run 1	Run 2	Run 3	Run 4
# inner itr	28	27	29	33
# new sce add itr	17,643	52,783	65,881	77,425
# 2nd stage recenter itr	314,253	207,982	186,794	174,610
# 2nd stage rectr calls	280,000	194,000	174,400	156,600
final objective	232.336743	232.336743	232.336743	232.336743
run time	3,223	2,490	2,373	2,321
avg new sce add itrs	1.76	5.28	6.59	7.74
avg recenter iters	1.12	1.07	1.07	1.12
infeasibility	8.73E-09	9.80E-09	9.70E-09	6.38E-09

TABLE 21
*Number of scenarios in different scenario addition strategies for MCL*33-10,000 *($p_{\%} = 50\%$, $\hat{\sigma}_{100}(10^{-3}, \hat{x}(10^{-3})) = 24.96$).*

Outer itr.	μ	Run 1	Run 2	Run 3	Run 4
1	1E+01	10,000	100	100	100
2	1E-01	10,000	100	100	100
3	1E-02	10,000	1,500	100	100
4	1E-03	10,000	10,000	1,500	100
5	1E-04	10,000	10,000	10,000	1,500
6	1E-05	10,000	10,000	10,000	10,000
7	1E-06	10,000	10,000	10,000	10,000
8	1E-07	10,000	10,000	10,000	10,000

TABLE 22
*Computational performance: MCL*33-10,000.

	Run 1	Run 2	Run 3	Run 4
# inner itr	26	26	29	81
# new sce add itr	34,777	96,439	108,506	119,561
# 2nd stage recenter itr	280,745	165,344	157,244	495,655
# 2nd stage rectr calls	270,000	165,200	157,000	494,300
final objective	204.112473	204.112473	204.112472	204.112473
run time	3,049	2,417	2,435	5,946
avg new sce add itrs	3.48	9.64	10.85	11.96
avg recenter iters	1.04	1.00	1.00	1.00
infeasibility	5.45E-12	5.75E-12	4.18E-10	5.19E-12

a 40-security instance of the problem setting $\tau_{2,i} = 0.25$. Decreasing the upper bound $\tau_{2,i}$ on the variation of the portfolio decomposition makes the two-stage Markowitz problem increasingly harder and eventually infeasible.

In Table 24 we report results for the $\tau_{2,i} = 0.05$ case. We performed two runs for each problem. In Run 1, we loaded all scenarios from the beginning and in Run 2 we added them adaptively following the strategies given in Table 23. For all problems, Run 2 uses the scenario addition heuristic of section 6 with $z_\alpha = 1.96$. With $\tau_{2,i} = 0.05$, strategy Warm-Start-1 was not stable. This is similar to the experience described in the previous section. Therefore, we performed addition of scenarios according to strategy Warm-Start-2. Since scenario additions require about 7 second stage iterations and the algorithm converges to optimality taking considerably fewer inner iterations than on the randomly generated test problems, here the adaptive scenario addition does not yield significant computational savings.

In Table 26, we report results for $\tau_{2,i} = 0.25$, where strategy Warm-Start-1 worked successfully. Our experience in this case was positively different. In Run 1 we loaded all scenarios in the first outer iteration. In Run 2 scenarios are added adaptively as discussed in section 6 with $z_\alpha = 1.96$. In Runs 3, 4, and 5 we added scenarios by warm-starting from the final iterate of a neighboring scenario at latest value of μ. Run 4 delays addition of scenarios one major iteration forward. Run 5 is very aggressive and delays addition of scenarios to the final two outer iterations. In Run 2 computational savings over Run 1 are about 30%. Due to more efficient addition of scenarios, savings in Run 3 increase by about 40%. Finally, Run 5 achieves 80% computational savings over Run 1. Also, note that even in Run 5, addition of scenarios was absorbed without taking additional inner iterations. All runs achieved optimality after 15 inner iterations. We estimate that with $\tau_{2,i} = 0.25$, a 15,000-scenario instance of the two-stage Markowitz problem with 40 securities can produce a "statistical accuracy" of

TABLE 23

Number of scenarios in different scenario addition strategies for two-stage Markowitz problems with 20, 30, *and* 40 *securities;* $\tau_{2,i} = 0.05$, $\hat{\sigma}_{100}(10^{-7}, \hat{x}(10^{-7})) = 6.65 \times 10^{-5}$, 4.92×10^{-5}, 3.68×10^{-5}, *for* $N = 20$, 30, *and* 40, *respectively.*

		dim $y_i = 20$		dim $y_i = 30$		dim $y_i = 40$	
Itr.	μ	Run 1	Run 2	Run 1	Run 2	Run 1	Run 2
1	1E-04	25,000	100	20,000	100	15,000	100
2	1E-05	25,000	100	20,000	100	15,000	100
3	1E-06	25,000	500	20,000	300	15,000	250
4	1E-07	25,000	12,500	20,000	7,820	15,000	5,200
5	1E-08	25,000	25,000	20,000	20,000	15,000	15,000
6	1E-09	25,000	25,000	20,000	20,000	15,000	15,000
7	1E-10	25,000	25,000	20,000	20,000	15,000	15,000
8	1E-11	25,000	25,000	20,000	20,000	15,000	15,000
9	1E-12	25,000	25,000	20,000	20,000	15,000	15,000

TABLE 24

Computational performance: with 20, 30, *and* 40 *securities;* $\tau_{2,i} = 0.05$.

	dim $y_i = 20$		dim $y_i = 30$		dim $y_i = 40$	
	Run 1	Run 2	Run 1	Run 2	Run 1	Run 2
# inner itr	17	21	18	21	18	18
# new sce add itr	78,997	178,461	26,075	140,021	15,861	101,362
# 2nd stage recenter itr	524,696	450,853	411,641	305,371	291,613	151,852
# new sce adds	25,000	49,900	20,000	20,000	15,000	15,000
# 2nd stage rectr calls	425,000	277,400	340,000	224,760	255,000	146,300
final objective	−2.07640524E-03	−2.07640524E-03	−1.88081179E-03	−1.88081179E-03	−1.68518238E-03	−1.68518238E-03
run time	1,932	1,896	1,634	1,518	1,754	1,211
avg new sce add itrs	3.16	7.14	1.30	7.00	1.06	6.76
avg recenter iters	1.23	1.63	1.21	1.36	1.14	1.04
infeasibility	2.60E-14	6.88E-14	4.29E-14	4.00E-14	9.41E-09	9.61E-09

TABLE 25
Number of scenarios in different scenario addition strategies for with 40 securities; $\tau_{2,i} = 0.25$, $\hat{\sigma}_{100}(10^{-7}, \hat{x}(10^{-7})) = 2.9610^{-5}$.

Itr.	μ	Run 1	Run 2	Run 3	Run 4	Run 5
1	1E-04	15,000	100	100	100	100
2	1E-05	15,000	100	100	100	100
3	1E-06	15,000	175	175	100	100
4	1E-07	15,000	4,150	4,150	175	100
5	1E-08	15,000	15,000	15,000	4,150	100
6	1E-09	15,000	15,000	15,000	15,000	100
7	1E-10	15,000	15,000	15,000	15,000	100
8	1E-11	15,000	15,000	15,000	15,000	15,000
9	1E-12	15,000	15,000	15,000	15,000	15,000

TABLE 26
Computational performance: with 40 securities; $\tau_{2,i} = 0.25$.

	Run 1	Run 2	Run 3	Run 4	Run 5
# inner itr	15	15	15	15	15
# new sce add itr	15,331	129,572	15,009	15,009	15,013
# 2nd stage recenter itr	281,249	129,320	159,122	129,320	46,470
# new sce adds	15,000	15,000	15,000	15,000	15,000
# 2nd stage rectr calls	225,000	114,050	143,850	114,050	46,200
final objective	−1.57070630E-03	−1.57070630E-03	−1.57070630E-03	−1.57070630E-03	−1.57070630E-03
run time	1,430	1,058	886	742	278
avg new sce add itrs	0.61	8.64	1.00	1.00	1.00
avg recenter iters	1.25	1.13	1.11	1.13	1.01
infeasibility	4.43E-14	4.45E-14	3.79E-14	4.35E-14	4.60E-14

about four digits. So, the relative easiness of the second stage problems are not because scenarios are closer to each other in terms of problem data. We think that the reason lies in the geometric simplicity of the problem that allows absorption of infeasibilities easily even when starting from neighboring scenario solutions for small values of μ.

Recall that the objective of the two-stage Markowitz problem is a concave quadratic function. We remark that our implementation successfully handled these quadratic terms without any numerical issues.

In summary, our experience suggests that adaptive addition of scenarios can be a very effective way of achieving computational savings, particularly when warm-starting the second stage problems from a nearby scenario is possible. However, it is also important to balance scenario addition with changes in the barrier function along the central path.

8. Conclusions. We have given a practical primal decomposition algorithm that follows the primal central path in the first stage. At each iteration, using approximate primal and dual solutions of the second stage barrier problems, it generates gradient and Hessian information for the first stage problem and takes a step along the Newton direction in the primal space. Several problems inherent in the context of primal algorithms were resolved using a preprocessing phase, and heuristics were developed for line search and scenario addition. These heuristics are based on empirically observed properties of the central path. A rigorous theoretical justification of these observed properties is a topic of future research.

Numerical experiments were conducted on a set of randomly generated problems and a two-stage extension of Markowitz's basic portfolio optimization model. Numerical experience suggests that we can solve the second stage centering problem approximately without compromising the performance of the decomposition algorithm. This experience also suggests that we need to follow the primal central path closely to develop stable implementations. Our results show that significant computational savings are possible for primal decomposition algorithms by adaptive addition of scenarios. These computational savings increase if it is possible to warm-start the solution of newly added second stage centering problems from the solution of a neighboring problem. The possibility of using primal predictor-corrector methods and integrating warm-starts with scenario generation may further improve the performance of primal decomposition methods and should be explored in the future.

It is possible to conceive other ways of adding scenarios in the adaptive scenario addition strategy. One example of such a strategy is a linear growth strategy where an equal number of scenarios are added in each major iteration. While a linear growth strategy is simple, and possibly useful, it is not consistent with what one would expect from the convergence rate of quasi–Monte Carlo methods. A linear growth strategy adds more scenarios in the early outer iterations, which is inefficient. In fact, it is easy to argue that with the linear growth strategy the maximum amount of computational savings cannot be more than 50% assuming that the average number of second stage Newton iterations per call is unchanged. Our goal here is to demonstrate that greater improvements in solution times are possible. The results on delayed addition of scenarios are used to demonstrate the importance of developing a strategy with a "reasonable" theoretical foundation.

REFERENCES

[1] F. Alizadeh and D. Goldfarb, *Second-order cone programming*, Math. Program., 95 (2003), pp. 3–51.

[2] F. Alizadeh and S. H. Schmieta, *Optimization with Semidefinite, Quadratic and Linear Constraints*, Technical report, RUTCOR, Rutgers University, 1997.

[3] O. Bahn, O. du Merle, J.-L. Goffin, and J. P. Vial, *A cutting plane method from analytic centers for stochastic programming*, Math. Programming, 69 (1995), pp. 45–73.

[4] J. R. Birge and F. Louveaux, *Introduction to Stochastic Programming*, Springer-Verlag, New York, 1997.

[5] P. Bratley and B. Fox, *Algorithm* 659*: Implementing Sobol's quasirandom sequence generator*, ACM Trans. Math. Software, 14 (1988), pp. 88–100.

[6] L. Faybusovich, *Euclidean Jordan algebras and interior-point algorithms*, Positivity, 1 (1997), pp. 331–357.

[7] L. Faybusovich, *Linear systems in Jordan algebras and primal-dual interior-point algorithms*, J. Comput. Appl. Math., 86 (1997), pp. 149–175.

[8] P. Glasserman, *Monte Carlo Methods in Financial Engineering*, Springer-Verlag, New York, 2003.

[9] D. Goldfarb and G. Iyengar, *Robust portfolio selection problems*, Math. Oper. Res., 28 (2003), pp. 1–38.

[10] M. Kojima, M. Shida, and S. Shindoh, *Global and Local Convergence of Predictor-Corrector Infeasible-Interior-Point Algorithms for Semidefinite Programs*, Research Reports on Information Sciences B-305, Department of Information Sciences, Tokyo Institute of Technology, Tokyo, Japan, 1995.

[11] O. Ledoit, P. Santa-Clara, and M. Wolf, *Flexible multivariate GARCH modeling with an application to international stock markets*, Rev. Econom. Statist., 85 (2003), pp. 735–747.

[12] J. Linderoth, A. Shapiro, and S. Wright, *The empirical behavior of sampling methods for stochastic programming*, Ann. Oper. Res., 142 (2006), pp. 215–241.

[13] I. J. Lustig, R. E. Marsten, and D. F. Shanno, *On implementing Mehrotra's predictor-corrector interior-point method for linear programming*, SIAM J. Optim., 2 (1992), pp. 435–449.

[14] H. M. Markowitz, *Portfolio selection*, J. Finance, 46 (1952), pp. 469–477.

[15] S. Mehrotra, *On the implementation of a primal-dual interior point method*, SIAM J. Optim., 2 (1992), pp. 575–601.

[16] S. Mehrotra, *Properties of a Weighted Barrier Function and a Decomposition Algorithm for Stochastic Programs with Continuous Support*, Technical report, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL, 2008.

[17] S. Mehrotra and M. G. Özevin, *Decomposition based interior point methods for two-stage stochastic convex quadratic programs with recourse*, Oper. Res., to appear.

[18] S. Mehrotra and M. G. Özevin, *Decomposition-based interior point methods for two-stage stochastic semidefinite programming*, SIAM J. Optim., 18 (2007), pp. 206–222.

[19] S. Mehrota and M. G. Özevin, *Convergence of a Weighted Barrier Decomposition Algorithm for Two Stage Stochastic Programming with Discrete Support*, Technical report, IEMS Department, Northwestern University, 2007.

[20] R. O. Michaud, *Efficient Asset Management: A Practical Guide to Stock Portfolio Management and Asset Allocation*, Financial Management Association Survey and Synthesis Series, HBS Press, Boston, 1998.

[21] Yu. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, Stud. Appl. Math. 13, SIAM, Philadelphia, 1994.

[22] Yu. E. Nesterov and M. J. Todd, *Primal-dual interior-point methods for self-scaled cones*, SIAM J. Optim., 8 (1998), pp. 324–364.

[23] F. A. Potra and R. Sheng, *A superlinearly convergent primal-dual infeasible-interior-point algorithm for semidefinite programming*, SIAM J. Optim., 8 (1998), pp. 1007–1028.

[24] J. Renegar, *A Mathematical View of Interior-Point Methods in Convex Programming*, MPS-SIAM Ser. Optim. 3, SIAM, Philadelphia, 2001.

[25] A. Shapiro, *Asymptotic analysis of stochastic programs*, Ann. Oper. Res., 30 (1991), pp. 169–186.

[26] J. Sturm, *Using SeDuMi* 1.02*, a MATLAB toolbox for optimization over symmetric cones*, Optim. Methods Softw., 11–12 (1999), pp. 625–653.

[27] J. Sturm, *Implementation of interior point methods for mixed semidefinite and second order cone optimization problems*, Optim. Methods Softw. 17 (2002), pp. 1105–1154.

[28] J. Sturm, *Avoiding numerical cancellation in the interior point method for solving semidefinite programs*, Math. Program., 95 (2003), pp. 219–247.
[29] M. J. Todd, K. C. Toh, and R. H. Tütüncü, *On the Nesterov–Todd direction in semidefinite programming*, SIAM J. Optim., 8 (1998), pp. 769–796.
[30] R. M. Van Slyke and R. Wets, *L-shaped linear programs with applications to optimal control and stochastic linear programming*, SIAM J. Appl. Math., 17 (1969), pp. 638–663.
[31] G. Zhao, *A log-barrier method with Benders decomposition for solving two-stage stochastic linear programs*, Math. Program., 90 (2001), pp. 507–536.

SIAM J. OPTIM.
Vol. 19, No. 4, pp. 1881–1893

A USE OF CONJUGATE GRADIENT DIRECTION FOR THE CONVEX OPTIMIZATION PROBLEM OVER THE FIXED POINT SET OF A NONEXPANSIVE MAPPING*

HIDEAKI IIDUKA† AND ISAO YAMADA‡

Abstract. In this paper, we discuss the convex optimization problem over the fixed point set of a nonexpansive mapping. The main objective of the paper is to accelerate the hybrid steepest descent method for the problem. To this goal, we present a new iterative scheme that utilizes the conjugate gradient direction. Its convergence to the solution is guaranteed under certain assumptions. In order to demonstrate the effectiveness, performance, and convergence of our proposed algorithm, we present numerical comparisons of the algorithm with the existing algorithm.

Key words. convex optimization problem, nonexpansive mapping, fixed point, hybrid steepest descent method, conjugate gradient direction

AMS subject classifications. 47H07, 47H09, 65K05, 65K10, 90C25, 90C30, 90C52

DOI. 10.1137/070702497

1. Introduction.

Let H be a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and its induced norm $\| \cdot \|$. It is well known that the standard *smooth convex optimization problem* [10, 24, 33, 45], given a convex, Fréchet differentiable function $f\colon H \to \mathbb{R}$ and a closed convex subset C of H,

$$\text{find a point } x^* \in C \text{ such that } f(x^*) = \min_{x \in C} f(x), \tag{1.1}$$

can be formulated equivalently as the *variational inequality problem* [23, 24, 36, 45] over C:

$$\text{find } x^* \in C \text{ such that } \langle v - x^*, \nabla f(x^*) \rangle \geq 0 \text{ for all } v \in C, \tag{1.2}$$

where $\nabla f\colon H \to H$ is the gradient of f. The simplest iterative scheme for (1.1) is the well-known *projected gradient method* [17]: $x_1 \in C$ and $x_{n+1} = P_C(x_n - \mu \nabla f(x_n))$ for every $n \in \mathbb{N}$, where P_C is the *metric projection* from H onto C (see section 2) and μ is a positive real number. This method requires repetitive use of P_C, although the closed form expression of P_C is not always known in many situations. To help resolve this problem, the following *hybrid steepest descent method* [42, 43, 44] for (1.2) when C is equal to the *fixed point set* $\mathrm{Fix}(T) := \{x \in H : T(x) = x\}$ of a *nonexpansive* mapping T [2, 3, 15, 16, 32, 37, 38] has been established: $x_1 \in H$ and

$$x_{n+1} = T(x_n - \mu \alpha_n \nabla f(x_n)) \text{ for every } n \in \mathbb{N}, \tag{1.3}$$

where $\mu > 0$, $(\alpha_n)_{n \in \mathbb{N}} \subset (0, 1]$ is a slowly diminishing constant sequence, and $\nabla f\colon H \to H$ is *strongly monotone* and *Lipschitz continuous* (see section 2). By the nonexpansivity of P_C, the method (1.3) is the same as the projected gradient method when

*Received by the editors September 10, 2007; accepted for publication (in revised form) October 27, 2008; published electronically February 27, 2009. This work was supported by the Japan Society for the Promotion of Science (JSPS) Grants-in-Aid (19001979).
http://www.siam.org/journals/siopt/19-4/70249.html

†Network Design Research Center, Kyushu Institute of Technology, Hibiya Kokusai Bldg. 1F 107, 2-2-3 Uchisaiwai-cho, Chiyoda-ku, Tokyo, 100-0011, Japan (iiduka@ndrc.kyutech.ac.jp).

‡Department of Communications and Integrated Systems, Tokyo Institute of Technology, Ookayama, Meguro-ku, Tokyo, 152-8552, Japan (isao@comm.ss.titech.ac.jp).

$T := P_C$ and $\alpha_n := 1$ $(n \in \mathbb{N})$. The projected gradient and hybrid steepest descent methods converge strongly to the uniquely existing solution of (1.2) when $C = \mathrm{Fix}(T)$ [42, 43, 44]. Recently, the method (1.3) has been applied successfully to signal processing, inverse problems, and so on [34, 35, 44]. Other algorithms for solving the problem (1.2) have been proposed in [5] and [22]. In [5], an effective scheme for solving the signal recovery problem has been proposed, and this method converges strongly to the solution without using a diminishing constant sequence. In [22], the problem which contains (1.2) and an iterative algorithm for this problem have been presented.

In the case where $C = H = \mathbb{R}^N$, iterative procedures for (1.1) [1, 6, 7, 8, 11, 12, 13, 14, 19, 20, 21, 25, 26, 27, 28, 29, 30, 31, 39, 40, 46] have a long history and have been studied extensively, for example, the *Newton method*, the *quasi-Newton methods*, the *steepest descent method*, and the *conjugate gradient methods*. These have the common form

$$x_{n+1} = x_n + \alpha_n d_n, \tag{1.4}$$

where $x_n \in \mathbb{R}^N$ is the nth approximation to the solution, $\alpha_n > 0$ is a step size, and $d_n \in \mathbb{R}^N$ is a search direction. The Newton method and the quasi-Newton methods are known as fast convergent iterative methods for solving (1.1) when $C = \mathbb{R}^N$. To compute the search direction, the Newton method requires us to solve the Newton equation $\nabla^2 f(x_n) d_n = -\nabla f(x_n)$ $(n \in \mathbb{N})$, that is, to invert the Hessian $\nabla^2 f(x_n)$. If $\nabla^2 f(x_n)$ is positive definite, then the solution d_n of this equation satisfies the descent condition, that is, $\langle d_n, \nabla f(x_n)\rangle < 0$ $(n \in \mathbb{N})$. Even when the objective function f does not have its positive definite Hessian, $\nabla^2 f(x_n)$ or $\nabla^2 f(x_n)^{-1}$ is often replaced by a simpler, positive definite matrix to define a simpler search direction. Such a method is called the quasi-Newton method. In particular, the Davidon–Fletcher–Powell and Broyden–Fletcher–Goldfarb–Shanno methods have been used as effective algorithms for solving (1.1) when $C = \mathbb{R}^N$. The steepest descent method does not need any matrix inversion because this method always utilizes the steepest descent direction $d_n = -\nabla f(x_n)$ $(n \in \mathbb{N})$. Acceleration of the steepest descent method has been of great interest. Much research along this direction covers, for example, the conjugate gradient methods [1, 7, 8, 13, 14, 19, 29, 30, 31, 46], the *three-term-recurrence method* [27], and the *memory gradient methods* [6, 25, 26]. In particular, the conjugate gradient methods have been used widely as an efficient accelerated version of the most gradient methods. We define the conjugate gradient direction as follows:

$$d_n = -\nabla f(x_n) + \beta_n d_{n-1}, \tag{1.5}$$

where $\beta_n \in \mathbb{R}$. Many design schemes for β_n have been developed, for example, the Fletcher–Reeves formula $\beta_n^{\mathrm{FR}} = \|\nabla f(x_n)\|^2/\|\nabla f(x_{n-1})\|^2$ [13, 1, 7, 14], the Polak–Ribière–Polyak formula $\beta_n^{\mathrm{PRP}} = \langle \nabla f(x_n), y_{n-1}\rangle/\|\nabla f(x_{n-1})\|^2$ [29, 30, 14], the Hestenes–Stiefel formula $\beta_n^{\mathrm{HS}} = \langle \nabla f(x_n), y_{n-1}\rangle/\langle d_{n-1}, y_{n-1}\rangle$ [19, 14], and the Dai–Yuan formula $\beta_n^{\mathrm{DY}} = \|\nabla f(x_n)\|^2/\langle d_{n-1}, y_{n-1}\rangle$ [8], where $y_n := \nabla f(x_{n+1}) - \nabla f(x_n)$. The FR method [1, 7, 14] and the DY method [8] determine β_n in such a way that the descent condition is satisfied.

The goal of this paper is to accelerate the hybrid steepest descent method (1.3). To achieve this goal, we present a new iterative scheme (Algorithm 3.4) by combining two ideas: One is the hybrid steepest descent method (1.3) for the variational inequality problem over the fixed point set of a nonexpansive mapping, and the other is the conjugate gradient method (1.4) and (1.5) for the unconstrained optimization problem.

The rest of this paper is divided into four sections. In section 2, we state preliminaries on fixed points, nonexpansive mapping, metric projection, convexity, continuity, and monotone operator. In section 3, we propose an iterative algorithm (Algorithm 3.4) that utilizes a new search direction. In section 4, we apply our method to the convex optimization problem and present convergence analysis (Theorem 4.1) under some assumptions. In section 5, to demonstrate the effectiveness, performance, and convergence of the proposed algorithm, we present numerical comparisons of the algorithm with the hybrid steepest descent method (1.3).

2. Preliminaries.

2.1. Convexity and monotonicity. A function $f\colon H \to \mathbb{R}$ is said to be *convex* if for any $x, y \in H$ and for any $\lambda \in [0,1]$, $f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda)f(y)$. In particular, a convex function $f\colon H \to \mathbb{R}$ is said to be *strongly convex* with $\alpha > 0$ (α-strongly convex) if $f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda)f(y) - (\alpha\lambda(1-\lambda)/2)\|x-y\|^2$ for all $x, y \in H$ and for all $\lambda \in [0,1]$. For example, $f(x) := (1/2)\langle x, Q(x)\rangle + \langle b, x\rangle$ ($x \in H$) is α-strongly convex, where $b \in H$ and $Q\colon H \to H$ is a self-adjoint, bounded, linear operator satisfying $\langle x, Q(x)\rangle \ge \alpha\|x\|^2$ for some $\alpha > 0$ and for all $x \in H$ (for details, see Example 3.2 of this paper). In particular, if Q is the identity mapping on H, then f is 1-strongly convex.

An operator $A\colon H \to H$ is said to be *monotone* if $\langle x-y, A(x)-A(y)\rangle \ge 0$ for all $x, y \in H$. $A\colon H \to H$ is called an *α-strongly monotone* operator if $\langle x-y, A(x)-A(y)\rangle \ge \alpha\|x-y\|^2$ for all $x, y \in H$.

Suppose that $f\colon H \to \mathbb{R}$ is a continuously Fréchet differentiable function. Then f is convex if and only if the gradient ∇f is monotone [4, 20, 21]. It is also known [20, 21] that f is α-strongly convex if and only if ∇f is α-strongly monotone. Let $f\colon \mathbb{R}^N \to \mathbb{R}$ be a twice continuously differentiable function. Then f is convex if and only if, for all $x \in \mathbb{R}^N$, the Hessian $\nabla^2 f(x)$ is positive semidefinite on $\mathbb{R}^N$ [4, 20, 21]. It is also known [11, 20, 21] that f is α-strongly convex if and only if, for all $x \in \mathbb{R}^N$, the matrix $\nabla^2 f(x) - \alpha E$ is positive semidefinite, that is, $\langle y, \nabla^2 f(x)(y)\rangle \ge \alpha\|y\|^2$ for all $y \in \mathbb{R}^N$.

2.2. Nonexpansive mapping and fixed point. A mapping $T\colon H \to H$ is said to be *Lipschitz continuous* with $L > 0$ (L-Lipschitz continuous) if $\|T(x) - T(y)\| \le L\|x-y\|$ for all $x, y \in H$. When $T\colon H \to H$ is 1-Lipschitz continuous, T is said to be *nonexpansive*. It is well known that the *fixed point set* $\mathrm{Fix}(T) := \{x \in H\colon T(x) = x\}$ of a nonexpansive mapping T is closed and convex [2, 16, 38]. $T\colon H \to H$ is called a *firmly nonexpansive* mapping if for all $x, y \in H$, $\|T(x)-T(y)\|^2 \le \langle x-y, T(x)-T(y)\rangle$. Given a nonempty, closed convex subset C of H, the mapping that assigns every point $x \in H$ to its unique nearest point in C is called the *metric projection* onto C and is denoted by P_C, that is, $\|x - P_C(x)\| = \inf_{y\in C}\|x-y\|$. The metric projection P_C is a typical example of firmly nonexpansive mapping satisfying $\mathrm{Fix}(P_C) = C$. Some closed convex set C, for example, a linear variety, a closed ball, a closed cone, and a closed polytope, is simple in the sense that the closed form expression of P_C is known, which implies that P_C can be computed within a finite number of arithmetic operations [2, 9, 41].

To prove the main theorem of this paper, we need the following lemma.

LEMMA 2.1 (see [42, 44]). *Let $T\colon H \to H$ be a nonexpansive mapping and $f\colon H \to \mathbb{R}$ a continuously Fréchet differentiable function. Suppose that $\nabla f\colon H \to H$ is α-strongly monotone and L-Lipschitz continuous and $\mu \in (0, 2\alpha/L^2)$. Define $T^\lambda(x) :=$*

$T(x - \mu\lambda\nabla f(x))$ for all $x \in H$, where $\lambda \in [0,1]$. Then for all $x, y \in H$,

$$\|T^{\lambda}(x) - T^{\lambda}(y)\| \leq (1 - \lambda\tau)\|x - y\|,$$

where $\tau := 1 - \sqrt{1 - \mu(2\alpha - \mu L^2)} \in (0, 1]$.

3. Optimization over the fixed point set of a nonexpansive mapping. In this paper, we consider the following constrained optimization problem.

PROBLEM 3.1. *Assume that*

(A1) *$T\colon H \to H$ is a nonexpansive mapping with $\mathrm{Fix}(T) \neq \emptyset$;*

(A2) *$f\colon H \to \mathbb{R}$ is continuously Fréchet differentiable;*

(A3) *$\nabla f\colon H \to H$ is α-strongly monotone and L-Lipschitz continuous.*

Our objective is to

$$\text{find a point } x^* \in \mathrm{Fix}(T) \text{ such that } f(x^*) = \min_{x \in \mathrm{Fix}(T)} f(x).$$

Note that under the assumptions (A1), (A2), and (A3), the existence and the uniqueness of the minimizer $x^* \in \mathrm{Fix}(T)$ of f over $\mathrm{Fix}(T)$ is guaranteed [42]. A well-known example of a convex function satisfying the assumptions (A2) and (A3) is the following.

Example 3.2 (see [42]). Let $b \in H$ and $Q\colon H \to H$ be a self-adjoint, bounded, linear operator and strongly positive; that is, there exists $\alpha > 0$ such that $\langle x, Q(x)\rangle \geq \alpha\|x\|^2$ for all $x \in H$. Define a quadratic function $f\colon H \to \mathbb{R}$ by

$$f(x) := \frac{1}{2}\langle x, Q(x)\rangle + \langle b, x\rangle \text{ for all } x \in H.$$

Then $\nabla f(\cdot) = Q(\cdot) + b$ is α-strongly monotone and $\|Q\|$-Lipschitz continuous, where $\|Q\| := \sup_{x \neq 0} |\langle x, Q(x)\rangle| \|x\|^{-2}$.

Remark 3.3. When H is finite dimensional, the above operator Q coincides with a positive definite matrix. Then, for all $x \in \mathbb{R}^N$, $\nabla^2 f(x) = Q$ and $\lambda_{\min}\|x\|^2 \leq \langle x, Q(x)\rangle \leq \lambda_{\max}\|x\|^2$, where $\lambda_{\min}$ and $\lambda_{\max}$ are, respectively, the minimum and maximum eigenvalues of Q. Hence $\alpha = \lambda_{\min} \leq \lambda_{\max} = \|Q\|$.

For Problem 3.1, we present an algorithm using a conjugate gradient direction.

ALGORITHM 3.4. *Let $f\colon H \to \mathbb{R}$ and $T\colon H \to H$ satisfy the conditions* (A1), (A2), *and* (A3) *in Problem* 3.1.

Step 0. *Take $\mu > 0$. Choose $x_1 \in H$ and $\alpha_1 \in (0, 1]$ arbitrarily, and set $d_1 := -\nabla f(x_1)$ and $n := 1$.*

Step 1. *Given $x_n \in H$ and $d_n \in H$, choose $\alpha_n \in (0, 1]$ (see Theorem* 4.1*) and define $x_{n+1} \in H$ by*

$$x_{n+1} := T(x_n + \mu\alpha_n d_n). \tag{3.1}$$

Step 2. *Choose $\beta_{n+1} \in [0, \infty)$ (see Theorem* 4.1*), and update the search direction as*

$$d_{n+1} := -\nabla f(x_{n+1}) + \beta_{n+1} d_n. \tag{3.2}$$

Put $n := n + 1$, and go to Step 1.

In the case where T is the identity mapping on H and $\mu := 1$, Algorithm 3.4 coincides with the conjugate gradient method for the unconstrained optimization problem. By (3.1) and (3.2), we can see that the search direction d_n is defined by combining the ideas of the hybrid steepest descent method (1.3) and the conjugate gradient method (1.5).

4. Convergence theorem for the iterative algorithm. We present convergence analysis on Algorithm 3.4.

THEOREM 4.1. *Suppose that* $\mu \in (0, 2\alpha/L^2)$, *and* $(\alpha_n)_{n\in\mathbb{N}} \subset (0,1]$ *and* $(\beta_n)_{n\geq 2} \subset [0,\infty)$ *satisfy* (i) $\lim_{n\to\infty} \alpha_n = 0$, (ii) $\sum_{n=1}^{\infty} \alpha_n = \infty$, (iii) $\sum_{n=1}^{\infty} |\alpha_{n+1} - \alpha_n| < \infty$, (iv) $\alpha_n/\alpha_{n+1} \leq \sigma$ *for some* $\sigma \geq 1$ *and for every* $n \in \mathbb{N}$, *and* (v) $\lim_{n\to\infty} \beta_n = 0$ (an example of satisfying (i)–(v) is $\alpha_n := 1/(n+1)^{\rho}$ and $\beta_{n+1} := 1/(n+1)^{\gamma}$ $(n \in \mathbb{N}, \rho \in (0,1]$, and $\gamma > 0)$). *If* $(\nabla f(x_n))_{n\in\mathbb{N}}$ *is bounded, the sequence* $(x_n)_{n\in\mathbb{N}}$ *generated by Algorithm* 3.4 *satisfies the following:*

(a) $(x_n)_{n\in\mathbb{N}}$ *and* $(d_n)_{n\in\mathbb{N}}$ *are bounded. Moreover,* $\lim_{n\to\infty} \|x_{n+1} - T(x_n)\| = 0$.
(b) $\lim_{n\to\infty} \|x_{n+1} - x_n\| = 0$ *and* $\lim_{n\to\infty} \|x_n - T(x_n)\| = 0$.
(c) *The sequence* $(x_n)_{n\in\mathbb{N}}$ *converges strongly to the uniquely existing solution of Problem* 3.1.

Proof. (a) We shall prove that $(d_n)_{n\in\mathbb{N}}$ in Algorithm 3.4 is bounded. By the condition (v), there exists $m_1 \in \mathbb{N}$ such that $\beta_n \leq 1/2$ for all $n \geq m_1$. Put $K_1 := \sup\{\|\nabla f(x_n)\| : n \in \mathbb{N}\} < \infty$ and $K_2 := \max\{K_1, \|d_{m_1}\|\}$. It is obvious from the definition of K_2 that $\|d_{m_1}\| \leq 2K_2$. From (3.2), we have

$$\|d_{n+1}\| \leq \|\nabla f(x_{n+1})\| + \beta_{n+1}\|d_n\| \leq K_2 + \frac{1}{2}\|d_n\| \tag{4.1}$$

for all $n \geq m_1$. Suppose that $\|d_n\| \leq 2K_2$ for some $n \geq m_1$. Then, by (4.1), we get $\|d_{n+1}\| \leq 2K_2$. By induction, we obtain $\|d_n\| \leq 2K_2$ for all $n \geq m_1$, and hence $(d_n)_{n\in\mathbb{N}}$ is bounded.

Let $x^* \in \mathrm{Fix}(T)$ be the solution of Problem 3.1, and let $\tau \in (0,1]$ be as in Lemma 2.1. Put $K_3 := \sup\{\|\beta_{n+1}d_n - \nabla f(x^*)\| : n \in \mathbb{N}\} < \infty$ and $K := \max\{\|\nabla f(x^*)\|, K_3\}$. Then, by Lemma 2.1, for every $n \geq 2$, we have

$$\begin{aligned}
\|x_{n+1} - x^*\| &= \|T(x_n + \mu\alpha_n d_n) - T(x^*)\| \\
&\leq \|(x_n + \mu\alpha_n(-\nabla f(x_n) + \beta_n d_{n-1})) - x^*\| \\
&= \|(x_n - \mu\alpha_n\nabla f(x_n)) - (x^* - \mu\alpha_n\nabla f(x^*)) \\
&\qquad + \mu\alpha_n(\beta_n d_{n-1} - \nabla f(x^*))\| \\
&\leq \|(x_n - \mu\alpha_n\nabla f(x_n)) - (x^* - \mu\alpha_n\nabla f(x^*))\| \\
&\qquad + \mu\alpha_n\|\beta_n d_{n-1} - \nabla f(x^*)\| \\
&\leq (1 - \tau\alpha_n)\|x_n - x^*\| + \left(\frac{\mu K}{\tau}\right)\tau\alpha_n.
\end{aligned}$$

By $x_2 = T(x_1 - \mu\alpha_1\nabla f(x_1))$ and the definition of K, the inequality above holds for $n = 1$. Therefore, by induction, we obtain

$$\|x_n - x^*\| \leq \max\left\{\|x_1 - x^*\|, \frac{\mu K}{\tau}\right\} \text{ for every } n \in \mathbb{N};$$

that is, $(x_n)_{n\in\mathbb{N}}$ is bounded.

By (3.1) and the nonexpansivity of T, we have $\limsup_{n\to\infty} \|x_{n+1} - T(x_n)\| \leq \mu \limsup_{n\to\infty} \alpha_n\|d_n\|$. The condition (i) and the boundedness of $(d_n)_{n\in\mathbb{N}}$ imply

$$\lim_{n\to\infty} \|x_{n+1} - T(x_n)\| = 0. \tag{4.2}$$

(b) We shall prove $\lim_{n\to\infty} \|x_{n+1} - x_n\| = 0$. Put $z_n := x_n + \mu\alpha_n d_n$ $(n \in \mathbb{N})$, $M_1 := \sup\{|\langle z_n - z_{n-1}, \nabla f(x_{n-1})\rangle| : n \geq 2\} < \infty$, $M_2 := \sup\{|\langle z_n - z_{n-1}, d_{n-1}\rangle|/\tau :$

$n \geq 2\} < \infty$, $M_3 := \sup\{\sigma|\langle z_n - z_{n-1}, d_{n-2}\rangle|/\tau : n \geq 3\} < \infty$, and $M := \max\{M_1, M_2, M_3\}$. It follows from (3.1), (3.2), Lemma 2.1, the nonexpansivity of T, and the condition (iv) that

$$\begin{aligned}
\|x_{n+1} - x_n\|^2 &= \|T(x_n + \mu\alpha_n d_n) - T(x_{n-1} + \mu\alpha_{n-1}d_{n-1})\|^2 \\
&\leq \|(x_n + \mu\alpha_n d_n) - (x_{n-1} + \mu\alpha_{n-1}d_{n-1})\|^2 \\
&= \|(x_n + \mu\alpha_n(-\nabla f(x_n) + \beta_n d_{n-1})) \\
&\qquad - (x_{n-1} - \mu\alpha_n\nabla f(x_{n-1})) - \mu\alpha_n\nabla f(x_{n-1}) - \mu\alpha_{n-1}d_{n-1}\|^2 \\
&= \|(x_n - \mu\alpha_n\nabla f(x_n)) - (x_{n-1} - \mu\alpha_n\nabla f(x_{n-1})) \\
&\qquad + \mu(\alpha_n\beta_n d_{n-1} - \alpha_n\nabla f(x_{n-1}) - \alpha_{n-1}d_{n-1})\|^2 \\
&\leq \|(x_n - \mu\alpha_n\nabla f(x_n)) - (x_{n-1} - \mu\alpha_n\nabla f(x_{n-1}))\|^2 \\
&\qquad + 2\mu\langle\alpha_n\beta_n d_{n-1} - \alpha_n\nabla f(x_{n-1}) - \alpha_{n-1}d_{n-1}, z_n - z_{n-1}\rangle \\
&\leq (1 - \tau\alpha_n)^2\|x_n - x_{n-1}\|^2 + 2\mu\langle\alpha_n\beta_n d_{n-1} - \alpha_n\nabla f(x_{n-1}) \\
&\qquad - \alpha_{n-1}(-\nabla f(x_{n-1}) + \beta_{n-1}d_{n-2}), z_n - z_{n-1}\rangle \\
&\leq (1 - \tau\alpha_n)\|x_n - x_{n-1}\|^2 + 2\mu(\alpha_{n-1} - \alpha_n)\langle z_n - z_{n-1}, \nabla f(x_{n-1})\rangle \\
&\qquad + 2\mu\alpha_n\beta_n\langle z_n - z_{n-1}, d_{n-1}\rangle + 2\mu\alpha_{n-1}\beta_{n-1}\langle z_n - z_{n-1}, -d_{n-2}\rangle \\
&\leq (1 - \tau\alpha_n)\|x_n - x_{n-1}\|^2 + 2\mu M|\alpha_{n-1} - \alpha_n| \\
&\qquad + 2\mu M\tau\alpha_n\beta_n + 2\mu M\tau\frac{\alpha_{n-1}}{\sigma}\beta_{n-1} \\
&\leq (1 - \tau\alpha_n)\|x_n - x_{n-1}\|^2 + 2\mu M|\alpha_{n-1} - \alpha_n| \\
&\qquad + 2\mu M\tau\alpha_n\beta_n + 2\mu M\tau\alpha_n\beta_{n-1}
\end{aligned}$$

for every $n \geq 3$. By the condition (v), for any $\varepsilon > 0$, there exists $n_1 \in \mathbb{N}$ such that $\beta_n \leq \varepsilon/4$ for all $n \geq n_1$. For all $n \geq n_1 + 1$, we have

$$\begin{aligned}
\|x_{n+1} - x_n\|^2 \leq (1 - \tau\alpha_n)\|x_n - x_{n-1}\|^2 &+ 2\mu M|\alpha_{n-1} - \alpha_n| \\
&+ \mu M\varepsilon(1 - (1 - \tau\alpha_n)).
\end{aligned}$$

So, we obtain that for all $n, m \geq n_1$,

$$\begin{aligned}
\|x_{n+m+1} - x_{n+m}\|^2 &\leq (1 - \tau\alpha_{n+m})\|x_{n+m} - x_{n+m-1}\|^2 + 2\mu M|\alpha_{n+m} - \alpha_{n+m-1}| \\
&\qquad + \mu M\varepsilon(1 - (1 - \tau\alpha_{n+m})) \\
&\leq (1 - \tau\alpha_{n+m})\{(1 - \tau\alpha_{n+m-1})\|x_{n+m-1} - x_{n+m-2}\|^2 \\
&\qquad + 2\mu M|\alpha_{n+m-1} - \alpha_{n+m-2}| + \mu M\varepsilon(1 - (1 - \tau\alpha_{n+m-1}))\} \\
&\qquad + 2\mu M|\alpha_{n+m} - \alpha_{n+m-1}| + \mu M\varepsilon(1 - (1 - \tau\alpha_{n+m})) \\
&\leq (1 - \tau\alpha_{n+m})(1 - \tau\alpha_{n+m-1})\|x_{n+m-1} - x_{n+m-2}\|^2 \\
&\qquad + 2\mu M(|\alpha_{n+m} - \alpha_{n+m-1}| + |\alpha_{n+m-1} - \alpha_{n+m-2}|) \\
&\qquad + \mu M\varepsilon\{1 - (1 - \tau\alpha_{n+m})(1 - \tau\alpha_{n+m-1})\} \\
&\leq \prod_{k=m}^{n+m-1}(1 - \tau\alpha_{k+1})\|x_{m+1} - x_m\|^2 + 2\mu M\sum_{k=m}^{n+m-1}|\alpha_{k+1} - \alpha_k| \\
&\qquad + \mu M\varepsilon\left(1 - \prod_{k=m}^{n+m-1}(1 - \tau\alpha_{k+1})\right).
\end{aligned}$$

Since the condition (ii) implies $\prod_{k=m}^{\infty}(1-\tau\alpha_{k+1})=0$, we have

$$\begin{aligned}\limsup_{n\to\infty}\|x_{n+1}-x_n\|^2 &= \limsup_{n\to\infty}\|x_{n+m+1}-x_{n+m}\|^2\\ &\le 2\mu M\limsup_{n\to\infty}\sum_{k=m}^{n+m-1}|\alpha_{k+1}-\alpha_k|+\mu M\varepsilon\end{aligned}$$

for every $m\ge n_1$. From the condition (iii), we obtain $\limsup_{n\to\infty}\|x_{n+1}-x_n\|^2\le \mu M\varepsilon$ for arbitrarily small $\varepsilon>0$, and hence

$$\lim_{n\to\infty}\|x_{n+1}-x_n\|=0. \tag{4.3}$$

Now, by (4.2) and (4.3), we have

$$\lim_{n\to\infty}\|x_n-T(x_n)\|=0. \tag{4.4}$$

(c) The first step is to show $\limsup_{n\to\infty}\langle x^*-x_n,\nabla f(x^*)\rangle\le 0$. Choose a subsequence $(x_{n_i})_{i\in\mathbb{N}}$ of $(x_n)_{n\in\mathbb{N}}$ such that

$$\limsup_{n\to\infty}\langle x^*-x_n,\nabla f(x^*)\rangle=\lim_{i\to\infty}\langle x^*-x_{n_i},\nabla f(x^*)\rangle.$$

Boundedness of $(x_{n_i})_{i\in\mathbb{N}}$ implies the existences of a subsequence $(x_{n_{i_j}})_{j\in\mathbb{N}}$ of $(x_{n_i})_{i\in\mathbb{N}}$ and a point $\hat{x}\in H$ such that $\lim_{j\to\infty}\langle x_{n_{i_j}},w\rangle=\langle\hat{x},w\rangle$ $(w\in H)$. We may assume without loss of generality that $\lim_{i\to\infty}\langle x_{n_i},w\rangle=\langle\hat{x},w\rangle$ $(w\in H)$. Assume $\hat{x}\ne T(\hat{x})$. By (4.4) and the nonexpansivity of T, we have a contradiction:

$$\begin{aligned}\liminf_{i\to\infty}\|x_{n_i}-\hat{x}\| &< \liminf_{i\to\infty}\|x_{n_i}-T(\hat{x})\| = \liminf_{i\to\infty}\|x_{n_i}-T(x_{n_i})+T(x_{n_i})-T(\hat{x})\|\\ &= \liminf_{i\to\infty}\|T(x_{n_i})-T(\hat{x})\|\le\liminf_{i\to\infty}\|x_{n_i}-\hat{x}\|.\end{aligned}$$

This implies $\hat{x}\in\mathrm{Fix}(T)$. Since $x^*\in\mathrm{Fix}(T)$ satisfies $\langle v-x^*,\nabla f(x^*)\rangle\ge 0$ $(v\in\mathrm{Fix}(T))$, we get

$$\begin{aligned}\limsup_{n\to\infty}\langle x^*-x_n,\nabla f(x^*)\rangle &= \lim_{i\to\infty}\langle x^*-x_{n_i},\nabla f(x^*)\rangle\\ &= \langle x^*-\hat{x},\nabla f(x^*)\rangle\le 0.\end{aligned}$$

Moreover, by the condition (v) and the boundedness of $(z_n)_{n\in\mathbb{N}}$ and $(d_n)_{n\in\mathbb{N}}$, we have

$$\limsup_{n\to\infty}\beta_n\langle z_n-x^*,d_{n-1}\rangle\le\limsup_{n\to\infty}\beta_n\|z_n-x^*\|\|d_{n-1}\|\le 0.$$

These facts and the condition (i) guarantee that, for any $\varepsilon>0$, there exists $m_0\in\mathbb{N}$ such that

$$\frac{\mu^2\alpha_n}{\tau}\langle d_n,-\nabla f(x^*)\rangle\le\frac{\varepsilon}{6},\ \frac{\mu}{\tau}\langle x^*-x_n,\nabla f(x^*)\rangle\le\frac{\varepsilon}{6}\ \text{and}\ \frac{\mu\beta_n}{\tau}\langle z_n-x^*,d_{n-1}\rangle\le\frac{\varepsilon}{6} \tag{4.5}$$

for all $n\ge m_0$.

Finally, we shall prove that $(x_n)_{n\in\mathbb{N}}$ converges strongly to the solution x^*. It holds from (3.1), (3.2), the nonexpansivity of T, and Lemma 2.1 that

$$\begin{aligned}
\|x_{n+1}-x^*\|^2 &= \|T(x_n+\mu\alpha_n d_n)-T(x^*)\|^2\\
&\le \|x_n+\mu\alpha_n(-\nabla f(x_n)+\beta_n d_{n-1})-x^*\|^2\\
&= \|(x_n-\mu\alpha_n\nabla f(x_n))-(x^*-\mu\alpha_n\nabla f(x^*))\\
&\qquad +\mu\alpha_n\beta_n d_{n-1}-\mu\alpha_n\nabla f(x^*)\|^2\\
&\le \|(x_n-\mu\alpha_n\nabla f(x_n))-(x^*-\mu\alpha_n\nabla f(x^*))\|^2\\
&\qquad +2\alpha_n\langle (x_n+\mu\alpha_n d_n)-x^*, \mu\beta_n d_{n-1}-\mu\nabla f(x^*)\rangle\\
&\le (1-\tau\alpha_n)\|x_n-x^*\|^2+2\tau\alpha_n\left\{\frac{\mu\beta_n}{\tau}\langle z_n-x^*, d_{n-1}\rangle\right.\\
&\qquad \left.+\frac{\mu}{\tau}\langle x^*-x_n,\nabla f(x^*)\rangle+\frac{\mu^2\alpha_n}{\tau}\langle d_n,-\nabla f(x^*)\rangle\right\}
\end{aligned}$$

for every $n\ge 2$. So, by (4.5), we have

$$\|x_{n+1}-x^*\|^2\le(1-\tau\alpha_n)\|x_n-x^*\|^2+\varepsilon\tau\alpha_n$$

for all $n\ge m_0$. By induction, we get

$$\|x_{n+1}-x^*\|^2\le\prod_{k=m_0}^{n}(1-\tau\alpha_k)\|x_{m_0}-x^*\|^2+\varepsilon\left(1-\prod_{k=m_0}^{n}(1-\tau\alpha_k)\right)$$

for all $n\ge m_0$. Thus, it follows from the condition (ii) that

$$\limsup_{n\to\infty}\|x_{n+1}-x^*\|^2\le\varepsilon.$$

Since $\varepsilon>0$ is arbitrary, we obtain $\limsup_{n\to\infty}\|x_{n+1}-x^*\|^2\le 0$ and hence $\lim_{n\to\infty}\|x_{n+1}-x^*\|^2=0$. This implies that the sequence $(x_n)_{n\in\mathbb{N}}$ converges strongly to the uniquely existing solution x^*. □

Remark 4.2. In convergence analysis [1, 7, 8, 14, 31, 46] for the conjugate gradient methods for unconstrained optimization problems, it is commonly assumed that the step size α_n satisfies the following well-known *Wolfe conditions* [39, 40]: $f(x_n)-f(x_n+\alpha_n d_n)\ge-\delta_1\alpha_n\langle d_n,\nabla f(x_n)\rangle$, $\langle d_n,\nabla f(x_n+\alpha_n d_n)\rangle\ge\delta_2\langle d_n,\nabla f(x_n)\rangle$, where $0<\delta_1<\delta_2<1$. In Theorem 4.1 on Algorithm 3.4 for Problem 3.1, the conditions (i)–(v) are assumed.

5. Numerical examples. In order to demonstrate the effectiveness, performance, and convergence of Algorithm 3.4, we first discuss Problem 3.1 for Example 3.2, where $Q\in\mathbb{R}^{64\times 64}$ is a positive definite matrix which is given in [18] and $b:=(-1,-2,\dots,-64)^T\in\mathbb{R}^{64}$.

PROBLEM 5.1. *Consider the following optimization problem:*

$$\textit{minimize } f(x):=\frac{1}{2}\langle x,Q(x)\rangle+\langle b,x\rangle \textit{ subject to } x\in C:=C_1\cap C_2,$$

where $C_1:=\{x\in\mathbb{R}^{64}\colon\|x\|^2\le 1\}$ *and* $C_2:=\{x\in\mathbb{R}^{64}\colon\|x-(1,1,0,\dots,0)^T\|^2\le 1\}$.

The matrix Q has $\lambda_{\min}\approx 8.5539\times 10^{-2}$ and $\lambda_{\max}\approx 9.9251\times 10$. The global minimizer $-Q^{-1}(b)\approx(24.4697,-23.8379,\dots,620.9782)^T\in\mathbb{R}^{64}$ of f is not in C, and

the exact solution of Problem 5.1 cannot be described because of the definitions of Q [18] and $C := C_1 \cap C_2$. We use $x_1 := (-0.5, -0.5, \ldots, -0.5)^T \in \mathbb{R}^{64}$, $\alpha_n := 1/\sqrt{n+1}$, and $\beta_{n+1} := 1/(n+1)^a$ $(n \in \mathbb{N},\ a > 0)$. From the definition of C_i $(i = 1, 2)$, we note that the exact computation of $P_{C_1 \cap C_2}$ is not necessarily easy, while the computations of P_{C_1} and P_{C_2} are tractable. So, we define a mapping $T\colon\ \mathbb{R}^{64} \to \mathbb{R}^{64}$ by $T(x) := (1/2)P_{C_1}(x) + (1/2)P_{C_2}(x)$ for all $x \in \mathbb{R}^{64}$. Then T is a nonexpansive mapping with $\mathrm{Fix}(T) = C := C_1 \cap C_2 \neq \emptyset$ and the computation of T is not hard [42].

Algorithm 3.4 requires us to choose $\mu \in (0, 2\lambda_{\min}/\lambda_{\max}^2)$ by Theorem 4.1. Even if $\mu \geq 2\lambda_{\min}/\lambda_{\max}^2$, by $\alpha_n := 1/\sqrt{n+1}$ $(n \in \mathbb{N})$, there exists $k_1 \in \mathbb{N}$ such that $\mu\alpha_n < 2\lambda_{\min}/\lambda_{\max}^2$ for all $n \geq k_1$. Moreover, Theorem 4.1 guarantees that Algorithm 3.4 with an initial point $x_{k_1} \in \mathbb{R}^{64}$ converges to the solution of Problem 5.1. Since the approximations of $\lambda_{\min}$ and $\lambda_{\max}$ are given, we have $2\lambda_{\min}/\lambda_{\max}^2 \approx 1.7367 \times 10^{-5}$, and hence $\mu := 10^{-5}$. Then the required iterations of Algorithm 3.4 versus the hybrid steepest descent method (HSDM) [42, 43, 44] are presented in Figure 1. Figure 1 shows that the values $\|x_n - T(x_n)\|$ $(n \in \mathbb{N})$ generated by Algorithm 3.4 when $a = 0.1, 0.01, 0.001$ and HSDM converge to 0; that is, the proposed method and HSDM converge to a point in $\mathrm{Fix}(T) = C$. At the same time, it can be observed from Figure 1 that, as compared with HSDM, Algorithm 3.4 succeeds in reducing the value of f, and Algorithm 3.4 has faster convergence than HSDM. It is also seen from Figure 1 that HSDM and Algorithm 3.4 when $a = 0.001$ converge to the same point. When a sequence $(\beta_{n+1})_{n\in\mathbb{N}}$ is a very slowly diminishing constant sequence $(a = 0.01, 0.001)$, Algorithm 3.4 has fast convergence. Figure 2 shows the behavior of $\|x_n - T(x_n)\|$ and $f(x_n)$ $(n = 1, 2, \ldots, 1000)$ for the methods when $\mu := 1$. It is seen from Figure 2 that by a large μ, $f(x_n)$ $(n = 1, 2, \ldots, 10)$ is severely decreasing, and $f(x_n)$ and $\|x_n - T(x_n)\|$ $(n = 11, 12, \ldots, 1000)$ stay about the same. The case when $\mu := 10^{-2}$ is presented in Figure 3. Figure 3 shows that, as compared with Figure 1 $(\mu := 10^{-5})$, HSDM and Algorithm 3.4 when $a = 0.1$ have faster convergence than the methods when $a = 0.01, 0.001$. By the discussions above, if we use $\mu \in (0, 2\lambda_{\min}/\lambda_{\max}^2)$ (see Lemma 2.1 and Theorem 4.1), then Algorithm 3.4 with a small $a > 0$ has fast convergence.

Next, we consider the case where Q is the identity matrix E and $b := (-1, -2, \ldots, -64)^T \in \mathbb{R}^{64}$: Minimize $f(x) := (1/2)\|x\|^2 + \langle b, x\rangle$ subject to $x \in C := C_1 \cap C_2$. In this case, the global minimizer of f is $-b = (1, 2, \ldots, 64)^T \notin C$. We use $\mu := 10^{-3}$ and suppose that $(\alpha_n)_{n\in\mathbb{N}}$ and $(\beta_{n+1})_{n\in\mathbb{N}}$ are the same as in the discussion above. Convergence to the solution in Problem 5.1 and the required iterations of the methods are provided in Figure 4. It can be observed from Figure 4 that HSDM has faster convergence than the proposed algorithm. In particular, for a small $a > 0$, the required iterations are large. In general, a use of the conjugate gradient method for the unconstrained optimization problem is the most effective when the ratio of the minimum and maximum eigenvalues of Q is small. In the case where $Q = E$, that is, the ratio $\lambda_{\min}/\lambda_{\max} = 1$, HSDM is very effective as compared with other methods. From Figure 4, we note that HSDM and Algorithm 3.4 when $a = 0.01$ converge to the same point.

Finally, we discuss the case where Q is a diagonal matrix which has eigenvalues $1, 2, \ldots, 64$ and $b := (0, 0, \ldots, 0)^T \in \mathbb{R}^{64}$ and consider the following: Minimize $f(x) := (1/2)\langle x, Q(x)\rangle$ subject to $x \in C := C_3 \cap C_4$, where $C_3 := \{x \in \mathbb{R}^{64}\colon \|x\|^2 \leq 4\}$ and $C_4 := \{x \in \mathbb{R}^{64}\colon \|x - (2, 0, 0, \ldots, 0)^T\|^2 \leq 1\}$. In this case, $x^* := (1, 0, 0, \ldots, 0)^T \in \{x \in C\colon f(x) = \min_{y\in C} f(y)\}$. Define $T(x) := (1/2)P_{C_3}(x) + (1/2)P_{C_4}(x)$ $(x \in \mathbb{R}^{64})$. We use $\mu := 10^{-4} < 2\lambda_{\min}/\lambda_{\max}^2$ and suppose that $(\alpha_n)_{n\in\mathbb{N}}$ and $(\beta_{n+1})_{n\in\mathbb{N}}$ are the

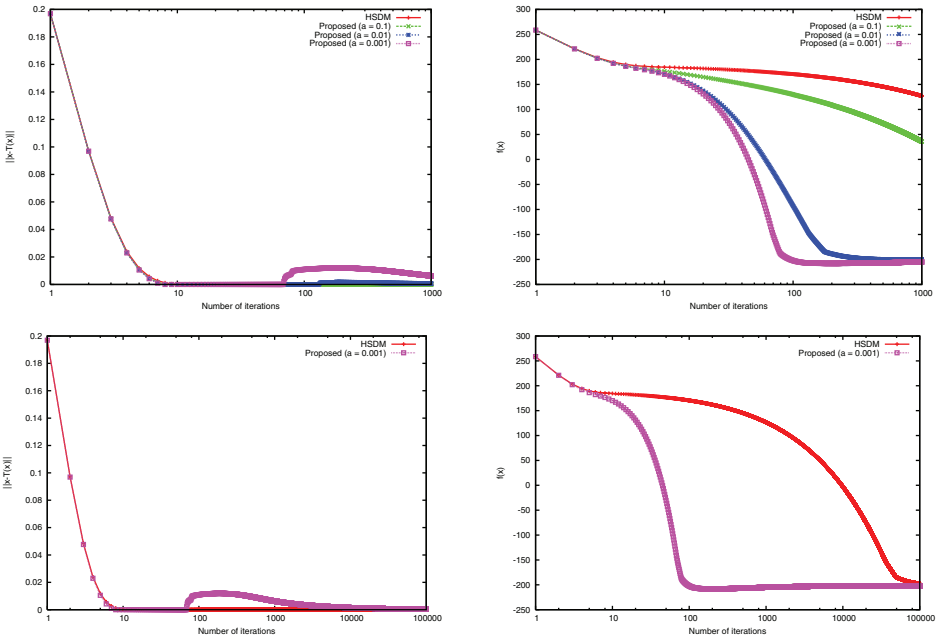


FIG. 1. *Convergence for the methods to the solution when Q is a positive definite matrix which is given in* [18], $\beta_{n+1} := 1/(n+1)^a$ $(a = 0, 1, 0.01, 0.001)$, *and* $\mu := 10^{-5}$.

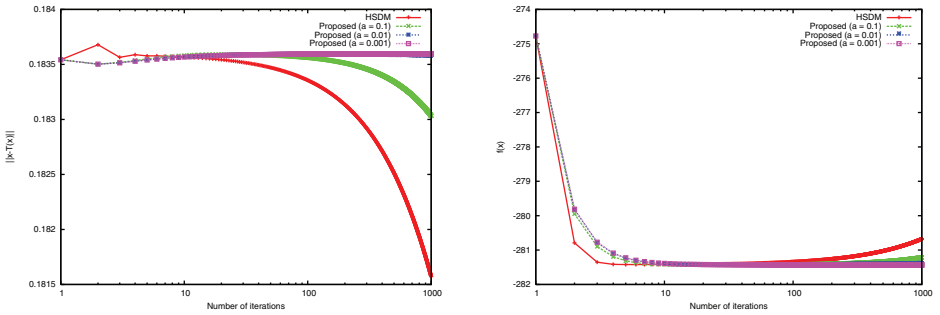


FIG. 2. *Behavior of the methods when Q is a positive definite matrix which is given in* [18], $\beta_{n+1} := 1/(n+1)^a$ $(a = 0, 1, 0.01, 0.001)$, *and* $\mu := 1$.

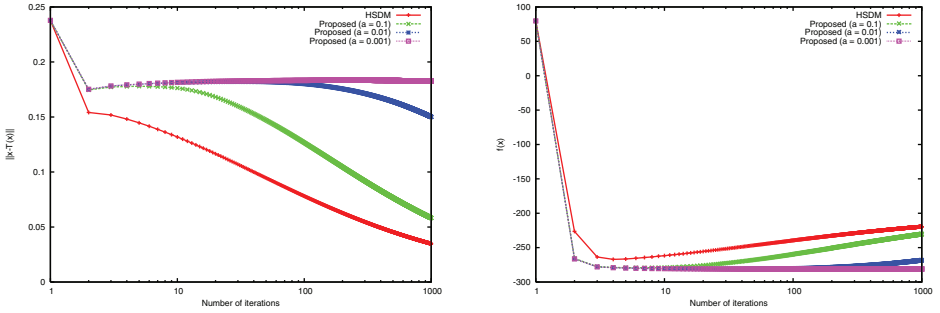


FIG. 3. *Behavior of the methods when Q is a positive definite matrix which is given in* [18], $\beta_{n+1} := 1/(n+1)^a$ $(a = 0, 1, 0.01, 0.001)$, *and* $\mu := 10^{-2}$.

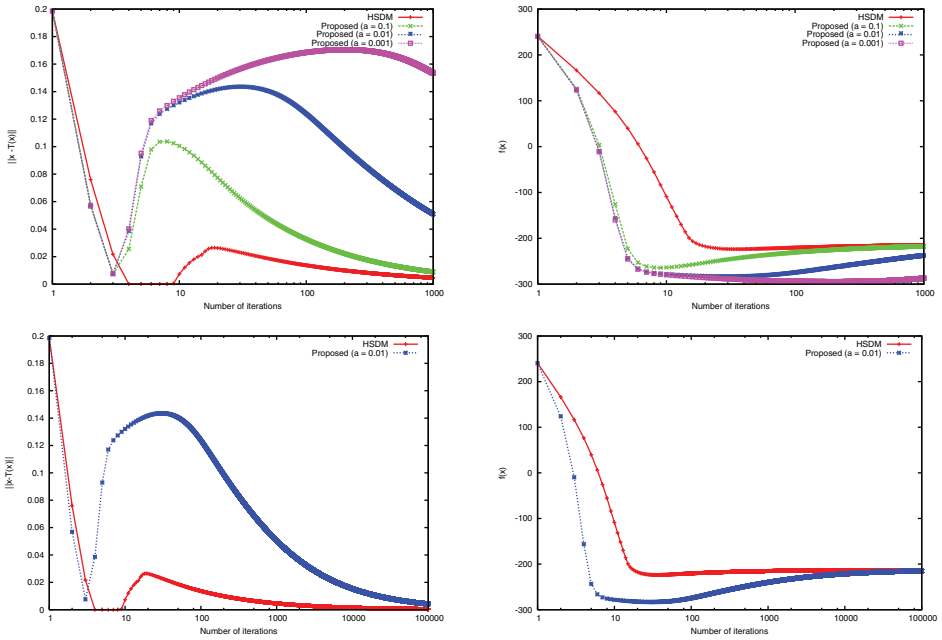


FIG. 4. *Convergence for the methods to the solution when Q is the identity matrix, $\beta_{n+1} := 1/(n+1)^a$ $(a = 0, 1, 0.01, 0.001)$, and $\mu := 10^{-3}$.*

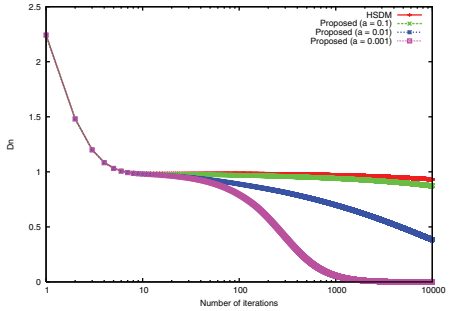


FIG. 5. *Behavior of $D_n := \|x_n - x^*\|^2$ $(n \in \mathbb{N}$, $\{x^*\} = \arg\min_{x \in \mathrm{Fix}(T)} f(x))$ for the methods when Q is a diagonal matrix which has eigenvalues $1, 2, \ldots, 64$, $\beta_{n+1} := 1/(n+1)^a$ $(a = 0, 1, 0.01, 0.001)$, and $\mu := 10^{-4}$.*

same as in the first discussion. The required iterations of the proposed method versus HSDM and the behavior of $D_n := \|x_n - x^*\|^2$ $(n \in \mathbb{N})$ are presented in Figure 5. From Figure 5, our method has faster convergence than HSDM and, in particular, Algorithm 3.4 when $a = 0.001$ succeeds in reducing severely the value D_n $(n \in \mathbb{N})$.

Acknowledgments. The authors would like to thank the anonymous reviewers who helped us improve the original manuscript.

REFERENCES

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

[2] H. H. Bauschke and J. M. Borwein, *On projection algorithms for solving convex feasibility problems*, SIAM Rev., 38 (1996), pp. 367–426.

[3] H. H. Bauschke, J. M. Borwein, and A. S. Lewis, *The method of cyclic projections for closed convex sets in Hilbert space*, Contemp. Math., 204 (1997), pp. 1–38.

[4] J. M. Borwein and A. S. Lewis, *Convex Analysis and Nonlinear Optimization: Theory and Examples*, Springer, New York, 2000.

[5] P. L. Combettes, *A block-iterative surrogate constraint splitting method for quadratic signal recovery*, IEEE Trans. Signal Process., 51 (2003), pp. 1771–1782.

[6] E. E. Cragg and A. V. Levy, *Study on a supermemory gradient method for the minimization of functions*, J. Optim. Theory Appl., 4 (1969), pp. 191–205.

[7] Y. H. Dai and Y. Yuan, *Convergence properties of the Fletcher-Reeves method*, IMA J. Numer. Anal., 16 (1996), pp. 155–164.

[8] Y. H. Dai and Y. Yuan, *A nonlinear conjugate gradient method with a strong global convergence property*, SIAM J. Optim., 10 (1999), pp. 177–182.

[9] F. Deutsch, *Best Approximation in Inner Product Spaces*, Springer, New York, 2001.

[10] I. Ekeland and R. Tĕmam, *Convex Analysis and Variational Problems*, Classics Appl. Math. 28, SIAM, Philadelphia, 1999.

[11] F. Facchinei and J. S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems* I, Springer, New York, 2003.

[12] F. Facchinei and J. S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems* II, Springer, New York, 2003.

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

[14] J. C. Gilbert and J. Nocedal, *Global convergence properties of conjugate gradient methods for optimization*, SIAM J. Optim., 2 (1992), pp. 21–42.

[15] K. Goebel and W. A. Kirk, *Topics in Metric Fixed Point Theory*, Cambridge Stud. Adv. Math. 28, Cambridge University Press, Cambridge, 1990.

[16] K. Goebel and S. Reich, *Uniform Convexity, Hyperbolic Geometry, and Nonexpansive Mappings*, Dekker, New York and Basel, 1984.

[17] A. A. Goldstein, *Convex programming in Hilbert space*, Bull. Amer. Math. Soc., 70 (1964), pp. 709–710.

[18] P. C. Hansen, *Regularization Tools Version* 4.0 *for Matlab Version* 7.3, Numer. Algorithms, 46 (2007), pp. 189–194.

[19] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 409–436.

[20] J. B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* I, Springer, New York, 1993.

[21] J. B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* II, Springer, New York, 1993.

[22] S. A. Hirstoaga, *Iterative selection methods for common fixed point problems*, J. Math. Anal. Appl., 324 (2006), pp. 1020–1035.

[23] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1980.

[24] Z. Q. Luo, J. S. Pang, and D. Ralph, *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press, New York, 1996.

[25] A. Miele and J. W. Cantrell, *Study on a memory gradient method for the minimization of functions*, J. Optim. Theory Appl., 3 (1969), pp. 459–470.

[26] Y. Narushima and H. Yabe, *Global convergence of a memory gradient method for unconstrained optimization*, Comput. Optim. Appl., 35 (2006), pp. 325–346.

[27] J. L. Nazareth, *A conjugate direction algorithm without line searches*, J. Optim. Theory Appl., 23 (1977), pp. 373–387.

[28] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Ser. Oper. Res., Springer, New York, 1999.

[29] E. Polak and G. Ribière, *Note sur la convergence de directions conjugées*, Rev. Francaise Informat Recherche Opertionelle, 3e année, 16 (1969), pp. 35–43.

[30] B. T. Polyak, *The conjugate gradient method in extremal problems*, USSR Comp. Math. Math. Phys., 9 (1969), pp. 94–112.

[31] M. J. D. Powell, *Nonconvex minimization calculations and the conjugate gradient method*, in Numerical Analysis, Lecture Notes in Math. 1066, Springer, Berlin, 1984, pp. 122–141.

[32] S. Reich, *Some problems and results in fixed point theory*, Contemp. Math., 21 (1983), pp. 179–187.
[33] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[34] K. Slavakis, I. Yamada, and K. Sakaniwa, *Computation of symmetric positive definite Toeplitz matrices by the hybrid steepest descent method*, Signal Process., 83 (2003), pp. 1135–1140.
[35] K. Slavakis and I. Yamada, *Robust wideband beamforming by the hybrid steepest descent method*, IEEE Trans. Signal Process., 55 (2007), pp. 4511–4522.
[36] G. Stampacchia, *Formes bilinéaires coercitives sur les ensembles convexes*, C. R. Acad. Sci. Paris, 258 (1964), pp. 4413–4416.
[37] H. Stark and Y. Yang, *Vector Space Projections: A Numerical Approach to Signal and Image Processing, Neural Nets, and Optics*, John Wiley & Sons, 1998.
[38] W. Takahashi, *Nonlinear Functional Analysis*, Yokohama, Yokohama, 2000.
[39] P. Wolfe, *Convergence conditions for ascent methods*, SIAM Rev., 11 (1969), pp. 226–235.
[40] P. Wolfe, *Convergence conditions for ascent methods*. II: *Some corrections*, SIAM Rev., 13 (1971), pp. 185–188.
[41] P. Wolfe, *Finding the nearest point in a polytope*, Math. Program., 11 (1976), pp. 128–149.
[42] I. Yamada, *The hybrid steepest descent method for the variational inequality problem over the intersection of fixed point sets of nonexpansive mappings*, in Inherently Parallel Algorithms for Feasibility and Optimization and Their Applications, D. Butnariu, Y. Censor, and S. Reich, eds., Elsevier, New York, 2001, pp. 473–504.
[43] I. Yamada and N. Ogura, *Hybrid steepest descent method for variational inequality problem over the fixed point set of certain quasi-nonexpansive mapping*, Numer. Funct. Anal. Optim., 25 (2004), pp. 619–655.
[44] I. Yamada, N. Ogura, and N. Shirakawa, *A numerical robust hybrid steepest descent method for the convexly constrained generalized inverse problems*, Contemp. Math., 313 (2002), pp. 269–305.
[45] E. Zeidler, *Nonlinear Functional Analysis ans Its Applications* III. *Variational Methods and Optimization*, Springer, New York, 1985.
[46] G. Zoutendijk, *Nonlinear programming*, in Integer and Nonlinear Programming, J. Abadie, ed., North–Holland, Amsterdam, 1970, pp. 37–86.

SIAM J. OPTIM.
Vol. 19, No. 4, pp. 1894–1917

AN INVITATION TO TAME OPTIMIZATION*

A. D. IOFFE†

Abstract. The word "tame" is used in the title in the same context as in expressions like "convex optimization," "nonsmooth optimization," etc.—as a reference to the class of objects involved in the formulation of optimization problems. Definable and tame functions and mappings associated with various o-minimal structures (e.g. semilinear, semialgebraic, globally subanalytic, and others) have a number of remarkable properties which make them an attractive domain for various applications. This relates both to the power of results that can be obtained and the power of available analytic techniques. The paper surveys certain ideas and recent results, some new, which have been or (hopefully) can be productively used in studies relating to variational analysis and nonsmooth optimization.

Key words. variational analysis, optimization, o-minimal structure, definable sets, functions and mappings, critical value, optimal control, semismooth mapping

AMS subject classifications. 26E99, 49J53, 58K05, 32B20, 49K05, 26D10, 65K10

DOI. 10.1137/080722059

1. Introduction.

The progress in variational analysis in the last quarter of the 20th century resulted in a complete reshaping of some chapters of analysis and optimization theory. I first of all mean the nonsmooth analysis per se, the theory of nonconvex subdifferentials that extended all rules of differential and convex subdifferential calculus as far as to lower semicontinuous functions and set-valued mappings.

However, it was recognized from the outset that the classes of all lower semicontinuous and even all (locally) Lipschitz functions are often unnecessarily broad and the general theory does not give any answer for which functions and sets one or another result indeed gives useful information. As a most striking example, I shall mention that the limiting subdifferential of a generic Lipschitz function with Lipschitz constant one is identically equal to the unit ball [6], so for a typical Lipschitz function no useful information of its behavior can be obtained from its subdifferential mapping.

Therefore, a search for suitable classes of functions that can be effectively used in one or another context has been one of the leading themes in variational analysis since the very beginning. It is enough to mention semismooth functions [12], prox-regular functions [28], composite and amenable functions [28], etc. More recently, an understanding that sets, functions, and mappings that usually appear in applications have some distinctive structural features which, taken into consideration, may substantially facilitate analysis has been gradually becoming a driving idea. The successful use of convex semialgebraic sets associated with linear matrix inequalities in numerical convex optimization [26] is probably the most impressive demonstration of this trend. Recent concepts of active sets, $\mathcal{U}$-Lagrangians, and some others [24, 22] reviving at a very new technical level an old heuristic idea of the "ravine method" of Gelfand and Zetlin [13] is another important example.

The results presented in the paper suggest that the classes of tame and, especially, definable functions may offer an appropriate domain on which the machinery of variational analysis works with full efficiency. The boundaries of the class have

*Received by the editors April 23, 2008; accepted for publication (in revised form) October 29, 2008; published electronically February 27, 2009.
http://www.siam.org/journals/siopt/19-4/72205.html
†Department of Mathematica, Technion, Haifa 32000, Israel (ioffe@math.technion.ac.il).

not been yet fully established, but what is already known shows that it encompasses an enormous variety of functions, sets, and mappings. Functions of this class are uniquely defined by their subdifferential mappings (up to an additive constant for each connected component of the graph). In a nutshell they can be characterized as "pathology free" objects. The meaning of this expression will be gradually clarified in the course of our discussions. We shall just give here a few illuminating examples.

A tame function need not be continuous. But the discontinuities it may have can only be of the "first" type with unilateral limits (perhaps infinite) along tame bounded curves always existing. A tame set need not be a manifold, but it can always be decomposed into a locally finite collection of manifolds which fit each other in a nice way. A tame set-valued mapping has tame selections without any additional topological assumptions. And, what is especially valuable—the class is closed with respect to practically all operations used in variational analysis: Boolean, topological, linear, structural, and most analytical (as composition or differentiation of mappings), including, e.g., partial minimization—a feature especially important for optimization. Finally, elements of the class have distinctive and recognizable structural properties which tremendously facilitate working with them.

Definable and tame sets, functions, and mappings are a product of model theory and algebraic geometry; they are the main concepts of the theory of so-called "o-minimal structures" that has been actively developing during last 20–25 years [8, 10]. Many results of the theory cannot be proved by means of analysis. This, however, is not a real obstacle for using them which is often (if not typically) simple and extremely convenient. The paper is a sort of a survey of some very recent results relating to optimization theory and heavily relying on o-minimality. However, it also contains some new results and proofs.

To the best of my knowledge, the first study in which o-minimality was directly applied to optimization is the 2002 paper by Graña Drammond and Peterzil [14]. They showed that under certain assumptions (of which analyticity of the data was the key, opening the possibility to apply o-minimality) central path trajectories in problems of semidefinite programming converge. We shall briefly describe their result in the next section in which we give definitions of o-minimal structures and tame and definable objects (sets, functions, and mappings) they generate and talk about them in some detail. For more substantial information we can refer to [8, 10, 32]. In the third section we show that the main maps of variational analysis, such as various subdifferentials, slopes, and moduli of metric regularity and surjection (as functions of a point of the graph) are definable or tame, provided so is the original mapping. These two sections can be viewed as an extended introduction. The main content of the paper is sections 4–10.

In the fourth section we present an extension of Sard's theorem to (as far as) tame set-valued mappings, recently proved in [18]. Sard (or Morse–Sard) theorem is among the very central (and technically difficult) results of analysis playing an important part in differential topology, dynamical systems, and singularity theory. It says that the set of critical values of a sufficiently smooth mapping between finite dimensional spaces has Lebesgue measure zero. It might have valuable applications also in optimization theory (especially in stability and sensitivity analysis). There were indeed a few (see, e.g., [29]), but, basically, this has not happened, mainly because of the strong smoothness requirements of the Sard theorem which are known to be sharp. It turns out that, with a natural extension of the concept of critical value based on the metric regularity/openness-at-a-linear-rate dichotomy of variational analysis, the extension of Sard theorem to tame set-valued mappings is possible. In the fifth section we

apply the Sard theorem of section 4 to get some good properties of mathematical programming problems with definable or tame data (normality of the feasible set for a typical right-hand side, finiteness of the number of critical values in a normal problem).

In the sixth section we prove a set-valued extension of the famous Łojasiewicz inequality describing a behavior of a real analytic function near a critical point. In 1998 Kurdyka extended Łojasiewicz's theorem to continuously differentiable definable functions [21] and recently a further extension to lower semicontinuous definable functions in terms of Clarke's generalized gradients was obtained by Bolte-Daniilidis-Lewis in [4]. The non-smooth extension of the Sard theorem plays a central part in our proof and actually leads to its noticeable simplification.

In the seventh section we apply the result of section 6 to get a nonsmooth extension of another famous theorem of Łojasiewicz [23] saying that bounded trajectories of gradient descent of a real analytic function have bounded length. Earlier extensions of this theorem were obtained in [21] for C^1 definable functions and in [3] for definable lower C^2 functions. We consider arbitrary lower semicontinuous definable functions and prove that bounded curves of maximal slope (a concept introduced in 1980 by DeGiorgi–Marino–Tosques [9] as an appropriate extension of trajectories of gradient descent) have bounded length. Again an application of the set-valued Sard theorem is an important element of the argument.

Then in the eighth section we give a simplified proof of a recent result of [5] about semismoothness of locally Lipschitz definable functions. Recall that semismoothness is the property that guarantees superlinear convergence of the Newton method. The problem with this concept was that no sufficiently universal sufficient criteria were known that made the problem of recognizing semismoothness rather complicated. The theorem of [5] basically fills the gap.

In the ninth section we prove a certain "definable" extension of Lyapunov's theorem on vector measures. Namely, it is shown that the integral of a set-valued mapping (that is the collection of integrals of its summable selections) coincides with the "definable" integral (the collection of integrals of tame selections), provided the set-valued mapping is tame.

Finally, in the tenth section we apply the obtained definable extension of Lyapunov's theorem to a class of optimal control problems which contains in particular optimal control of systems guided by linear (nonstationary) equations. This class of problems plays also an important part in mathematical economics (e.g., the famous Aumann–Perles problem [1] belongs to this class). The main result we present here shows that if the data of the problem (e.g., the integrand) are definable, then the problem contains a piecewise continuous optimal control (a very desirable property), provided a certain optimal control exists.

2. Definable and tame functions and sets. There is a sort of hierarchy of classes of tame sets, functions, and maps. In each case it is usually sufficient to describe the collection of sets and then define a corresponding class of functions and mappings by their graphs being elements of the class of sets. The simplest class is formed by so-called semilinear sets (functions, mappings). A semilinear set in $\mathbb{R}^n$ is defined as a finite union of open polyhedra. An open polyhedron is the intersection of a finite number of affine sets and open half-spaces:

$$P = \{x \in \mathbb{R}^n : \ a_i \cdot x = \alpha_i, \ i = 1, \dots, k; \ \langle a_i, x\rangle < \alpha_i, \ i = k+1, \dots, m\}.$$

The dimension of (a nonempty polyhedron) P is $n - \operatorname{rank}(a_1, \dots, a_k)$. A closed polyhedron is clearly a semilinear set (union of all its open faces), so a closed semilinear

set can be defined as a finite union of closed polyhedra. One can easily check that the collection of semilinear sets in $\mathbb{R}^n$ is closed under Boolean operations (union, intersection, complement) and the image and preimage of a semilinear set under a linear (or affine) mapping is also a semilinear set.

A function or a map is called semilinear (sometimes piecewise linear) if its graph is a semilinear set. Another easily verifiable property of a semilinear mapping is that its domain can be decomposed in a finite union of polyhedra such that the restriction of the mapping to each of them is an affine mapping. The latter can be viewed as another definition of semilinear mappings and functions.

The class of semilinear sets, functions, and mappings is of course sufficiently narrow. A much richer collection of tame objects is provided by *semialgebraic* sets, mappings, and functions. A set in $\mathbb{R}^n$ is semialgebraic if it is a finite union of sets of the form

$$\{x \in \mathbb{R}^n : \ p_i(x) = 0, \ i = 1, \dots, k; \ p_i(x) < 0, \ i = k+1, \dots, m\}, \tag{2.1}$$

where all p_i are polynomials. An important for optimization class of convex semialgebraic sets are sets of solutions of a linear-matrix inequalities

$$\sum_{i=1}^{n} x_i A_i \succcurlyeq 0,$$

where A_i are square (symmetric) matrices and $\succcurlyeq$ means positive semidefiniteness. Semialgebraic functions and sets, as their semilinear counterparts, enjoy a number of remarkable properties:

(i) *the class of semialgebraic sets is closed with respect to Boolean operators; a Cartesian product of semialgebraic sets is a semialgebraic set;*

(ii) *the closure and the interior of a semialgebraic set is a semialgebraic set;*

(iii) *semialgebraic subsets of* $\mathbb{R}$ *are precisely finite unions of points and open intervals (hence the same as semilinear);*

(iv) *the projection of a semialgebraic subset of* $\mathbb{R}^n$ *onto* $\mathbb{R}^{n-1}$*, e.g.,*

$$(x_1, \dots, x_n) \to (x_1, \dots, x_{n-1})$$

is a semialgebraic set.

The last of these statements, unlike in the semilinear case, is a deep fact known as the Tarski–Seidenberg theorem. The proofs of the other three are much simpler (although the closedness property in (ii) is a consequence of Tarski–Seidenberg theorem).

The key question behind the further developments is whether there are broader classes of functions and sets for which these (and some other related properties) hold. It turns out that the answer is positive. It is given in terms of so-called *o-minimal structures.*[1]

The latter is a system $\mathcal{S} = \{\mathcal{S}_n\}$, with $\mathcal{S}_n$ being a collection of subsets of $\mathbb{R}^n$ containing all semialgebraic subsets, satisfying (i)–(iv) if we replace there "semialgebraic" by "elements of $\mathcal{S}$." The o-minimal structure of semialgebraic sets is usually

[1] To be more precise we talk here about what is formally defined as an o-minimal structure over $(\mathbb{R}, +, \cdot)$. This means that we consider the subsets of $\mathbb{R}^n$ and make full use of the ring structure of reals. O-minimal structures over other rings can also be considered. On the other hand, semilinear sets are an o-minimal structure which makes no use of the multiplicative structure of reals.

denoted $\mathcal{S}_{alg}$. If we add level and sublevel sets of restrictions of analytic functions to balls, we get the structure $\mathcal{S}_{an}$ of so-called globally subanalytic sets, etc. We refer to [8, 10] for details.

If we have an o-minimal structure, its elements are called *definable sets* (with a reference to the o-minimal structure if necessary). Functions and mappings whose graphs are definable sets are also called definable.

Given an o-minimal structure, we can be sure that

(a) *if $f(x,y)$ is a definable function, then so is $\varphi(x) = \inf_y f(x,y)$;*

(b) *a composition of definable mappings is a definable mapping;*

(c) *the image and preimage of a definable set under a definable mapping are definable sets.*

The three properties we have just stated heavily rely on (iv), the definable counterpart of the Tarski–Seidenberg theorem. To better understand ramifications of the latter, we notice that as an immediate implication of (iv), we get definability of any set $\{x \in P : \exists y \in Q,\ (x,y) \in S\}$, provided P, Q, and S are definable sets in the corresponding spaces. It follows that also $\{x \in P : \forall y \in Q,\ (x,y) \in S\}$ is a definable set as its complement is the union of the complement of P and the set $\{x \in P : \exists y \in Q,\ (x,y) \notin S\}$. Thus, if we have a finite collection of definable sets, then any set obtained from them with the help of a finite chain of quantifiers is also definable.

A consequence of (a)–(c) very useful in any optimization context is that

(d) *if Q is a definable set, then the distance function $d(x,Q)$ is also definable.*

Indeed, the function $\varphi(x,u) = \|u - x\|$ is semialgebraic, and hence definable as well as $\delta_Q(x)$, the indicator of Q; hence $d(x,Q) = \inf_u(\|x-u\| + \delta_Q(u))$ is a definable function by (a),(b).

In case we are interested only in local results and what happens near the infinity is of a limited interest, the following formal definition of tameness is justified: a set is called *tame* (with respect to a certain o-minimal structure) if its intersection with any ball is definable. A function or (set-valued) mapping is *tame* if its graph is tame.

Thus, for instance, the function $\sin t$, not definable in any o-minimal structure (since the nonempty preimage of a point is a countable set), is a tame function with respect to the o-minimal structure $\mathcal{S}_{an}$ of globally subanalytic sets. However, the function $\sin t^{-1}$ is not tame on $(0,1)$.

It is appropriate to note here that *a projection of a tame set may fail to be tame as well as composition of tame mappings* unless the external map is coercive (that is such that preimages of bounded sets are bounded). In particular, the sum of two tame functions may fail to be tame.[2]

Here are some fundamental properties of definable objects.

THEOREM 2.1 (monotonicity theorem). *Let f be a definable function on a real interval (a,b), where $-\infty \le a < b \le \infty$. Then there is a finite number of points $a = t_0 < t_1 < \ldots < t_k < b = t_{k+1}$ such that on every interval (t_i, t_{i+1}) f is continuous and either strictly monotone or constant.*

As an immediate consequence of this theorem we get that a bounded definable trajectory in $\mathbb{R}^n$ defined on an open interval $(0,T)$ converges when, e.g., $t \to 0$. We

[2]An alternative definition void of this inconvenience would be, e.g., a mapping is tame if its restriction to any ball is definable. However, such a definition would also be problematic because in this case (a) the inverse to a tame mapping may not be tame and (b) there would be less tame mappings. Hopefully, future developments will clarify which localization of definability is more convenient, at least for the needs of optimization theory.

shall have a couple of chances to see the power of this seemingly simple result. In particular, this is the key argument in the proof of convergence of the central path in nonlinear semidefinite programming in the quoted paper [14].

We shall briefly explain the idea of the proof in a somewhat more general situation of the problem

$$\text{minimize } f(x), \quad \text{s.t. } g(x,u) \le 0,\ \forall u \in U,$$

where $x \in \mathbb{R}^n$ and $U \subset \mathbb{R}^m$ is a compact set. The functions f and $g(\cdot, u)$ are assumed convex continuous. Continuation methods for solving such problems (under the strict feasibility, or Slater assumption that there is an x such that $\sup_{u\in U} g(x,u) < 0$) consist in solving auxiliary problems

$$\text{minimize } f(x) + \mu\beta(x),$$

where $\beta(x)$ is a barrier function (finite for strictly feasible x and equal to ∞ otherwise) and studying the behavior of its solution $x(\mu)$ as $\mu \to 0$. The barrier function is usually strictly convex, so there may be only one solution $x(\mu)$ for every μ. Theory of convex programming provides conditions that guarantee that $x(\mu)$ (the central path) is well-defined and bounded. The question is when $x(\mu)$ converges as $\mu \to 0$. It turns out that, even for problems of linear programming, the proof of convergence needs methods of algebraic geometry [16]. Monotonicity theorem allows one to easily solve the problem under tameness/definability assumptions. Namely, if *U is a definable set, the functions f and g are tame, and β is definable, all in the same o-minimal structure, then the central path (if it is defined for small μ and bounded) converges.*[3]

THEOREM 2.2 (definable selection theorem). *Let $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ be a set-valued mapping with a definable graph. Let* dom $F = \{x :\ F(x) \neq \emptyset\}$ *stand for the domain of F. Then there is a single-valued definable mapping φ from $\mathbb{R}^n$ into $\mathbb{R}^m$, defined at least on* dom F*, such that $\varphi(t) \in F(t)$ for all $t \in$* dom F.

Again it is worth paying attention to how easily the selection problem (very painful in many situations) is solved in the realm of definable objects.

THEOREM 2.3 (stratification theorem [11]). (a) *Let $Q \in \mathbb{R}^n$ be a definable set. Then Q admits a C^k-Whitney stratification for any k: there is a finite partition*[4] *of Q into C^k manifolds M_i such that*

- *if $M_j \cap \mathrm{cl} M_i \neq \emptyset$, then $M_j \subset \mathrm{cl} M_i \backslash M_i$;*
- *if $x \in M_j$ and $x_k \in M_i$ converge to x as $k \to \infty$, then $T_x M_j$, the tangent space to M_j at x, is contained in the lower limit of $T_{x_k} M_i$.*

(b) *If F is a definable mapping from $\mathbb{R}^n$ into $\mathbb{R}^m$, then for any k there is a C^k-Whitney stratification (M_i) of* dom F *such that the restriction of F to every M_i is k times continuously differentiable.*

This is the most fundamental structural characterization of definable sets. Of course a set which admits even a C^∞ stratification may not be definable. So the important element of the theorem is the possibility to get strata that are both definable and having the desirable order of smoothness.

3. Variational analysis and tameness. Let us consider first an extended-real-valued function f on $\mathbb{R}^n$. The set dom $f = \{x :\ |f(x)| < \infty\}$ is the *domain* of f and the set epi $f = \{(x,\alpha) \in \mathbb{R}^n \times R :\ \alpha \ge f(x)\}$ is the *epigraph* of f. The

[3] In [14] this argument was actually applied to the structure $\mathcal{S}_{\text{an exp}}$ and real analytic f and g.

[4] Note that the standard definition of Whitney stratification involves locally finite partitions.

function is lower semicontinuous if its epigraph is closed. The function $\bar f$ defined by epi $\bar f = \mathrm{cl}(\mathrm{epi}\ f)$ is called the *lower closure* of f. If f is definable (tame), then so is $\bar f$.

Recall that the *proximal* and the *Fréchet* subdifferentials of f at x are defined, respectively, by

$$\partial_p f(x) = \{y :\ \liminf_{\|h\|\to 0} \|h\|^{-2}\big(f(x+h) - f(x) - \langle y, h\rangle\big) > -\infty\},$$

and

$$\partial_F f(x) = \{y :\ \liminf_{\|h\|\to 0} \|h\|^{-1}\big(f(x+h) - f(x) - \langle y, h\rangle\big) \ge 0\}$$

for sufficiently small h.

Along with the two subdifferential set-valued mappings $x \mapsto \partial_p f(x)$ and $x \mapsto \partial_F f(x)$, we can also consider the corresponding *subjets* which are the sets

$$[\partial_p f] = \{(x, \alpha, y) :\ y \in \partial_p f(x),\ \alpha = f(x)\}$$

and

$$[\partial_F f] = \{(x, \alpha, y) :\ y \in \partial_F f(x),\ \alpha = f(x)\}.$$

A remarkable fact is that the closures of the two subjets coincide. The *limiting* subdifferential $\partial f(x)$ is the set-valued mapping defined by the condition: $y \in \partial f(x)$ if and only if $(x, f(x), y)$ belongs to the closure of the just defined subjets. Finally, if f satisfies the Lipschitz condition, then $\partial_c f(x) = \mathrm{conv}\ \partial f(x)$ is the *Clarke* subdifferential (or generalized gradient) of f at x. The limiting and Clarke subjets are defined in the same way as the subjets for the first two subdifferentials.

Observe that for a continuously differentiable function, the Fréchet, limiting, and Clarke subgradients coincide and contain a unique element—the derivative of the function. If, in addition, the function is twice differentiable at the point, then the proximal subdifferential coincides with the other three. Likewise, if f is a convex function, all four subdifferentials coincide and are equal to the subdifferential of the function in the sense of convex analysis. We refer to [28] for the basic facts concerning finite dimensional subdifferential calculi.

Another important characteristic of a local behavior of a function (which actually makes sense even for functions on arbitrary metric spaces) is the *slope* of f at x:

$$|\nabla f|(x) = \limsup_{h\to 0,\ h\ne 0} \|h\|^{-1}(f(x) - f(x+h))^+$$

(where $\alpha^+ = \max\{\alpha, 0\}$) which is the maximal "speed" of decrease of function from x. The connection between slopes and subdifferentials is determined by the following inequalities:

$$\inf\{\|y\| :\ y \in \partial_F f(x)\} \ge |\nabla f|(x) \ge \inf\{\|y\| :\ y \in \partial f(x)\}$$

(see [17] for details). In particular, if f is differentiable at x, then $|\nabla f|(x)$ is equal to the norm of the gradient of f at x; if f is convex, then the slope is equal to the distance from $\partial f(x)$ to zero.

Proposition 3.1. (a) *If F is a definable mapping $\mathbb{R}^n \to \mathbb{R}^m$, then its Fréchet derivative is a definable mapping (into the space of linear operators);*

(b) *If f is a definable function, then so are the four subdifferential mappings and the slope function $|\nabla f|$;*

(c) *If f is a tame function, then each of the four subjets is a tame set.*

(For the case of the generalized gradient we assume that f is locally Lipschitz.[5]*)*

Proof. We first observe the following: let $\varphi(x,h)$ be a definable function of two variables (of different dimensions in general). Then the function

$$\psi(x) = \liminf_{\substack{h\to 0\\ h\neq 0}} \varphi(x,h)$$

is also definable since

$$\psi(x) = \sup_{\varepsilon>0} \inf_{0<\|h\|<\varepsilon} \varphi(x,h).$$

Clearly, the same is true also for

$$\limsup_{\substack{h\to 0\\ h\neq 0}} \varphi(x,h).$$

Now to prove (a) we take

$$\varphi(x,A,h) = \|h\|^{-1}\|F(x+h) - F(x) - Ah\|$$

and notice that the graph of F' is the zero level set of the function $\rho(x,y)$ equal to the upper limit of $\varphi(x,y,h)$ when $0 \neq h \to 0$.

Likewise to prove (b), we take

$$\varphi_r(x,y,h) = \|h\|^{-r}(f(x+h) - f(x) - \langle y,h\rangle)$$

with $r=2$ for proximal subdifferentials and $r=1$ for Fréchet subdifferentials and

$$\varphi(x,h) = \|h\|^{-1}(f(x) - f(x+h))^+$$

for slopes. If now

$$\psi_r(x,y) = \liminf_{\substack{h\to 0\\ h\neq 0}} \varphi_r(x,y,h),$$

then

$$\text{Graph } \partial_p f = \{(x,y):\ \psi_2(x,y) > -\infty\};\ \text{Graph } \partial_F f = \{(x,y):\ \psi_1(x,y) > -\infty\}$$

and, of course,

$$|\nabla f|(x) = \limsup_{\substack{h\to 0\\ h\neq 0}} \varphi(x,h).$$

As long as we know that $\partial_p f$ and ∂f are definable maps, it is an easy matter to show that the corresponding subjets are definable sets. Consider, for instance, the case of the Fréchet subdifferential. Set

$$Q = \{(x,\alpha,z,y):\ \alpha = f(x),\ y \in \partial_F f(z)\} = (\text{Graph } f) \times (\text{Graph } (\partial_F f)).$$

[5] Although the result is valid without the assumption if we use Rockafellar's representation formula for the generalized gradient.

Then $[\partial_F f]$ is the projection of the set $Q \cap \{(x,\alpha,z,y) : \ z = x\}$ to the space of (x,α,y).

Finally, the graph of ∂f is by definition $\{(x,y) : (x,f(x),y) \in \mathrm{cl}[\partial_F f]\}$, that is to say, the preimage of the definable set $\mathrm{cl}[\partial_F f]$ under the definable mapping $(x,y) \to (x,f(x),y)\}$, hence a definable set.

To prove (c), we first consider the case of a bounded f. Then the restriction of f to any open ball is a definable function (as the graph of the restriction is the intersection of the graph of f with a bounded definable set). Hence the subdifferentials and the corresponding subjets of the restriction are definable, so the subjets of the function are tame sets.

Let now f be an arbitrary lower semicontinuous tame function. For any N set

$$B_N = \{(x,\alpha,y) : \ \max\{\|x\|,|\alpha|,\|y\|\} \le N\}, \quad f^K(x) = \min\{K, \max\{f(x),-K\}\}.$$

Then, e.g., $[\partial_F f^{2N}] \cap B_N = [\partial_F f] \cap B_N$ (as is easy to verify) and the same equality holds for the proximal subdifferential. This completes the proof of the proposition. □

Remarks. 1. The derivative or subdifferential of a tame function may not be a tame mapping: consider for instance the function $f(x) = x^{-1} - \sin x^{-1}$ on $(0,\infty)$.

2. We do not need in this paper coderivatives and derivatives of set-valued mappings as well as tangent and normal cones, all playing an important role in variational analysis. Similar properties (definability or tameness) can be also established for them.

We shall next consider the property of metric regularity, one of the central in variational analysis. Recall that F is called *metrically regular* near $(\bar{x},\bar{y}) \in \mathrm{Graph}\ F$ if there is a positive K such that

$$d(x, F^{-1}(y)) \le K d(y, F(x))$$

for all (x,y) sufficiently close to $(\bar{x},\bar{y})$. The lower bound $\mathrm{reg}F(\overline{x}|\overline{y})$ of such K is called *modulus of metric regularity* of F near $(\bar{x},\bar{y})$. F is called *open at a linear rate* near $(\bar{x},\bar{y})$ if there is an $r > 0$ such that

$$B(y,rt) \subset F(B(x,t))$$

if $(x,y) \in \mathrm{Graph}\ F$ are sufficiently close to $(\bar{x},\bar{y})$ and $t > 0$ is sufficiently small. The upper bound $\mathrm{sur}F(\overline{x}|\overline{y})$ of such r is called the *modulus of surjection* of F near $(\bar{x},\bar{y})$. A well-known and important fact is that

$$\mathrm{sur}F(\overline{x}|\overline{y}) \cdot \mathrm{reg}F(\overline{x}|\overline{y}) = 1. \tag{3.1}$$

If we agree to set $0 \cdot \infty = 1$, the property is valid unconditionally. Another useful and rather elementary fact is that

$$\mathrm{sur}(F \circ H)(x|z) \le \|H'(x)\| \mathrm{sur}F(H(x)|z) \tag{3.2}$$

if $H : \mathbb{R}^k \to \mathbb{R}^n$ is continuously differentiable at x.

Finally, we give two formulas that allow to compute the modulus of surjection (metric regularity) for a set-valued mapping $F : \ \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ with closed graph (hence upper semicontinuous). The first (which is in principle valid in arbitrary Banach spaces) is given in terms of slopes:

$$\mathrm{sur}F(\overline{x}|\overline{y}) = \liminf_{\substack{(x,y) \to F\ (\overline{x},\overline{y}) \\ y \notin F(x)}} |\nabla d(y, F(\cdot))|(x) \tag{3.3}$$

(see [17], Theorem 2.2b). Here $(x,y) \to_F (\overline{x},\overline{y})$ means that $x \to \overline{x}$, $y \to \overline{y}$, and $d(y,F(x)) \to 0$.

The second formula, often more convenient for practical calculation, is finite dimensional:

$$\mathrm{sur}F(\overline{x},\overline{y}) = \inf\{\|x^*\| :\ x^* \in D^*F(\bar{x},\bar{y})(y^*),\ \|y^*\| = 1\} \tag{3.4}$$

(see, e.g., [28], Theorem 9.43). Here $D^*F(x,y)$ is the limiting coderivative of F at (x,y).[6]

PROPOSITION 3.2. *If F is definable (tame) set-valued mapping, then so are the functions $(x,y) \to \mathrm{reg}F(x|y)$ and $(x,y) \to \mathrm{sur}F(x,y)$.*

Proof. We shall prove only the "definable" part of the proposition. The "tame" part will then be immediate as the definitions of the moduli deal with bounded portions of the graph. On the other hand, as long as we consider only definable mappings, it is sufficient, according to (3.1), to establish the definability property only for one of the moduli. So we shall talk about modulus of metric regularity of a definable set-valued mapping.

For such a mapping the function $\varphi(x,y) = d(x,F^{-1}(y))$ is definable as its epigraph is the closure of the projection onto the (x,y,α)-space of the set

$$\{(x,y,u,\alpha) :\ \alpha \geq \|x-u\|,\ (u,y) \in \mathrm{Graph}\ F\}.$$

Similar argument shows that also the function $d(y,F(x))$ is definable.

Setting

$$\varphi(u,v) = \begin{cases} \dfrac{d(u,F^{-1}(v))}{d(v,F(u))}, & \text{if } v \notin F(u); \\ 0, & \text{if } v \in F(u) \end{cases}$$

and

$$\psi(x,y,u,v,\varepsilon) = \begin{cases} 0, & \text{if } \|x-u\| < \varepsilon,\ \|y-v\| < \varepsilon, \\ -\infty & \text{otherwise,} \end{cases}$$

we see that

$$\mathrm{reg}F(x|y) = \lim_{\varepsilon \to 0} \sup_{u,v} (\varphi(u,v) + \psi(x,y,u,v,\varepsilon))$$

is a definable function. □

Remark. If F is single valued, we usually write $\mathrm{sur}F(x)$ and $\mathrm{reg}F(x)$ rather than $\mathrm{sur}F(x|F(x))$ and $\mathrm{reg}F(x|F(x))$. The functions $x \to \mathrm{sur}F(x)$ and $x \to \mathrm{reg}F(x)$ are definable if so is F. However if F is only tame, the functions may not be tame: tameness is preserved only by the sets $\{(x,y,\alpha) :\ y = F(x), \alpha = \mathrm{sur}F(x)\}$ and $\{(x,y,\alpha) : y = F(x), \alpha = \mathrm{reg}F(x)\}$.

4. A nonsmooth extension of the Sard theorem. Let $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$, and let $y \in F(x)$. We say that (x,y) is a *critical point* of F if $\mathrm{sur}F(x|y) = 0$. If F is single-valued and continuously differentiable at x, this definition reduces to the

[6] This is the set-valued mapping which with every y^* associates the set of x^* such that $(x^*,-y^*)$ belongs to the limiting normal cone for the graph of F at (x,y). The latter is the limiting subdifferential of the indicator of Graph F (which is the function equal to zero on Graph F and $+\infty$ outside of Graph F). We shall not use the formula in the sequel.

standard concept of a critical (or singular) point: x is critical if $F'(x)$ maps $\mathbb{R}^n$ not to the whole of $\mathbb{R}^m$. We say that $y \in \mathbb{R}^m$ is a *critical value* of F if there is an $x \in \mathbb{R}^n$ such that (x, y) is a critical point of F. The famous Sard theorem (see, e.g., [15]) says that the collection of critical values of a C^k-mapping $F : \mathbb{R}^n \to \mathbb{R}^m$ has Lebesgue measure zero, provided $k > \max\{n - m, 0\}$. The theorem basically says that only for a very meager collection of right-hand side vectors y, the set of solutions of the equation $F(x) = y$ may sharply react to small changes of y. We refer to [32] for applications of the Sard theorem and discussions of some recent developments.

Sard's theorem is a sharp result. The first counterexample of a continuously differentiable function on $\mathbb{R}^2$ whose set of critical values contains an open interval was suggested by Whitney [31] even before the Sard theorem was proved in 1942. The final results due to Bates [2] and Norton [27] say that the conclusion of the theorem is still valid for $C^{n-m,1}$ mappings (having Lipschitz continuous $(n-m)$th derivative) but not for mappings of the Hölder classes $C^{n-m,\alpha}$ with $\alpha < 1$.

The following general result, however, holds true [18].

THEOREM 4.1. *Let $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ $(n \geq m)$ be a set-valued mapping whose graph is locally closed. Assume that* Graph F *admits a C^k-Whitney stratification with $k > \dim(\text{Graph } F) - m$.*[7] *Then the collection of critical values of F has m-dimensional Lebesgue measure zero. In particular, if the graph of F is a definable set in a certain o-minimal structure, then the set of critical values of F is also a definable subset of $\mathbb{R}^m$ whose dimension is strictly smaller than m.*

The proof of the theorem relies on the Sard theorem and is relatively simple. We omit it mainly because it requires introducing rather many technical details involving manifolds. The basic idea of the proof is to show that every critical point of the mapping (in the sense of variational analysis, just defined) is a critical point (in the classical sense) of the restriction of the mapping to the stratum (of the Whitney stratification) to which the point belongs and then to use the Sard theorem. It is worth noting that the inclusion can be strict: Consider for instance the set-valued mapping $\mathbb{R} \rightrightarrows \mathbb{R}$ whose graph is the set $\{(x, y) : \ |x| = |y|\}$. Then zero is a regular value of it in the sense of variational analysis but, for any stratification, a critical value of the union of the restrictions of the projection $(x, y) \to y$ to the strata.

5. Typical normality of optimization problems. Consider the standard minimization problem with finite number of equality and inequality constraints:

$$\mathbf{P}(a) \qquad \left.\begin{array}{ll} \textit{minimize} & f_0(x) \\ \textit{subject to} & f_i(x) \leq \alpha_i,\ i = 1, \dots, k; \\ & f_i(x) = \alpha_i,\ i = k+1, \dots, m. \end{array}\right\}$$

Here $x \in \mathbb{R}^n$ and $a = (\alpha_1, \dots, \alpha_m) \in \mathbb{R}^m$. We denote by $\mathcal{F}(a)$ the set of *feasible* vectors for $\mathbf{P}(a)$. A vector $x \in \mathcal{F}(a)$ is called *normal* for $\mathbf{P}(a)$ if the following condition is satisfied: For no nonzero vector $l = (\lambda_1, \dots, \lambda_m)$ such that $\lambda_i \geq 0$ for $i = 1, \dots, k$ and $\lambda_i(f_i(x) - \alpha_i) = 0$ for all indices $i = 1, \dots, m$, zero belongs to $\partial(\lambda_1 f_1 + \dots + \lambda_m f_m)(x)$. (In the case when all f_i are continuously differentiable, this is the famous *Mangasarian–Fromowitz condition*.) An $x \in \mathcal{F}(a)$ that is not normal is called *abnormal*.

We say that $\mathbf{P}(a)$ is a *normal problem* if every feasible $x \in \mathcal{F}(a)$ is a normal vector for $\mathbf{P}(a)$.

[7] It is natural to define $\dim(\text{Graph } F)$ as the maximal dimension of the strata. Clearly, this does not depend on the specific stratification.

THEOREM 5.1. *Assume that all functions f_i are locally Lipschitz and definable in a certain o-minimal structure. Then the collection of the right-hand side vectors a for which the problem* P(a) *is not normal is a definable set of dimesion strictly smaller than m.*

Proof. We view $\mathcal{F}$ as a set-valued mapping from $\mathbb{R}^m$ into $\mathbb{R}^n$. This is obviously a definable set-valued mapping and so is its inverse $\mathcal{F}^{-1}$ which associates with every $x \in \mathbb{R}^n$ the collection of $a \in \mathbb{R}^m$ such that $x \in \mathcal{F}(a)$. By the definable Sard theorem (Theorem 4.1), the set of critical values of $\mathcal{F}^{-1}$ is a definable set of dimension strictly smaller than m. So to prove the theorem we need to verify that a is a critical value if and only if $\mathbf{P}(a)$ is not a normal problem.

Set $F(x) = (f_1(x), \dots, f_m(x))$. Set further

$$K = \{y = (\eta_1, \dots, \eta_m) \in \mathbb{R}^m : \eta_i \geq 0,\ i = 1, \dots, k;\ \eta_i = 0,\ i = k+1, \dots, m\}.$$

Then $\mathcal{F}^{-1}(x) = F(x) + K$. We have for any $a \in \mathbb{R}^m$

$$\begin{aligned} d(y, F(x)+K) &= \inf\{\|y - F(x) - w\| :\ w \in K\} \\ &= \inf\{\sup_{\|l\|\leq 1} \langle l, y - F(x) - w\rangle :\ w \in K\} \\ &= \sup_{\|l\|\leq 1} \inf\{\langle l, y - F(x) - w\rangle :\ w \in K\} \\ &= \sup\{\langle l, y - F(x)\rangle :\ \|l\| \leq 1,\ l \in K^\circ\}. \end{aligned} \tag{5.1}$$

Here $K^\circ = \{l :\ \langle l, y\rangle \leq 0,\ \forall\, y \in K\}$ is the polar of K.

By definition $\overline{a}$ is a critical value of $\mathcal{F}^{-1}$ if and only if there is an $\overline{x} \in \mathcal{F}(\overline{a})$ such that $\mathrm{sur}\mathcal{F}^{-1}(\overline{x}|\overline{a}) = 0$. By (3.3) this means that for any $\varepsilon > 0$ there is a pair (x, a) arbitrarily close to $(\overline{x}, \overline{a})$ and with $a \notin F(x)+K$ such that $|\nabla d(a, F(\cdot)+K)|(x) < \varepsilon/2$.

Let $l(x, y)$ be the l for which the last supremum in (5.1), for the given x and y, is attained. As $a \neq F(x)$, we have $\|l(x, a)\| = 1$ and, clearly $\langle l(x, a), y - F(x)\rangle \geq 0$. Furthermore, for any u sufficiently close to x

$$\varepsilon \geq |\nabla d(a, F(\cdot) + K)|(x) \geq \frac{\langle l(x, a), F(x) - F(u)\rangle}{\|x - u\|}.$$

It follows that the function $u \to \langle l(x, a), F(u)\rangle + \varepsilon\|u - x\|$ attains a local minimum at x. Hence

$$0 \in \partial(l(x, a) \circ F)(x) + \varepsilon B. \tag{5.2}$$

We may assume that $l(x, a)$ converges to some l as we choose ever smaller ε and (x, a) closer to $(\overline{x}, \overline{a})$. Clearly $\|l\| = 1,\ l \in K^\circ$, and $\langle l, F(\overline{x}) - \overline{a}\rangle \geq 0$. As $\overline{a} - F(\overline{x}) \in K$, the latter two relations imply that $\langle l, F(\overline{x}) - \overline{a}\rangle = 0$. Finally, from (5.2) taking into account that F is Lipschitz near $\overline{x}$ we get that $0 \in \partial(l \circ F)(\overline{x})$. □

Theorem 5.1 applies rather to the constraints of the problem. In [19] a similar problem will be considered in a more general context of constraint systems. The second result, also from [19], deals with critical values of the problem itself. Recall that the first order necessary optimality conditions in P(a) at $x \in \mathcal{F}(a)$ is the existence of a nonzero set of multipliers $\lambda_0, \lambda_1, \dots, \lambda_n$ satisfying

$$\lambda_i \geq 0,\ \lambda_i(f_i(x) - \alpha_i) = 0,\ i = 0, 1, \dots, k;\ 0 \in \partial(\lambda_0 f_0 + \lambda_1 f_1 + \cdots + \lambda_n f_n). \tag{5.3}$$

We shall say that x is a *critical point* of P(a) if (5.3) holds for some nonzero set of multipliers. We shall further say that α is a *critical value* of P(a) if $f_0(x) = \alpha$ for some critical point x of P(a). The following theorem will be proved in [19].

THEOREM 5.2. *Assume that the functions $f_0, f_1, \dots, f_n$ are locally Lipschitz and definable in a certain o-minimal structure. Assume further that* P(a) *is a normal problem. Then the set of critical values of* P(a) *is finite.*

Of course, if $\mathbf{P}(a)$ is normal and (5.1) holds, then $\lambda_0 > 0$.

6. A set-valued Łojasiewicz inequality. In 1965 Łojasiewicz proved that a bounded trajectory of the gradient dynamic system

$$\dot{x} = -\nabla f(x)$$

is finite and converges to a critical point of f if f is a real analytic function. The proof was based on the inequality obtained by Łojasiewicz for real analytic functions a few years earlier: If f is a real analytic function and x is a critical point of f, then there is an $\alpha \in (0,1)$ such that in a neighborhood of x

$$\|\nabla f(u)\| \geq |f(u) - f(x)|^{\alpha}.$$

(The inequality is of course trivial if x is not a critical point.)

In 1998 Kurdyka [21] extended the Łojasiewicz inequality to continuously differentiable functions definable in an o-minimal structure. Kurdyka's result is stated as follows. Let f be a C^1 function on a bounded open set $U \subset \mathbb{R}^n$ which is definable in some o-minimal structure. Suppose that $f(x) > 0 = \inf_U f$ for all $x \in U$. Then there is a $\rho > 0$ and a strictly increasing definable function ψ on $(0, \infty)$ such that for all $x \in U$ with $f(x) < \rho$ we have[8]

$$\|\nabla(\psi \circ f)(x)\| \geq 1. \tag{6.1}$$

Finally, very recently Bolte–Daniilidis–Lewis–Shiota [4] extended the last theorem to arbitrary lower semicontinuous definable functions by proving that a more general version of (6.1) with $\psi'(f(x))\|y\|$ in the left-hand side, where y stands for an arbitrary element of $\partial_c f(x)$.

The theorem below is a Kurdyka-type result for set-valued mappings into $\mathbb{R}$.

THEOREM 6.1. *Let $F : \mathbb{R}^n \rightrightarrows \mathbb{R}$ be a tame set-valued mapping with closed graph, and let U be a bounded open definable set. Take a $\tau \in \mathbb{R}$ and assume that $(\tau, \tau+\delta) \cap F(U) \neq \emptyset$ for any $\delta > 0$. Then there are $\rho > 0$ and a nonnegative continuous function $\psi(t)$ defined at least on $[\tau, \tau + \rho)$ which is continuously differentiable and strictly increasing on $(\tau, \tau + \rho)$ and such that the inequality*

$$\operatorname{sur}(\psi \circ F)(x|\psi(\tau + h)) \geq 1$$

holds for all $x \in U$ and all $h \in (0, \rho)$ such that $\tau + h \in F(x)$.

Proof. Fix a certain $T > 0$. The intersection of the graph of F with $U \times (0, T)$ is a definable set by definition of tameness. Let

$$\varphi(t) = \inf\{\operatorname{sur} F(x|t) : \ x \in U, \ t \in F(x)\}, \quad t \in (0, T).$$

It follows from Proposition 3.2 that φ is a definable function. By the monotonicity theorem (Theorem 2.1) the limit $r = \lim_{t \to 0} \varphi(t)$ exists.

[8] The proof that (6.1) implies the Łojasiewicz inequality for real-analytic functions is based on a highly nontrivial property of subanalytic functions, the so-called Puiseux expansion. We refer to [21] for explanations.

If $r > 0$, then $\psi(t) = kt$ with $k > r^{-1}$ is the desired function. So we assume that $r = 0$ in the rest of the proof.

By the monotonicity theorem (Theorem 2.1), there is a $\delta > 0$ such that either $\varphi(t) \equiv 0$ on $(0, \delta)$ or φ is a strictly increasing, positive, and continuously differentiable function. In the first case, we would have that for any $t \in (0, \delta)$ there is an x in the closure of U such that $\mathrm{sur}F(x|t) = 0$. This would mean that every $t \in (0, \delta)$ is a critical value of F. But the latter contradicts to Theorem 4.1. Thus φ is a continuously differentiable positive and strictly increasing function on $(0, \delta)$ going to zero as $t \to 0$.

To prove the theorem it is enough to verify that $\eta(t) = [\varphi(t)]^{-1}$ is summable on $(0, \rho)$ for some positive ρ. Indeed, in this case

$$\psi(t) = \int_0^t \frac{1}{\varphi(\xi)} d\xi$$

satisfies $\psi'(t)\varphi(t) \equiv 1$ on some interval $(0, \rho)$ and $\mathrm{sur}(\psi \circ F)(x|t)) = \psi'(t)\mathrm{sur}F(x|t) \geq \psi'(t)\varphi(t) \geq 1$ which is immediate from the definition of the modulus of surjection.

Consider the set-valued mapping $\Phi : (0, \delta) \mapsto \mathbb{R}^n$ defined by

$$t \mapsto \{x : \ x \in U, \ t \in F(x), \ 2\varphi(t) \geq \mathrm{sur}F(x|t)\}.$$

This is a definable mapping with nonempty values. By the definable choice theorem (Theorem 2.2) there is a definable selection $x(t)$ for Φ. Applying the monotonicity theorem (and taking again a smaller δ if necessary) we can be sure that $x(t)$ is continuously differentiable on $(0, \delta)$.

As U is a bounded set, it follows from the monotonicity theorem (applied to each component of $x(t)$) that $x(t)$ has a finite length. Set $H(t) = F(x(t))$. We also have $t \in H(t)$ for all positive t close to zero. This means that $\mathrm{sur}H(t|t) \geq 1$ for such t.

On the other hand, as $x(t)$ is continuously differentiable,

$$\mathrm{sur}H(t|t) \leq \|\dot{x}(t)\|\mathrm{sur}F(x(t)|t) \leq 2\varphi(t)\|\dot{x}(t)\|$$

(see (3.2)), so that

$$\int_0^\delta \frac{1}{\varphi(t)} dt \leq 2 \int_0^\delta \|\dot{x}(t)\| dt < \infty$$

as desired. □

COROLLARY 6.2. *Let f be a lower semicontinuous function on $\mathbb{R}^n$, and let U be an open bounded subset of $\mathbb{R}^n$. If f is tame in some o-minimal structure and U is a definable set in the same structure, then for any $\tau \in \mathbb{R}$ there are $\rho > 0$ and a nonnegative continuous function $\psi(t)$ defined at least on $[\tau, \tau+\rho)$ which is continuously differentiable and strictly increasing on $(\tau, \tau + \rho)$ and such that the inequality*

$$|\nabla(\psi \circ f)|(x) \geq 1$$

holds for all $x \in U$ such that $0 < f(x) - \tau < \rho$.

Proof. Consider the set-valued mapping $F(x) = \{\alpha \in \mathbb{R} : \ \alpha \geq f(x)\}$, that is, the mapping whose graph is epi f. As f is lower semicontinuous, the graph of F is closed and we can apply the theorem. Fix a certain $\overline{x} \in \mathrm{dom}\ f$ and let $\overline{y} = f(\overline{x}) \in F(\overline{x})$. If $y \notin F(\overline{x})$, then the distance from y to $F(x)$ for all x sufficiently close to $\overline{x}$ is

$f(x) - y$. Therefore by (3.3) $\mathrm{sur}F(\overline{x}|\overline{y}) \le |\nabla f|(\overline{x})$. This inequality holds for all $(\bar{x}, \bar{y}) \in \mathrm{Graph}\, F$, so the application of the theorem proves the claim. □

The corollary also follows from the mentioned result of [4]; moreover, we do not claim that the function ψ is definable. The advantage of Theorem 6.1 and the corollary is simplicity of the proof which does not rely either on the results of [21][9] or any advanced results on definable functions, beyond those quoted in section 3. We shall apply the corollary in the next section to extend Łojasiewicz's gradient descent theorem to slopes of lower semicontinuous definable functions.

7. Curves of maximal slope. As we mentioned in the beginning of the previous section, one of the most important applications of the Łojasiewicz inequality was his proof that bounded trajectories of gradient descent of real analytic functions have bounded length. Kurdyka showed in [21] that the same is true for continuously differentiable definable functions, and recently Bolte–Daniilidis–Lewis [3] established that a similar property is shared by trajectories of subgradient descent of definable "lower C^2-functions."[10]

Here we show that the property can be further extended to arbitrary lower semicontinuous functions. The key to this extension is the concept of a curve of maximal slope introduced by DeGiorgi–Marino–Tosques in [9]. Namely, given an extended-real-valued function f on $\mathbb{R}^n$ (the original definition actually applies to functions on metric spaces), a curve γ in $\mathbb{R}^n$ is called a *curve of maximal slope for* f if there is a parameterization $u(t)$ for γ such that

(i) $u(0) \in \mathrm{dom}\, f$ and $(f \circ u)(t)$ does not increase with t;

(ii) $u(t)$ is absolutely continuous and $\|u'(t)\| = |\nabla f|(u(t))$ for almost every $t \in (0, T)$;

(iii) $f \circ u$ is absolutely continuous and $(f \circ u)'(t) = -[|\nabla f|(u(t)]^2$ almost everywhere on $(0, T)$.

We leave aside the question about the existence of curves of maximal slope and refer to [9, 25] for the basic existence theorems. In the nutshell, curves exist for any initial $u(0)$, provided the slope function $|\nabla f|(x)$ behaves sufficiently well. In particular, if f is a continuously differentiable function, then curves of maximal slope for f are precisely trajectories of the antigradient equation $\dot{x} = -\nabla f(x)$; if f is a convex or, more generally, lower C^1 function, then curves of maximal slope are precisely trajectories of the antisubgradient inclusion $\dot{x} \in -\partial f(x)$ (see, e.g., [25], Theorem 1.11).

Our purpose is to prove the following theorem.

THEOREM 7.1. *Let f be a definable lower semicontinuous function on $\mathbb{R}^n$, and let U be an open and bounded subset of $\mathbb{R}^n$. Then there is a number N such that the length of any curve of maximal slope for f lying in U does not exceed N.*

Proof. The proof below is basically a readjustment of that of [21] for our more general setting.

1. To begin with, we observe the following. Let γ be a curve of maximal slope for f, and let φ be a strictly increasing continuously differentiable function on a certain interval containing the values of $f(x)$ for $x \in \gamma$. Then γ is a curve of maximal slope

[9] The idea to study the mapping Φ has been borrowed from [21] but the method of the subsequent analysis is different here.

[10] A lower C^k, $k \ge 1$, function is defined as a function that can be locally represented by difference of a convex and C^k function. We observe that for lower C^k functions the four subdifferentials mentioned in section 3 are identically equal if $k \ge 2$, and for $k = 1$ the same is true for Fréchet, limiting and Clarke subdifferentials.

for $\varphi \circ f$. Indeed, let $u(t)$ be a parameterization of γ satisfying (i)–(iii). Then, as

$$|\nabla(\varphi \circ f)|(x) = \varphi'(t)|\nabla f|(x),$$

the function $v(\tau) = u(t(\tau))$ with $\tau(t)$ defined by

$$\frac{d\tau}{dt} = \big(\varphi'(f(u(t))\big)^{-1}$$

is a parameterization of γ showing that (i)–(iii) holds for $\varphi \circ f$; that is, γ is also a curve of maximal slope for $\varphi \circ f$.

As the first consequence of this observation, we may assume that f is bounded on U: Otherwise we can replace f by, say, $\varphi \circ f$, where $\varphi(t) = t/\sqrt{1+t^2}$.

2. Set

$$\varphi(s) = \inf\{|\nabla f|(x) : \ x \in U, \ f(x) = s\}.$$

Again we see that φ is a definable function. So the monotonicity theorem implies that, up to a finite number of points, the domain of φ is the union of finitely many, say k open intervals (α_i, β_i) such that φ is continuous and either strictly monotone or constant on each of them. Let c_i be the lower bound of φ on the ith interval. We observe that $\varphi(f(s))$ is strictly positive on each interval. This is obvious if $c_i > 0$. But if $c_i = 0$ for some i, then φ cannot be constant on the interval because of Theorem 4.1 (as in the proof of Theorem 5.1).

3. Let now γ be a curve of maximal slope and $u(t)$ a corresponding parameterization of γ. Then $f(u(t))$ does not increase with t. In principle, $f(u(t))$ can be constant on finitely many intervals on which it is equal to a critical value of f (that is to s such that $\varphi(s) = 0$) and $u(t)$ is constant. But we can eliminate such intervals by reparameterizing γ in an obvious way and assume that $f(u(t))$ is strictly decreasing. Define ξ_i and η_i by $f(u(\xi_i)) = \beta_i$, $f(u(\eta_i)) = \alpha_i$, and let l_i stand for the length of the piece of γ between $u(\xi_i)$ and $u(\eta_i)$. We have

$$l_i = \int_{\xi_i}^{\eta_i} \|u'(t)\|dt = \int_{\xi_i}^{\eta_i} |\nabla f|(u(t))dt \le \Big((\eta_i - \xi_i)\int_{\xi_i}^{\eta_i} |\nabla f|(u(t))^2 dt\Big)^{1/2}.$$

We also have (in view of (iii))

$$\int_{\xi_i}^{\eta_i} |\nabla f|(u(t))^2 dt = f(u(\xi_i)) - f(u(\eta_i)) = \beta_i - \alpha_i.$$

If $c_i > 0$, then $l_i \ge c_i(\eta_i - \xi_i)$ which, combined with the two inequalities above gives

$$l_i \le \frac{\beta_i - \alpha_i}{c_i}.$$

4. If, on the other hand $c_i = 0$, then by Corollary 6.2 (assuming, to be certain, that φ is strictly increasing on (α_i, β_i)) we shall find a function $\psi_i(s)$ which is defined and continuous on $[\alpha_i, \alpha_i + \rho)$ for some $\rho > 0$ continuously differentiable and strictly increasing on $(\alpha_i, \alpha_i + \rho)$ and such that $|\nabla(\psi_i \circ f)|(u(t)) \ge 1$ if $x \in U$ and $0 < f(x) - \alpha_i < \rho$. As $\varphi(s)$ is strictly increasing on (α_i, β_i), so that $|\nabla f|(x) \ge \varphi(\alpha_i + \rho) > 0$ if $\beta_i > f(x) \ge \alpha_i + \rho$, we can extend ψ_i to a continuous function on the entire segment

$[\alpha_i, \beta_i]$ continuously differentiable and strictly increasing in the interior of the segment with the inequality $|\nabla(\psi \circ f)|(x) \geq 1$ being valid for all $x \in U$ with $\alpha_i < f(x) < \beta_i$.

Then, as we have seen in the beginning of the proof, the piece of γ corresponding to the values of $u(t)$ for $t \in (\xi_i, \eta_i)$ is a curve of maximal slope for $\psi_i \circ f$. As the slope of the function is not smaller than one for all values of f between α_i and β_i, we get as above that $l_i \leq \psi_i(\beta_i) - \psi_i(\alpha_i)$.

If we now denote by I the collection of indices for which $c_i > 0$, then we conclude that the length of the curve does not exceed

$$N = \sum_{i \in I} \frac{\beta_i - \alpha_i}{c_i} + \sum_{i \in \{1,\dots,k\} \setminus I} (\psi_i(\beta_i) - \psi_i(\alpha_i)),$$

which completes the proof as α_i, β_i, c_i, and ψ_i are determined by f and do not depend on the choice of γ. □

8. Semismoothness of tame mappings. A mapping $F : \mathbb{R}^n \to \mathbb{R}^m$ is called semismooth near $\overline{x}$ if it satisfies the Lipschitz condition near $\overline{x}$ and for each x of the neighborhood it is differentiable along every direction and

$$\|F'(x + h; h) - F'(x; h)\| = o(\|h\|).$$

The latter amounts to

$$r(t) = t^{-1} \max_{d \in S^{n-1}} \|F'(x + td; d) - F'(x; d)\| \to 0$$

as $t \to 0$.

The role of the semismoothness property introduced in 1977 by Mifflin is determined by the fact that it guarantees superlinear convergence of the Newton method. We refer to [12] for details.

THEOREM 8.1 ([5]). *A locally Lipschitz tame mapping $F : \mathbb{R}^n \to \mathbb{R}^m$ is semismooth near every point of its domain.*

LEMMA 8.2. *Let $\varphi(t)$ and $\psi(t)$ be defined on a certain interval $(0, T)$ and have the following properties:*

(a) $\varphi(0) = \psi(0) = 0$;

(b) *both functions are definable in a certain o-minimal structure;*

(c) $\psi'(t) > 0$ *for* $t > 0$ *sufficiently close to zero. Then*

$$\lim_{t \to 0} \frac{\varphi(t)}{\psi(t)} = r \;\Rightarrow\; \lim_{t \to 0} \frac{\varphi'(t)}{\psi'(t)} = r.$$

Proof. Fix a $\delta > 0$ and consider the functions $\eta_{\mp}(t) = \varphi(t) - (r \pm \delta)\psi(t)$. Clearly $\eta_{\pm}(0) = 0$ and there is an $\varepsilon > 0$ such that $\eta_-(t) < 0$ and $\eta_+(t) > 0$ for $t \in (0, \varepsilon)$. It follows that any neighborhood of zero must contain points at which $\eta'_-(t) < 0$ and points at which $\eta'_+(t) > 0$. Applying the monotonicity lemma to $\eta'_{\pm}$ and taking a smaller ε , if necessary, we can be sure that the inequalities hold for all $t \in (0, \varepsilon)$ and, consequently,

$$\left| \frac{\varphi'(t)}{\psi'(t)} - r \right| \leq \delta, \ \forall\, t \in (0, \varepsilon),$$

and the lemma follows. □

Proof of the theorem. Fix an $h \in S^{n-1}$. The mapping $t \to F(x+th) - F(x)$ is definable and Lipschitz in a neighborhood of zero, so applying the monotonicity lemma to each of its components, we conclude that the directional derivative $F'(x;h)$ exists for all x of the domain of F and all $h \in S^{n-1}$.

To prove the theorem we have to show that, given an x in the domain of F, the quantity $F'(x+th;h) - F'(x;h) \to 0$ uniformly for $h \in S^{n-1}$ as $t \to +0$. Assuming the contrary, we shall find, using the definable choice theorem, an $\varepsilon > 0$ and a definable mapping $h : (0,\varepsilon) \mapsto S^{n-1}$ such that

$$\|F'(x+th(t);h(t)) - F'(x;h(t))\| \geq \alpha > 0 \tag{8.1}$$

for all sufficiently small t. Then $h(t)$ converge to some $h \in S^{n-1}$ as $t \to 0$, so that $th(t) = th + \xi(t)$, where $\|\xi(t)\| = o(t)$ and $\xi(t)$ is obviously a definable mapping. By Lemma 8.2, $\xi'(t) \to 0$ as $t \to 0$.

The mapping $t \to F(x+th(t))$ is definable as a composition of definable mappings. We can therefore assume (taking a smaller ε, if necessary) that it is differentiable for all $t \in (0,\varepsilon)$ We claim that

$$\lim_{t\to 0} \frac{d}{dt} F(x+th(t)) = F'(x;h) \tag{8.2}$$

Indeed, let K be the Lipschitz constant of F in a neighborhood of x. Then

$$t^{-1}\|F(x+th(t)) - F(x+th)\| \leq Kt^{-1}\|\xi(t)\| \to 0$$

as $t \to 0$, so that

$$\lim_{t\to 0} t^{-1}(F(x+th(t)) - F(x)) = F'(x;h),$$

which, together with the lemma implies (8.2).

On the other hand, for $t \in (0,\varepsilon)$ we have

$$\frac{d}{dt} F(x+th(t)) = F'(x+th(t); h+\xi'(t)).$$

As $F(z;\cdot)$ is Lipschitz continuous with constant K for all z of a neighborhood of x, it follows that $\|F'(x,h(t)) - F'(x;h)\| \to 0$ and

$$F'(x+th(t); h+\xi'(t)) - F'(x+th(t);h(t)) \to 0$$

as $t \to 0$ which together with (8.2) leads to a contradiction with (8.1). □

9. Definable Lyapunov theorem. The famous Lyapunov theorem on vector measures says that the image set of a finite nonatomic $\mathbb{R}^n$-valued measure is a convex compact subset of $\mathbb{R}^n$. A particular case of this theorem is the following well-known result having important applications in optimal control theory: *If $F : [0,T] \rightrightarrows \mathbb{R}^n$ is a measurable set-valued mapping with closed values and there is a summable $r(t)$ on $[0,T]$ such that for almost every t the norm of every element of $F(t)$ does not exceed $r(t)$, then the integral $\int_0^T F(t)dt$ is a convex compact set.* (Here as usual the integral of a set-valued mapping is defined as a set of integrals of its summable selections.) As a consequence of this latter result, we get that the integral of any measurable set-valued mapping into $\mathbb{R}^n$ is either an empty set or a convex set.

Now let $\mathcal{S}$ be a certain o-minimal structure. We denote by

$$(\mathcal{S})\int_0^T F(t)dt$$

the collection of integrals of all definable (in $\mathcal{S}$) selections of F (which are piecewise continuous, hence measurable). Let us agree to call this set the *definable* or $\mathcal{S}$*-integral* of F.

THEOREM 9.1. *Suppose that $F : [0,T] \rightrightarrows \mathbb{R}^n$ is a set-valued mapping with closed values which is definable in a certain o-minimal structure $\mathcal{S}$. Then*

$$(\mathcal{S})\int_0^T F(t)dt = \int_0^T F(t)dt.$$

Proof. We first observe that the $\mathcal{S}$-integral is a subset of the standard integral, so that only the opposite inclusion needs proof. We note next that $(\mathcal{S})\int F(t)dt$ is nonempty if so is $\int F(t)dt$. Indeed, let the latter integral be nonempty, that is there is a summable selection $x(t)$ of F. The function $\rho(t) = \min\{\|x\| : x \in F(t)\}$ is then definable and summable, so the set-valued mapping $F_0(t) = \{x \in F(t) : \ \|x\| \le \rho(t)\}$ is also definable, $F_0(t) \neq \emptyset$ almost everywhere and any definable selection of F_0 is summable. Finally, we can assume of course that $T = 1$. With these three points in mind, we start the proof.

In what follows we set

$$\Phi = \int_0^1 F(t)dt$$

and assume that $\Phi \neq \emptyset$.

1. By the Lyapunov theorem Φ is a convex set. We shall prove the theorem using induction by the dimension of Φ. First we prove the theorem for the case $\dim \Phi = 0$. This is immediate from the above remark that the $\mathcal{S}$-integral is nonempty if so is Φ. Indeed, if the dimension of Φ is zero, then, being a convex set, Φ must be a singleton. This may happen only if $F(t)$ is single-valued up to a set of measure zero and therefore the integral of any of its definable selections must coincide with the unique element of Φ.

2. Assume that the theorem is valid for closed-valued definable mappings into $\mathbb{R}^k$ with $k < n$ (or equivalently, for closed-valued definable mappings with $\dim \Phi = k < n$). We claim that the theorem is true if it is true under the additional boundedness assumption: $0 \in F(t)$ almost everywhere and there is an $N > 0$ such that for almost every t the norm of any $x \in F(t)$ does not exceed N.

Indeed, given F, let F_N be defined by

$$F_N(t) = (F(t) - z(t)) \bigcap NB,$$

where $z(t)$ is any definable and summable selection of F and B stands for the unit ball, and let

$$\Psi_N = \int_0^1 F_N(t)dt, \quad \Psi = \bigcup \Psi_N.$$

Let z be the integral of $z(t)$. By the Lyapunov theorem, (Ψ_N) is a nondecreasing sequence of convex closed bounded sets, so Ψ is convex. Clearly, $\Psi \subset \Phi - z$. There

is no loss of generality in assuming that $z = 0$, which we shall do in the discussion to follow. We intend to show that $\Phi \setminus \Psi \subset (\mathcal{S}) \int F(t)dt$ which is of course sufficient to prove the claim since, according to our assumption, every element of Ψ belongs to $(\mathcal{S}) \int F_N dt$ for some N.

Suppose $\overline{x} \in \Phi \setminus \Psi$. Then $\overline{x}$ can be separated from Ψ by a nonzero vector, that is, there is a $p \in \mathbb{R}^n$ such that $\langle p, \overline{x} \rangle \geq \langle p, x \rangle$ for all $x \in \Psi$. Set $F_p(t) = \{x \in F(t) : \langle p, x \rangle = m_p(t)\}$ and $\Phi_p = \int F_p(t)dt$, where

$$m_p(t) = \sup\{\langle p, x \rangle : \ x \in F(t)\}.$$

Then $m_p(t)$ is a definable function and therefore F_p is a definable mapping with closed values (which in principle can be empty). We claim that

$$\langle p, \overline{x} \rangle = \int_0^1 m_p(t)dt. \tag{9.1}$$

Clearly $\langle p, \overline{x} \rangle \leq \int m_p(t)dt$. Assuming that $\langle p, \overline{x} \rangle < \int m_p(t)dt$, we would be able to find a $w \in \Phi$ and an $\varepsilon > 0$ such that $\langle p, w \rangle \geq \langle p, x \rangle + \varepsilon$ for any $x \in \Psi$. This would mean that

$$\int_0^1 m_p(t) \geq \int_0^1 \max\{\langle p, x \rangle : \ x \in F_N(t)\}dt + \varepsilon, \quad \forall\, N,$$

which is certainly untrue as the maximum under the integral pointwise and non-decreasingly converges to $m_p(t)$ as $N \to \infty$. So (9.1) indeed holds.

Let now $\overline{x}(t)$ be a selection of Φ whose integral is $\overline{x}$. By (9.1) we necessarily have $\langle p, \overline{x}(t) \rangle = m_p(t)$ almost everywhere, which means that the supremum in the definition of m_p is attained for almost every t and the projection of the graph of F_p onto $[0, 1]$ is a set of full measure. As F_p is definable, the projection is also a definable set and, as it has full measure, it must coincide with $[0, 1]$ up to finitely many points. This means, in turn, that the supremum in the definition of m_p is actually attained for all t except for a finite number of points. Thus we have $\overline{x} \in \Phi_p$. But the dimension of Φ_p must be strictly smaller than n, so by the induction assumption $\overline{x} \in (\mathcal{S}) \int F_p(t)dt$.

3. So we assume henceforth that there is an $N > 0$ such that for almost every t we have $\|x\| \leq N$ for all $x \in F(t)$. In this case by the Lyapunov theorem Φ is a convex compact set. Repeating the induction arguments of the previous step, we can easily conclude that every boundary point of Φ belongs to $(\mathcal{S}) \int F(t)dt$, so we have to show that this is true for all interior points of Φ.

So let $\overline{x} \in \operatorname{int} \Phi$. Choose $n + 1$ affinely independent points $x_1, \ldots, x_{n+1}$ on the boundary of Φ such that $\overline{x}$ belongs to the interior of their convex hull and let $r > 0$ be such that $B(\overline{x}, r) \subset \operatorname{conv}\{x_1, \ldots, x_{n+1}\}$. Being boundary points, $x_i \in (\mathcal{S}) \int F(t)dt$, so there are definable selections $x_i(t)$ of F whose integrals give x_i. Applying the monotonicity theorem to each component of every $x_i(t)$, we can find finitely many points $0 = \tau_0 < \tau_1 < \ldots < \tau_m = 1$ such that every $x_i(t)$ is continuous on each interval (τ_j, τ_{j+1}) with each component function of every $x_i(t)$ being either constant or strictly monotone on each interval. Adding more points, if necessary, we can be sure that $\|x_i(\tau_j + 0) - x_i(\tau_{j+1} - 0)\| < \varepsilon$ for a chosen positive ε. (Here $x(\tau + 0) = \lim_{t \searrow \tau} x(t)$ and $x(\tau - 0) = \lim_{t \nearrow \tau} x(t)$.)

We denote by $\Sigma_n = \{a = (\alpha_1, \ldots, \alpha_{n+1}) : \ \alpha_i \geq 0, \ \sum \alpha_i = 1\}$ the standard n-simplex and for any $j \in \{1, \ldots, m\}$ and any $a \in \Sigma_n$ split (τ_j, τ_{j+1}) into n consecutive intervals $\Delta_{ij} = \Delta_{ij}(a)$ such that $|\Delta_{ij}| = \alpha_i |\Delta_j|$ (where $|\cdot|$ stands for the length of an

interval). In other words,

$$\Delta_{ij} = (\tau_j + (\alpha_0 + \cdots + \alpha_{i-1})(\tau_{j+1} - \tau_i), \tau_j + (\alpha_0 + \cdots + \alpha_i)(\tau_{j+1} - \tau_i))$$

(where $\alpha_0 = 0$).

Finally we define ξ_{ij} as the arithmetic mean of the values of $x_i(t)$ at the ends of Δ_j (or to be more precise of the limits of $x_i(t)$ when t tends to the ends of Δ_j from within the interval) and set

$$\Delta_i(a) = \bigcup_j \Delta_{ij}(a); \quad u_i(t,a) = \begin{cases} x_i(t), & \text{if } t \in \Delta_i(a)\ , \\ 0, & \text{otherwise.} \end{cases}$$

We have $\|\xi_{ij} - x_i(t)\| < \varepsilon/2$ if $t \in (\tau_j, \tau_{j+1})$, that is

$$\|\xi_{ij}(\tau_{j+1} - \tau_j) - \int_{\tau_j}^{\tau_{j+1}} x_i(t)dt\| < \frac{\varepsilon}{2}(\tau_{j+1} - \tau_j),$$

so that

$$\|\xi_{ij}|\Delta_{ij}| - \alpha_i \int_{\tau_j}^{\tau_{j+1}} x_i(t)dt\| < \alpha_i \frac{\varepsilon}{2}(\tau_{j+1} - \tau_j).$$

On the other hand,

$$\|\xi_{ij}|\Delta_{ij}| - \int_{\Delta_{ij}} x_i(t)dt\| < \frac{\varepsilon}{2}|\Delta_{ij}| = \alpha_i \frac{\varepsilon}{2}(\tau_{j+1} - \tau_j),$$

and by comparing the last two inequalities, we get

$$\| \int_0^1 u_i(t,a)dt - \alpha_i x_i\| = \| \int_0^1 u_i(t,a)dt - \alpha_i \int_0^1 x_i(t)dt\| < \alpha_i \varepsilon.$$

As all $x_i(t)$ are continuous on every (τ_j, τ_{j+1}), it follows from the definition of Δ_{ij} that integrals of $u_i(t,a)$ depend continuously on a. Consider the following two mappings from Σ_n into $\mathbb{R}^n$:

$$\varphi(a) = \sum_{i=1}^n \alpha_i x_i, \qquad \psi(a) = \sum_{i=1}^n \int_0^1 u_i(t,a)dt.$$

The first is just a linear homeomorphism of Σ_n onto conv $\{x_1, \ldots, x_{n+1}\}$ (recall that $x_1, \ldots, x_{n+1}$ are affinely independent). The second mapping is continuous, and as follows from the last inequality $\|\varphi(a) - \psi(a)\| < \varepsilon$ for all $a \in \Sigma_n$. The latter can be rewritten as $\|(\psi \circ \varphi^{-1})(x) - x\| < \varepsilon$ for all $x \in$ conv $\{x_1, \ldots, x_{n+1}\}$.

It remains to recall that $\varepsilon < r/2$ and $B(\overline{x}, r) \subset$ conv $\{x_1, \ldots, x_{n+1}\}$. Now take an $w \in B(\overline{x}, r/2)$ and consider the mapping $x \mapsto F_w(x) = F(x) - w$, where we set $F = \psi \circ \varphi^{-1}$. Then for any x with $\|x - \overline{x}\| = r$ we have

$$\langle x - \overline{x}, F_w(x)\rangle = r^2 + \langle x - \overline{x}, F(x) - x\rangle + \langle x - \overline{x}, \overline{x} - w\rangle > 0,$$

so by the Borsuk antipodal theorem there is an $x \in B(\overline{x}, r)$ such that $F(x) = w$. Thus the ball $B(\overline{x}, r/2)$ is covered by the image of conv $\{x_1, \ldots, x_{n+1}\}$ under $\psi \circ \varphi^{-1}$. In other words, there is an $a \in \Sigma_n$ such that $\overline{x} = \psi(a)$. Setting $u_i(t) = u_i(t,a)$ and

$$u(t) = \sum_{i=1}^n u_i(t)dt,$$

we easily observe that $u(t)$ is a definable selection of F (as its graph is the union of finitely many definable pieces of graphs of $x_i(t)$) and the integral of $u(t)$ over $[0,1]$ is $\overline{x}$. This completes the proof of the theorem. ∎

10. A class of variational problems. Most general existence theorems in calculus of variations and optimal control, going back to Tonelli, give existence of absolutely continuous solutions. Proofs of existence in "friendlier" classes of functions usually need some additional analytic properties of the data (e.g., positive definiteness of certain second order derivatives). We refer to [7, 30] for details and related discussions.

In this concluding section we apply the definable Lyapunov Theorem 9.1 to prove the existence of piecewise smooth solutions in a class of variational problem. In the simplest form problems of this class are stated as follows:

$$\text{minimize} \quad \int_0^T f(t,u(t))dt, \text{ s.t. } \int_0^T u(t)dt = a. \tag{10.1}$$

The problem, with all simplicity of its formulation, represents a wide array of variational and optimal control problems, including state-linear problems of optimal control (see, e.g., [20]) or many variational problems arising in mathematical economics for economies with a measure space of agents (e.g., the famous Aumann–Perles problem [1] and its various extensions).

We shall consider the problem under the assumption that the integrand is a definable function. The immediate gain we shall get from Theorem 9.1 is the existence of definable (hence, piecewise smooth) solutions. Namely, Theorem 9.1 implies the following result.

THEOREM 10.1. *Suppose the integrand f in* (10.1) *is a definable (in a certain o-minimal structure) extended-real-valued function. Then the value of the problem coincides with its value if we restrict the problem to the class of definable (in the same o-minimal structure) controls $u(t)$.*

Moreover, if the problem has a solution, then it has also a definable solution.

Proof. Indeed, consider the mapping

$$\varphi(t,u) = \begin{pmatrix} f(t,u) \\ u \end{pmatrix}$$

from $[0,T] \times \mathbb{R}^n$ into $\mathbb{R}^n$. (Strictly speaking, φ is a set-valued mapping which is single-valued on its domain, defined as above when $f(t,u) < \infty$ and equal to $\emptyset$ if $f(t,u) = \infty$.) If f is definable, then so is φ. So if a $u(t)$ is such that

$$\int_0^T u(t)dt = a, \quad \int_0^T f(t,u(t))dt = \alpha,$$

then Theorem 9.1, applied to the set-valued mapping $F(t) = \varphi(t,\mathbb{R}^n)$, guarantees the existence of a definable $v(t)$ such that

$$\int_0^T v(t)dt = a, \quad \int_0^T f(t,v(t))dt = \alpha,$$

and the theorem follows. □

Combining Theorem 10.1 with the existence theorem for (10.1) ([20], Theorem 10.3.1), we arrive at the following conclusion. Set

$$S^*(p) = \int_0^T f^*(t,p)dt,$$

where $f^*(t,p)$ is the Fenchel transform of the function $u \to f(t,u)$.

THEOREM 10.2. *Assume that f is definable in a certain o-minimal structure and lower semicontinuous in u. Set*

$$S(x) = \sup_p (p \cdot x - S^*(p)),$$

and let $\bar{p} \in \partial S(a)$. If $a \in \mathrm{ri}(\mathrm{dom}\ S)$ and $\bar{p} \in \mathrm{int}\ (\mathrm{dom}\ S^*)$, then (10.1) *has a definable solution.*

Thus, under the conditions of the last theorem, the problem has a piecewise smooth optimal control.

Acknowledgments. I am thankful to A. S. Lewis and the referees for careful reading and useful remarks.

REFERENCES

[1] R.J. AUMANN AND M. PERLES, *A variational problem arising in economics*, J. Math. Anal. Appl., 11 (1965).
[2] S.M. BATES, *Toward a precise smoothness hypothesis in Sard's theorem*, Proc. Amer. Math. Soc., 117 (1993), pp. 279–283.
[3] J. BOLTE, A. DANIILIDIS, AND A.S. LEWIS, *The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems*, SIAM J. Optim., 17 (2007), pp. 1205–1223.
[4] J. BOLTE, A. DANIILIDIS, A.S. LEWIS, AND M. SHIOTA, *Clarke subgradients of stratifiable functions*, SIAM J. Optim., 18 (2007), pp. 556–572.
[5] J. BOLTE, A. DANIILIDIS, AND A.S. LEWIS, *Tame mappings are semismooth*, Math. Program. Ser. B, 117 (2009), pp. 5–17.
[6] J.M. BORWEIN AND X. WANG, *Distinct differentiable functions may share the same Clarke subdifferential at all points*, Proc. Amer. Math. Soc., 125 (1997), pp. 807–813.
[7] G. BUTTAZZO, M. GIAQUINTA, AND S. HILDEBRANDT, *One-Dimensional Variational Problems*, Clarendon Press, Oxford, 1998.
[8] M. COSTE, *An Introduction to o-Minimal Geometry*, Inst. Rech. Math., Univ. de Rennes, 1999, http://name.math.univ-rennes1.fr/michel.coste/polyens/OMIN.pdf.
[9] E. DE GIORGI, A. MARINO, AND M. TOSQUES, *Problemi di evoluzione in spazi metrici e curve di massima pendenza*, Atti Acad. Naz. Lincei, Rend. Cl. Sci. Fiz. Mat. Natur., 68 (1980), pp. 180–187.
[10] L. VAN DEN DRIES, *Tame Topology and O-minimal Structures*, Cambridge University Press, 1998.
[11] L. VAN DEN DRIES AND C. MILLER, *Geometric categories and o-minimal structures*, Duke Math. J., 84 (1996), pp. 497–540.
[12] F. FACCHINEI AND J.-S. PANG, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Volumes I and II, Springer, New York, 2003.
[13] I.M. GELFAND AND M. ZEJTLIN, *Printszip nelokalnogo poiska v sistemah avtomatich. Optimizatsii*, Dokl. AN SSSR, 137 (1961), pp. 295–298 (in Russian).
[14] L.M. GRAÑA DRUMMOND AND Y. PETERZIL, *The central path in smooth convex semidefinite programs*, Optimization, 51 (2002), pp. 207–233.
[15] V. GUILLEMIN AND A. POLLACK, *Differential Topology*, Prentice Hall, Englewood Cliffs, NJ, 1974.
[16] M. HALICKA, E. DE KLERK, AND C. ROOS, *On the convergence of the central path in semidefinite optimization*, SIAM J. Optim., 12 (2002), pp. 1090–1099.
[17] A.D. IOFFE, *Metric regularity and subdifferential calculus*, Uspehi Mat. Nauk, 55 (2000), pp. 103–162 (in Russian), English translation: Russian Math. Surveys, 55 (2000), pp. 501–558.
[18] A.D. IOFFE, *Critical values of set-values maps with stratifiable graphs. Extensions of Sard and Smale-Sard theorems*, Proc. AMS, 136 (2008), pp. 3111–3119.
[19] A.D. IOFFE, *Typical normality of optimization problems with tame data*, Pacific J. Optim., to appear.
[20] A.D. IOFFE AND V.M. TIKHOMIROV, *Theory of Extremal Problems*, "Nauka", Moscow 1974, English translation: North Holland, Amsterdam, 1979.
[21] K. KURDYKA, *On gradients of functions definable in o-minimal structures*, Ann. Inst. Fourier, 48 (1998), pp. 769–783.

[22] C. Lemarechal, F. Oustry, and C. Sagastizabal, *The U-Lagrangian of a convex function*, Trans. Amer. Math. Soc., 352 (2000), pp. 711–729.
[23] S. Łojasiewicz, *Sur la géométrie semi- et sous-analytique*, Ann. Inst. Fourier, 43 (1993), pp. 1575–1595.
[24] A.S. Lewis, *Active sets, nonsmoothness and sensitivity*, SIAM J. Optim., 13 (2002), pp. 702–725.
[25] A. Marino, C. Saccon, and M. Tosques, *Curves of maximal slope and parabolic variational inequalities on non-convex constraints*, Ann. Scuola Normale Pisa, XVI (1989), pp. 281–330.
[26] Y. Nesterov and A.S. Nemirovski, *Interior point polynomial algorithms in convex programming*, SIAM Studies in Appl. Math, 13 (1994).
[27] A. Norton, *Functions not constant on fractal quasi-arcs of critical points*, Proc. Amer. Math. Soc., 106 (1989), pp. 397–405.
[28] R.T. Rockafellar and R.J.B. Wets, *Variational Analysis*, Springer, New York, 1998.
[29] J.E. Spingarn and R.T. Rockafellar, *The generic nature of optimality conditions in nonlinear programming*, Math. Oper. Res., 4 (1979), pp. 425–430.
[30] R.B. Vinter, *Optimal Control*, Birkhauser, Boston, Cambridge, MA, 2000.
[31] H. Whitney, *A function not constant on a connected set of critical points*, Duke Math. J., 1 (1935), pp. 514–517.
[32] Y. Yomdin and G. Comte, *Tame Geometry with Applications to Smooth Analysis*, Springer LNM series, vol. 1834, 2004.

SIAM J. OPTIM.
Vol. 19, No. 4, pp. 1918–1946

A POLYNOMIAL PREDICTOR-CORRECTOR TRUST-REGION ALGORITHM FOR LINEAR PROGRAMMING*

GUANGHUI LAN†, RENATO D. C. MONTEIRO†, AND TAKASHI TSUCHIYA‡

Abstract. In this paper we present a scaling-invariant, interior-point, predictor-corrector type algorithm for linear programming (LP) whose iteration-complexity is polynomially bounded by the dimension and the logarithm of a certain condition number of the LP constraint matrix. At the predictor stage, the algorithm either takes the step along the standard affine scaling (AS) direction or a new trust-region type direction, whose construction depends on a scaling-invariant bipartition of the variables determined by the AS direction. This contrasts with the layered least squares direction introduced in S. Vavasis and Y. Ye [*Math. Program.*, 74 (1996), pp. 79–120], whose construction depends on multiple-layered partitions of the variables that are not scaling-invariant. Moreover, it is shown that the overall arithmetic complexity of the algorithm (weakly) depends on the right-hand side and the cost of the LP in view of the work involved in the computation of the trust region steps.

Key words. interior-point algorithms, primal-dual algorithms, path-following, trust-region, central path, layered steps, condition number, polynomial complexity, predictor-corrector, affine scaling, strongly polynomial, linear programming

AMS subject classifications. 65K05, 68Q25, 90C05, 90C51, 90C60

DOI. 10.1137/070693461

1. Introduction.

We consider the linear programming (LP) problem

$$\begin{array}{ll} \text{minimize}_x & c^T x \\ \text{subject to} & Ax = b,\ x \geq 0 \end{array} \tag{1}$$

and its associated dual problem

$$\begin{array}{ll} \text{maximize}_{(y,s)} & b^T y \\ \text{subject to} & A^T y + s = c,\ s \geq 0, \end{array} \tag{2}$$

where $A \in \Re^{m\times n}, c \in \Re^n$, and $b \in \Re^m$ are given, and the vectors $x, s \in \Re^n$, and $y \in \Re^m$ are the unknown variables.

Karmarkar in his seminal paper [4] proposed the first polynomially convergent interior-point method with an $\mathcal{O}(nL)$ iteration-complexity bound, where L is the size of the LP instance (1). The first path-following interior-point algorithm was proposed by Renegar in his breakthrough paper [17]. Renegar's method closely follows the primal central path and exhibits an $\mathcal{O}(\sqrt{n}L)$ iteration-complexity bound. The first path-following algorithm that simultaneously generates iterates in both the primal and dual spaces has been proposed by Kojima, Mizuno, and Yoshise [5] and Tanabe [19], based on ideas suggested by Megiddo [7]. In contrast to Renegar's algorithm, Kojima et al.'s algorithm has an $\mathcal{O}(nL)$ iteration-complexity bound. A primal-dual path-following with an $\mathcal{O}(\sqrt{n}L)$ iteration-complexity bound was subsequently obtained by

*Received by the editors June 1, 2007; accepted for publication (in revised form) November 6, 2008; published electronically February 27, 2009.

http://www.siam.org/journals/siopt/19-4/69346.html

†School of ISyE, Georgia Institute of Technology, Atlanta, Georgia 30332 (glan@isye.gatech.edu, monteiro@isye.gatech.edu). The first author was supported by NSF Grant CCR-0430644 and ONR Grants N00014-05-1-0183 and N00014-08-1-0033. The second author was supported in part by NSF Grants CCR-0430644 and CCF-0808863 and ONR Grant N00014-05-1-0183 and N00014-08-1-0033.

‡The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-Ku, Tokyo, 106-8569, Japan (tsuchiya@sun312.ism.ac.jp).

Kojima, Mizuno, and Yoshise [6] and Monteiro and Adler [11, 12] independently. Following these developments, many other primal-dual interior-point algorithms for linear programming have been proposed.

An outstanding open problem in optimization is whether there exists a strongly polynomial algorithm for linear programming, that is one whose complexity is bounded by a polynomial of m and n only. A major effort in this direction is due to Tardos [20] who developed a polynomial-time algorithm whose complexity is bounded by a polynomial of m, n, and L_A, where L_A denotes the size of A. Such an algorithm gives a strongly polynomial method for the important class of linear programming problems where the entries of A are either 1, -1, or 0, e.g., LP formulations of network flow problems. Tardos' algorithm consists of solving a sequence of "low-sized" LP problems by a standard polynomially convergent LP method and using their solutions to obtain the solution of the original LP problem.

The development of a method which works entirely in the context of the original LP problem and whose complexity is also bounded by a polynomial of m, n, and L_A is due to Vavasis and Ye [28]. Their method is a primal-dual, path-following, interior-point algorithm similar to the ones mentioned above except that it uses once in a while a crucial step, namely the least layered square (LLS) direction. They showed that their method has an $O(n^{3.5}(\log \bar{\chi}_A + \log n))$ iteration-complexity bound, where $\bar{\chi}_A$ is a condition number associated with A having the property that $\log \bar{\chi}_A = O(L_A)$. The number $\bar{\chi}_A$ was first introduced implicitly by Dikin [2] in the study of primal affine scaling (AS) algorithms, and was later studied by several researchers including Vanderbei and Lagarias [27], Todd [21], and Stewart [18]. Properties of $\bar{\chi}_A$ are studied in [3, 25, 26].

The complexity analysis of Vavasis and Ye's algorithm is based on the notion of crossover event, a combinatorial event concerning the central path. Intuitively, a crossover event occurs between two variables when one of them is larger than the other at a point in the central path and then becomes smaller asymptotically as the optimal solution set is approached. Vavasis and Ye showed that there can be at most $n(n-1)/2$ crossover events and that a distinct crossover event occurs every $O(n^{1.5}(\log \bar{\chi}_A + \log n))$ iterations, from which they deduced the overall $O(n^{3.5}(\log \bar{\chi}_A + \log n))$ iteration-complexity bound. In [10], an LP instance is given where the number of crossover events is $\Theta(n^2)$.

One difficulty of Vavasis and Ye's method is that it requires the explicit knowledge of $\bar{\chi}_A$ in order to determine a partition of the variables into layers used in the computation of the LLS step. This difficulty was remedied in a variant proposed by Megiddo, Mizuno, and Tsuchiya [8] which does not require the explicit knowledge of the number $\bar{\chi}_A$. They observed that at most n types of partitions arise as $\bar{\chi}_A$ varies from 1 to ∞, and that one of these can be used to compute the LLS step. Based on this idea, they developed a variant which computes the LLS steps for all these partitions and picks the one that yields the greatest duality gap reduction at the current iteration. Another approach that also remedies the above difficulty was proposed by Monteiro and Tsuchiya [14]. Their algorithm computes only one LLS step per iteration without any explicit knowledge of $\bar{\chi}_A$. This method is a predictor-corrector type algorithm like the one described in [9] except that at the predictor stage it takes a step along either the primal-dual AS step or the LLS step. In contrast to the LLS step used in Vavasis and Ye's algorithm, the partition of variables used for computing the LLS step is constructed from the information provided by the AS direction and hence does not require any knowledge on $\bar{\chi}_A$. Both of these variants ([8], [14]) have exactly the same overall complexity as Vavasis and Ye's algorithm.

Another disadvantage associated with Vavasis and Ye's algorithm, as well as its variants in [8] and [14], is that they are not scaling-invariant under the change of variables $(x, y, s) = (D\tilde{x}, \tilde{y}, D^{-1}\tilde{s})$, where D is a positive diagonal matrix. Hence, when these algorithms are applied to the scaled pair of LP problems, the number of iterations performed by it generally changes and is now bounded by $O(n^{3.5}\log(\bar{\chi}_{AD} + n))$, as AD is the coefficient matrix for the scaled pair of LP problems. On the other hand, using the notion of crossover events, LLS steps and a few other nontrivial ideas, Monteiro and Tsuchiya [15] have shown that, for the Mizuno-Todd-Ye predictor-corrector (MTY P-C) algorithm, the number of iterations needed to approximately traverse the central path from μ_0 to μ_f is bounded by $\mathcal{O}(n^{3.5}\log(\bar{\chi}_A^* + n) + T(\mu_0/\mu_f))$, where $\bar{\chi}_A^*$ is the infimum of $\bar{\chi}_{AD}$ as D varies over the set of positive diagonal matrices and $T(t) \equiv \min\{n^2\log(\log t), \log t\}$ for all $t > 0$. The condition number $\bar{\chi}_A^*$ is clearly scaling-invariant and the ratio $\bar{\chi}_A^*/\bar{\chi}_A$, as a function of A, can be arbitrarily small (see [15]). Hence, while the iteration-complexity obtained in [15] for the MTY P-C algorithm has the extra term $T(\mu_0/\mu_f)$, its first term can be considerably smaller than the bound obtained by Vavasis and Ye. Also note that, as μ_0/μ_f grows to ∞, the iteration-complexity bound obtained in [15] is smaller than the classical iteration-complexity bound of $\mathcal{O}(\sqrt{n}\log(\mu_0/\mu_f))$ established in [9] for the MTY P-C algorithm.

An interesting open problem is whether one can develop a scaling-invariant interior-point algorithm for linear programming whose iteration-complexity and arithmetic-complexity are bounded by a polynomial of n and $\log\bar{\chi}_A^*$. In this paper, we partially answer the above question by presenting a predictor-corrector type algorithm, referred to as the predictor-corrector trust-region (PC-TR) algorithm, which has $\mathcal{O}(n^{3.5}\log(\bar{\chi}_A^* + n))$ iteration-complexity bound. It is a predictor-corrector algorithm similar to the one developed in [9] except that, at the predictor stage, it takes a step along either the AS direction or a trust-region (TR) type step. Unlike the LLS direction used in the predictor-corrector algorithm of [14], the TR direction depends on a scaling-invariant bipartition of the variables and hence it is a scaling-invariant direction. Its iteration can be briefly described as follows. First, the AS direction is computed and a test involving this direction is performed to determine whether the TR step is needed. If the TR direction is not needed, a step along the AS direction, followed by a standard corrector step, is taken as usual. Otherwise, the AS direction determines a scaling-invariant bipartition of the variables which allows to construct a pair of primal and dual trust region subproblems whose optimal solutions yield the TR direction. Then the algorithm takes a step along either the AS or the TR direction whichever yields the largest duality gap reduction. Moreover, we show that the overall arithmetic complexity of the PC-TR algorithm (weakly) depends also on b and c due to work involved in the computation of the trust region steps.

The organization of the paper is as follows. Section 2 consists of six subsections. In subsection 2.1, we review the notion of the primal-dual central path and its associated two norm neighborhoods. Subsection 2.2 introduces the notion of the condition number $\bar{\chi}_A$ of a matrix A and describes the properties of $\bar{\chi}_A$ that will be useful in our analysis. Subsection 2.3 reviews the AS step and the corrector (or centrality) step which are the basic ingredients of several well-known, interior-point algorithms. Subsection 2.4 motivates and formally introduces the TR step. Subsection 2.5 describes an interior-point algorithm based on these TR steps, which we refer to as the predictor-corrector trust-region (PC-TR) algorithm, and states one of main results of this paper which gives an upper bound on the iteration-complexity of the PC-TR algorithm. Subsection 2.6 introduces a variant of the PC-TR algorithm with the same iteration-complexity as the latter one and discusses a procedure for computing the TR

steps used by this variant. It also states the other main result of this paper regarding the overall arithmetic complexity of the above variant of the PC-TR algorithm. Section 3, which consists of three subsections, introduces some basic tools which are used in our convergence analysis. Subsection 3.1 discusses the notion of crossover events. Subsection 3.2 introduces the LLS direction and states an approximation result that provides an estimation of the closeness between the AS direction and the LLS direction. Subsection 3.3 reviews an important result, which basically provides sufficient conditions for the occurrence of crossover events. Section 4 is dedicated to the proof of the main result stated in subsection 2.5. Section 5 provides the proof of the other main result stated in subsection 2.6 regarding the arithmetic complexity of the variant of the PC-TR algorithm. Finally, the Appendix gives the proof of an important lemma used in subsection 2.4 to motivate the definition of the TR step.

The following notation is used throughout our paper. We denote the vector of all ones by e. Its dimension is always clear from the context. The symbols $\Re^n$, $\Re^n_+$, and $\Re^n_{++}$ denote the n-dimensional Euclidean space, the nonnegative orthant of $\Re^n$, and the positive orthant of $\Re^n$, respectively. The set of all $m \times n$ matrices with real entries is denoted by $\Re^{m\times n}$. If J is a finite index set, then $|J|$ denotes its cardinality, that is the number of elements of J. For $J \subseteq \{1,\ldots,n\}$ and $w \in \Re^n$, we let w_J denote the subvector $[w_i]_{i\in J}$; moreover, if E is an $m \times n$ matrix, then E_J denotes the $m \times |J|$ submatrix of E corresponding to J. For a vector $w \in \Re^n$, we let $\max(w)$ and $\min(w)$ denote the largest and the smallest component of w, respectively; $\mathrm{Diag}(w)$ denote the diagonal matrix whose ith diagonal element is w_i for $i = 1,\ldots,n$; and for an arbitrary $\alpha \in \Re$, w^α denote the vector $[\mathrm{Diag}(w)]^\alpha e$ whenever it is well-defined. For two vectors $u, v \in \Re^n$, uv denotes their Hadamard product, i.e., the vector in $\Re^n$ whose ith component is $u_i v_i$. The Euclidean norm, the 1-norm, and the ∞-norm are denoted by $\|\cdot\|$, $\|\cdot\|_1$, and $\|\cdot\|_\infty$, respectively. For a matrix E, $\mathrm{Im}(E)$ denotes the subspace generated by the columns of E, and $\mathrm{Ker}(E)$ denotes the subspace orthogonal to the rows of E. The superscript ^{T} denotes transpose.

2. The problem and algorithm. In this section we propose a predictor-corrector, primal-dual, interior-point algorithm with trust-region steps for solving linear programming (1) and (2). We also present the main convergence results for the algorithm. One result establishes a polynomial iteration-complexity bound, namely, $O(n^{3.5}\log(\bar\chi_A^* + n + \varepsilon_0^{-1}))$, where ε_0 is a constant and $\bar\chi_A^*$ is a certain scaling-invariant condition number associated with the constraint matrix A, and the other result establishes a polynomial arithmetic complexity bound for the algorithm.

2.1. The central path. In this subsection we introduce the pair of primal and dual linear programs and the assumptions used in our development. We also describe the associated primal-dual central path and its corresponding two-norm neighborhoods.

Given $A \in \Re^{m\times n}$, $c \in \Re^n$, and $b \in \Re^m$, consider the pairs of linear programs (1) and (2), where $x \in \Re^n$ and $(y,s) \in \Re^m \times \Re^n$ are their respective variables. The set of strictly feasible solutions for these problems are

$$\mathcal{P}^{++} \equiv \{x \in \Re^n : Ax = b,\ x > 0\},$$
$$\mathcal{D}^{++} \equiv \{(y,s) \in \Re^{m\times n} : A^T y + s = c,\ s > 0\},$$

respectively. Throughout the paper we make the following assumptions on the pair of problems (1) and (2).

A.1 $\mathcal{P}^{++}$ and $\mathcal{D}^{++}$ are nonempty.

A.2 The rows of A are linearly independent.

Under the above assumptions, it is well known that for any $\nu > 0$ the system,

$$xs = \nu e, \tag{3}$$

$$Ax = b, \quad x > 0, \tag{4}$$

$$A^T y + s = c, \quad s > 0, \tag{5}$$

has a unique solution (x, y, s), which we denote by $(x(\nu), y(\nu), s(\nu))$. The central path is the set consisting of all these solutions as ν varies in $(0, \infty)$. As ν converges to zero, the path $(x(\nu), y(\nu), s(\nu))$ converges to a primal-dual optimal solution (x^*, y^*, s^*) for problems (1) and (2). Given a point $w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$, its duality gap and its normalized duality gap are defined as $x^T s$ and $\mu = \mu(x, s) \equiv x^T s/n$, respectively, and the point $(x(\mu), y(\mu), s(\mu))$ is said to be the central point associated with w. Note that $(x(\mu), y(\mu), s(\mu))$ also has normalized duality gap μ. We define the proximity measure of a point $w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ with respect to the central path by

$$\eta(w) \equiv \|xs/\mu - e\|.$$

Clearly, $\eta(w) = 0$ if and only if $w = (x(\mu), y(\mu), s(\mu))$, or equivalently w coincides with its associated central point. The two-norm neighborhood of the central path with opening $\beta > 0$ is defined as

$$\mathcal{N}(\beta) \equiv \{w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++} : \eta(w) \le \beta\}.$$

Finally, for any point $w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$, we define

$$\delta(w) \equiv s^{1/2} x^{-1/2} \in \Re^n. \tag{6}$$

The following propositions provide important estimates which are used throughout our analysis.

PROPOSITION 2.1. *For every $0 < \nu_1 \le \nu_2$, we have*

$$s(\nu_1) \le n s(\nu_2) \quad \text{and} \quad x(\nu_1) \le n x(\nu_2). \tag{7}$$

Proof. Please refer to Lemma 16 of Vavasis and Ye [28]. □

PROPOSITION 2.2. *Let $w = (x, y, s) \in \mathcal{N}(\beta)$ for some $\beta \in (0, 1)$ be given and define $\delta \equiv \delta(w)$. Let $w(\mu) = (x(\mu), y(\mu), s(\mu))$ be the central point associated with w. Then:*

$$\frac{1-\beta}{(1+\beta)^{1/2}}\,\delta \le \frac{s(\mu)}{\sqrt{\mu}} \le \frac{(1+\beta)^{1/2}}{1-\beta}\,\delta.$$

Proof. This result is summarized in Proposition 2.1 in [14]. □

2.2. Condition number. In this subsection we define a certain condition number associated with the constraint matrix A and state the properties of $\bar{\chi}_A$ which will play an important role in our analysis.

Let $\mathcal{D}$ denote the set of all positive definite $n \times n$ diagonal matrices and define

$$\bar{\chi}_A \equiv \sup\Big\{ \|A^T (A\tilde{D}A^T)^{-1} A\tilde{D}\| \ : \tilde{D} \in \mathcal{D} \Big\}$$

$$= \sup\Big\{ \tfrac{\|A^T y\|}{\|c\|} : y = \operatorname{argmin}_{\tilde{y} \in \Re^n} \|\tilde{D}^{1/2}(A^T \tilde{y} - c)\| \text{ for some } 0 \ne c \in \Re^n \text{ and } \tilde{D} \in \mathcal{D} \Big\}. \tag{8}$$

The parameter $\bar{\chi}_A$ plays a fundamental role in the complexity analysis of algorithms for linear programming and least square problems (see [28] and references therein). Its finiteness has been firstly established by Dikin and Zorkalcev [2]. Other authors have also given alternative derivations of the finiteness of $\bar{\chi}_A$ (see, for example, Stewart [18], Todd [21], and Vanderbei and Lagarias [27]).

We summarize in the next proposition a few important facts about the parameter $\bar{\chi}_A$.

PROPOSITION 2.3. *Let $A \in \Re^{m\times n}$ with full row rank be given. Then, the following statements hold:*

(a) $\bar{\chi}_{GA} = \bar{\chi}_A$ *for any nonsingular matrix* $G \in \Re^{m\times m}$;

(b) $\bar{\chi}_A = \max\{\|G^{-1}A\| : G \in \mathcal{G}\}$ *where* $\mathcal{G}$ *denotes the set of all* $m\times m$ *nonsingular submatrices of* A;

(c) *if the* $m \times m$ *identity matrix is a submatrix of* A *and* $\tilde{A}$ *is an* $r \times n$ *submatrix of* A, *then* $\|\tilde{G}^{-1}\tilde{A}\| \le \bar{\chi}_A$ *for every* $r \times r$ *nonsingular submatrix* $\tilde{G}$ *of* $\tilde{A}$.

Proof. Statement (a) readily follows from the definition (8). The inequality $\bar{\chi}_A \ge \max\{\|G^{-1}A\| : G \in \mathcal{G}\}$ is established in Lemma 3 of [28] while the proof of the reverse inequality is given in [21] (see also Theorem 1 of [22]). Hence, (b) holds. A proof of (c) can be found in [14]. □

The condition number $\bar{\chi}_A^*$, defined by taking the infimum of the condition number $\bar{\chi}_{AD}$ as D varies over the set of positive diagonal matrices, that is, $\bar{\chi}_A^* \equiv \inf\{\bar{\chi}_{AD} : D \in \mathcal{D}\}$, also plays an important role in the convergence analysis for our algorithm. Note that by definition, $\bar{\chi}_A^*$ is a scaling-invariant quantity.

2.3. Predictor-corrector step. In this subsection we describe the well-known predictor-corrector (P-C) iteration which is used by several interior-point algorithms (see for example Mizuno et al. [9]). We also describe the properties of this iteration which will be used in our analysis.

The P-C iteration consists of two steps, namely the predictor (or AS) step and the corrector (or centrality) step. The search direction used by either step from a current point in $(x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ is the solution of the following linear system of equations

$$\begin{aligned} S\Delta x + X\Delta s &= \sigma\mu e - xs, \\ A\Delta x &= 0, \\ A^T\Delta y + \Delta s &= 0, \end{aligned} \tag{9}$$

where $\mu = \mu(x, s)$ and $\sigma \in \Re$ is a prespecified parameter, commonly referred to as the centrality parameter. When $\sigma = 0$, we denote the solution of (9) by $(\Delta x^{\mathrm{a}}, \Delta y^{\mathrm{a}}, \Delta s^{\mathrm{a}})$ and refer to it as the primal-dual affine scaling direction at w; it is the direction used in the predictor step of the P-C iteration. When $\sigma = 1$, we denote the solution of (9) by $(\Delta x^{\mathrm{c}}, \Delta y^{\mathrm{c}}, \Delta s^{\mathrm{c}})$ and refer to it as the corrector direction at w; it is the direction used in the corrector step of the P-C iteration.

We are now ready to describe the entire predictor-corrector iteration. Suppose that a constant $\beta \in (0, 1/4]$ and a point $w = (x, y, s) \in \mathcal{N}(\beta)$ is given. The P-C iteration generates another point $(x^+, y^+, s^+) \in \mathcal{N}(\beta)$ as follows. It first moves along the direction $(\Delta x^{\mathrm{a}}, \Delta y^{\mathrm{a}}, \Delta s^{\mathrm{a}})$ until it hits the boundary of the enlarged neighborhood $\mathcal{N}(2\beta)$. More specifically, it computes the point $w^{\mathrm{a}} = (x^{\mathrm{a}}, y^{\mathrm{a}}, s^{\mathrm{a}}) \equiv (x, y, s) + \alpha_{\mathrm{a}}(\Delta x^{\mathrm{a}}, \Delta y^{\mathrm{a}}, \Delta s^{\mathrm{a}})$ where

$$\alpha_{\mathrm{a}} \equiv \sup\{\alpha \in [0, 1] : (x, y, s) + \alpha(\Delta x^{\mathrm{a}}, \Delta y^{\mathrm{a}}, \Delta s^{\mathrm{a}}) \in \mathcal{N}(2\beta)\}. \tag{10}$$

Next, the P-C iteration generates a point inside the smaller neighborhood $\mathcal{N}(\beta)$ by taking a unit step along the corrector direction $(\Delta x^{\rm c}, \Delta y^{\rm c}, \Delta s^{\rm c})$ at the point $w^{\rm a}$; that is, it computes the point $(x^+, y^+, s^+) \equiv (x^{\rm a}, y^{\rm a}, s^{\rm a}) + (\Delta x^{\rm c}, \Delta y^{\rm c}, \Delta s^{\rm c}) \in \mathcal{N}(\beta)$. The successive repetition of this iteration leads to the so-called Mizuno–Todd–Ye (MTY) predictor-corrector algorithm (see [9]).

Our method is very similar to the algorithm of [9] except that it sometimes replaces the AS step by the trust-region step described in the next subsection. The insertion of the trust region step in the above MTY predictor-corrector algorithm guarantees that the modified method has the finite termination property. The trust-region step is taken only when it yields a point with a smaller duality gap than the one obtained from the AS step as described above.

In the remaining part of this subsection, we discuss some properties of the P-C iteration and the primal-dual AS direction. For a proof of the next two propositions, we refer the reader to [9].

PROPOSITION 2.4 (predictor step). *Suppose that $w = (x, y, s) \in \mathcal{N}(\beta)$ for some constant $\beta \in (0, 1/2]$. Let $\Delta w^{\rm a} = (\Delta x^{\rm a}, \Delta y^{\rm a}, \Delta s^{\rm a})$ denote the affine scaling direction at $w^{\rm a}$ and let $\alpha_{\rm a}$ be the step-size computed according to* (10)*. Then the following statements hold:*

(a) *the point $w + \alpha \Delta w^{\rm a}$ has normalized duality gap $\mu(\alpha) = (1-\alpha)\mu$ for all $\alpha \in \Re$;*

(b) *$\alpha_{\rm a} \geq \sqrt{\beta/n}$ and hence $\mu(\alpha_{\rm a})/\mu \leq 1 - \sqrt{\beta/n}$.*

PROPOSITION 2.5 (corrector step). *Suppose that $w = (x, y, s) \in \mathcal{N}(2\beta)$ for some constant $\beta \in (0, 1/4]$ and let $(\Delta x^{\rm c}, \Delta y^{\rm c}, \Delta s^{\rm c})$ denote the corrector step at w. Then, $w + \Delta w^{\rm c} \in \mathcal{N}(\beta)$. Moreover, the (normalized) duality gap of $w + \Delta w^{\rm c}$ is the same as that of w.*

For the sake of future usage, we mention the following alternative characterization of the primal-dual AS direction whose verification is straightforward:

$$\Delta x^{\rm a} \equiv \mathrm{argmin}_{p \in \Re^n} \left\{ \|\delta(x+p)\|^2 : Ap = 0 \right\}, \tag{11}$$

$$(\Delta y^{\rm a}, \Delta s^{\rm a}) \equiv \mathrm{argmin}_{(r,q) \in \Re^m \times \Re^n} \left\{ \|\delta^{-1}(s+q)\|^2 : A^T r + q = 0 \right\}, \tag{12}$$

where $\delta \equiv \delta(w)$. For a search direction $(\Delta x, \Delta y, \Delta s)$ at a point (x, y, s), the quantity

$$\begin{aligned} (Rx(w), Rs(w)) &\equiv \left(\frac{\delta(x+\Delta x)}{\sqrt{\mu}}, \frac{\delta^{-1}(s+\Delta s)}{\sqrt{\mu}} \right) \\ &= \left(\frac{x^{1/2}s^{1/2} + \delta \Delta x}{\sqrt{\mu}}, \frac{x^{1/2}s^{1/2} + \delta^{-1}\Delta s}{\sqrt{\mu}} \right) \end{aligned} \tag{13}$$

appears quite often in our analysis. We refer to it as the *residual* of $(\Delta x, \Delta y, \Delta s)$. Note that if $(Rx^{\rm a}(w), Rs^{\rm a}(w))$ is the residual of $(\Delta x^{\rm a}, \Delta y^{\rm a}, \Delta s^{\rm a})$, then

$$Rx^{\rm a}(w) = -\frac{1}{\sqrt{\mu}} \delta^{-1} \Delta s^{\rm a}, \quad Rs^{\rm a}(w) = -\frac{1}{\sqrt{\mu}} \delta \Delta x^{\rm a}, \tag{14}$$

and

$$Rx^{\rm a}(w) + Rs^{\rm a}(w) = \frac{x^{1/2}s^{1/2}}{\sqrt{\mu}}, \tag{15}$$

due to the fact that $(\Delta x^{\rm a}, \Delta y^{\rm a}, \Delta s^{\rm a})$ satisfies the first equation in (9) with $\sigma = 0$. The following quantity is used in the test to determine when the trust-region step should be used in place of the AS step:

$$\varepsilon_\infty^{\rm a}(w) \equiv \max_i \left\{ \min \left\{ |Rx_i^{\rm a}(w)|, |Rs_i^{\rm a}(w)| \right\} \right\}. \tag{16}$$

We end this section by providing some estimates involving the residual of the AS direction.

LEMMA 2.6. *Suppose that $w = (x, y, s) \in \mathcal{N}(\beta)$ for some $\beta \in (0, 1/4]$. Then, for all $i = 1, \ldots, n$, we have*

$$\max\{|Rx_i^{\mathrm{a}}(w)|, |Rs_i^{\mathrm{a}}(w)|\} \geq \frac{\sqrt{1-\beta}}{2} \geq \frac{1}{4}.$$

Proof. Assume for a contradiction that for some $i \in \{1, \ldots, n\}$, $\max\{|Rx_i^{\mathrm{a}}(w)|, |Rs_i^{\mathrm{a}}(w)|\} < \sqrt{1-\beta}/2$. Then, using (15), we obtain the following contradiction:

$$\frac{x_i^{1/2} s_i^{1/2}}{\sqrt{\mu}} = Rx_i^{\mathrm{a}}(w) + Rs_i^{\mathrm{a}}(w) \leq |Rx_i^{\mathrm{a}}(w)| + |Rs_i^{\mathrm{a}}(w)| < \sqrt{1-\beta} \leq \frac{x_i^{1/2} s_i^{1/2}}{\sqrt{\mu}}. \qquad \square$$

2.4. Trust region step. In this subsection we introduce a new type of search step, namely, the trust-region (TR) step, and describe some properties about it.

The definition of the TR step is motivated by the following result regarding the duality gap reduction obtained by moving along a search direction satisfying certain conditions. This result can be viewed as a generalization of Lemma 4.6 in [14], and its proof is given in the Appendix.

LEMMA 2.7. *Let $w \in \mathcal{N}(\beta)$ with $\beta \in (0, 1/2]$ and a direction $\Delta w = (\Delta x, \Delta y, \Delta s)$ satisfying $A\Delta x = 0$ and $A^T \Delta y + \Delta s = 0$ be given. Then, for any positive scalar γ satisfying*

$$\left(4\sqrt{2} + \sqrt{2(1+\beta)}\right)\gamma \leq \frac{\beta - 2\beta^2}{1 + 2\beta} \tag{17}$$

and for any bipartition (B, N) of $\{1, 2, \ldots, n\}$, the condition

$$\max\left\{\frac{\|\delta_B \Delta x_B\|}{\sqrt{\mu}}, \frac{\|\delta_N^{-1} \Delta s_N\|}{\sqrt{\mu}}\right\} \leq \gamma \tag{18}$$

implies that

$$\frac{\mu(w + \alpha_\tau \Delta w)}{\mu(w)} \leq \frac{\sqrt{1+\beta} + \gamma}{2\gamma} \max\{\|Rx_N\|, \|Rs_B\|\}, \tag{19}$$

where $\alpha_\tau \equiv \sup\{\alpha \in [0, 1] : w + \alpha \Delta w \in \mathcal{N}(2\beta)\}$ and $(Rx(w), Rs(w))$ is defined in (13).

A trivial application of Lemma 2.7 is as follows. Let (B, N) be the AS-bipartition at w, i.e.,

$$\begin{aligned} B = B(w) &\equiv \{i : |Rs_i^a(w)| \leq |Rx_i^a(w)|\}, \\ N = N(w) &\equiv \{i : |Rs_i^a(w)| > |Rx_i^a(w)|\}. \end{aligned} \tag{20}$$

and let

$$\varepsilon_2^{\mathrm{a}}(w) := \max\{\|Rx_N^a\|, \|Rs_B^a\|\}. \tag{21}$$

Then, in view of Lemma 2.7 and identity (14), the condition $\varepsilon_2^{\mathrm{a}}(w) \leq \gamma$ implies that

$$\frac{\mu(w + \alpha \Delta w^{\mathrm{a}})}{\mu(w)} \leq \frac{\sqrt{1+\beta} + \gamma}{2\gamma} \varepsilon_2^{\mathrm{a}}(w),$$

showing that the smaller the quantity $\varepsilon_2^{\mathrm{a}}(w)$ is, the larger the reduction of the duality gap will be, as it moves from w along the AS direction $\Delta w^{\mathrm{a}}(w)$.

However, a more interesting application of Lemma 2.7 is towards deriving a new scaling-invariant search direction, which we refer to as the trust-region direction (see the definition below). This direction is the one which minimizes the right-hand side of (19) subject to the condition (18) when $(B, N) = (B(w), N(w))$. To define this direction, set $(B, N) = (B(w), N(w))$ and consider the two subproblems:

$$\begin{array}{ll} \text{minimize} & \|\delta_N(x_N + \Delta x_N)\| \\ \text{subject to} & \|\delta_B \Delta x_B\|/\sqrt{\mu} \le \gamma_p \\ & A\Delta x = 0 \end{array} \tag{22}$$

and

$$\begin{array}{ll} \text{minimize} & \|\delta_B^{-1}(s_B + \Delta s_B)\| \\ \text{subject to} & \|\delta_N^{-1}\Delta s_N\|/\sqrt{\mu} \le \gamma_d \\ & A^T\Delta y + \Delta s = 0 \end{array} . \tag{23}$$

DEFINITION. *Given $w \in \mathcal{N}(\beta)$ and positive scalars γ_p and γ_d, let $\Delta x^\tau(w;\gamma_p)$ and $(\Delta y^\tau(w;\gamma_d), \Delta s^\tau(w;\gamma_d))$ denote optimal solutions of subproblems* (22) *and* (23)*, respectively. The direction $\Delta w^\tau(w;\gamma_p,\gamma_d) \equiv (\Delta x^\tau(w;\gamma_p), \Delta y^\tau(w;\gamma_d), \Delta s^\tau(w;\gamma_d))$ is then referred to as a trust-region direction at w with radius pair (γ_p,γ_d).*

We now make a few observations regarding the above definition. First, it can be easily shown that both subproblems (22) and (23) must have optimal solutions although their optimal solutions are not necessarily unique. We will refer to any pair of optimal solutions of subproblems (22) and (23) as a trust-region step corresponding to the triple $(w;\gamma_p,\gamma_d)$. Second, if $\varepsilon_2^{\mathrm{a}}(w) \le \min\{\gamma_p,\gamma_d\}$, then the quantity $\varepsilon_2^\tau(w;\gamma_p,\gamma_d)$ defined as

$$\varepsilon_2^\tau(w;\gamma_p,\gamma_d) := \max\left\{\|Rx_N^\tau(w)\|, \|Rs_B^\tau(w)\|\right\}, \tag{24}$$

where $(Rx^\tau(w), Rs^\tau(w))$ denotes the residual pair for the the TR direction $\Delta w^\tau(w;\gamma_p,\gamma_d)$, satisfies $\varepsilon_2^\tau(w;\gamma_p,\gamma_p) \le \varepsilon_2^{\mathrm{a}}(w)$. In other words, whenever the AS direction is a reasonably good direction in the sense that $\varepsilon_2^{\mathrm{a}}(w)$ is sufficiently small, then the TR step is likely to be an even better direction in that it makes the right-hand side of (19) smaller. Third, even though our definition of a TR step does not uniquely characterize it, one can easily modify the definition to make it uniquely defined in the following way. Without loss of generality, we consider only the primal direction, which previously was defined as an optimal solution of (22). This clearly implies that Δx_N^τ is uniquely defined. Now, minimizing the quantity $\|\delta_B\Delta x_B\|$ under the condition that $A\Delta x = 0$ and $\Delta x_N = \Delta x_N^\tau$ uniquely determines the component Δx_B, and hence the whole primal TR step. We note, however, that our analysis does not require that the TR step be uniquely determined and in fact works for any pair of optimal solutions of (22) and (23).

2.5. Main algorithm and the convergence results. In this subsection, we describe our algorithm, namely, the predictor-corrector trust-region (PC-TR) algorithm, to solve the linear programming problem (1) and (2), and then state the main result of this paper which guarantees the convergence of the method in a strong sense. More specifically, this result states that the outer iteration-complexity bound for our method depends only on n and the scaling-invariant condition number $\bar{\chi}_A^*$.

We start by stating our predictor-corrector trust-region algorithm.

PC-TR Algorithm:

Let $0 < \beta \le 1/4$ and $\gamma > 0$ satisfying (17), $w^0 \in \mathcal{N}(\beta)$ and a scalar $\varepsilon_0 \in (0, \gamma/3]$ be given.
Set $\mu_0 \equiv \mu(w^0)$ and $k = 0$.

1) Set $w = w^k$, compute the AS step $\Delta w^{\rm a}$ at w and the residual $\varepsilon_2^{\rm a}(w)$ as defined in (21);
2) If $\varepsilon_2^{\rm a}(w) > \varepsilon_0$, then set $w \leftarrow w + \alpha_a \Delta w^{\rm a}$, where α_a is defined as in (10) and go to 6);
3) Otherwise, compute the TR step $\Delta w^\tau = \Delta w^\tau(w; \gamma_p, \gamma_d)$, for scalars $\gamma_p, \gamma_d \in [\gamma/2, 2\gamma]$;
4) Let $w^\tau = w + \alpha_\tau \Delta w^\tau$, where $\alpha_\tau \equiv \sup\{\alpha \in [0,1] : w + \alpha \Delta w^\tau \in \mathcal{N}(2\beta)\}$;
5) If $\mu(w^\tau) < (1 - \alpha_{\rm a})\mu$, then set $w \leftarrow w^\tau$, or else set $w \leftarrow w + \alpha_{\rm a} \Delta w^{\rm a}$;
6) If $\mu(w) = 0$, then **stop**;
7) Compute the corrector step $\Delta w^{\rm c}$ at w and set $w \leftarrow w + \Delta w^{\rm c}$;
8) Set $w^{k+1} = w$, increment k by 1 and go to 1).

End

We now make a few comments about the above algorithm. In the main body of the algorithm, step 2 followed by step 7 is a standard predictor-corrector iteration of the type described in subsection 2.3. This iteration is always performed in those iterations for which $\varepsilon_2^{\rm a}(w) > \varepsilon_0(w)$. In order to save computation time, the TR step is computed only in those iterations for which the current iterate w satisfies $\varepsilon_2^{\rm a}(w) \le \varepsilon_0$. In these iterations, the algorithm performs either a standard predictor-corrector iteration or a TR-corrector iteration depending on which of the two iterations gives the lower reduction of the duality gap. This test is performed in step 5 since the term $(1-\alpha_{\rm a})\mu$ is the normalized duality gap obtained when the AS step is taken (see Proposition 2.4(a)).

For the sake of future reference, we note that (17) and the assumption that $\beta \in (0, 1/4]$ imply that

$$\gamma \le \frac{1}{20}, \quad \varepsilon_0 \le \frac{\gamma}{3} \le \frac{1}{60}. \tag{25}$$

We refer to an iteration where the TR step is computed as a *TR-iteration*. The following result is immediate from Lemma 2.7 and the definition of a TR-iteration.

PROPOSITION 2.8. *Let w be an iterate of the PC-TR algorithm and assume that the next iterate w^+ after w is obtained by means of a TR-iteration. Then,*

$$\frac{\mu(w^+)}{\mu(w)} \le \frac{\sqrt{1+\beta}+\gamma}{2\gamma}\varepsilon_2^\tau(w;\gamma_p,\gamma_d) \le \frac{\sqrt{1+\beta}+\gamma}{2\gamma}\varepsilon_2^{\rm a}(w) \le \frac{\sqrt{1+\beta}+\gamma}{2\gamma}\varepsilon_0 \le \frac{1}{4}. \tag{26}$$

Proof. First note that if the iteration from w is a TR-iteration, then we have $\varepsilon_2^{\rm a}(w) \le \varepsilon_0 \le \gamma/3 \le \min\{\gamma_p, \gamma_d\}$. The first three inequalities in (26) follow from Lemma 2.7, the previous observation, and the second observation after (23). Moreover, the last inequality in (26) follows from (25) and the fact that $\beta \le 1/4$. ∎

We have the following convergence result for the above algorithm.

THEOREM 2.9. *The PC-TR algorithm described above finds a primal-dual optimal solution $w = (x^*, s^*, y^*)$ of (1) and (2) in at most $\mathcal{O}(n^{3.5}\log(\bar{\chi}_A^* + n + \varepsilon_0^{-1}))$ iterations, of which $\mathcal{O}(n^3 \log(\bar{\chi}_A^* + n + \varepsilon_0^{-1})/\log \varepsilon_0^{-1})$ are TR-iterations. In particular, if $\varepsilon_0^{-1} = \mathcal{O}((n + \bar{\chi}_A^*)^\kappa)$ for some constant $\kappa > 0$, then the total number of iterations is bounded*

by $\mathcal{O}(n^{3.5}\log(\bar{\chi}_A^* + n))$. *Also, if* $\varepsilon_0^{-1} = \Omega((n+\bar{\chi}_A^*)^\kappa)$ *for some constant* $\kappa > 0$, *then the number of TR-iterations is bounded by* $\mathcal{O}(n^3)$.

Note that the PC-TR algorithm is scaling-invariant; i.e., if the change of variables $(x, y, s) = (D\tilde{x}, \tilde{y}, D^{-1}\tilde{s})$ for some $D \in \mathcal{D}$ is performed on the pair of problems (1) and (2) and the PC-TR algorithm is applied to the new dual pair of scaled problems, then the sequence of iterates $\tilde{w}^k$ generated satisfies $(x^k, y^k, s^k) = (D\tilde{x}^k, \tilde{y}^k, D^{-1}\tilde{s}^k)$ for all $k \geq 1$, as long as the initial iterate $\tilde{w}^0 \in \mathcal{N}(\beta)$ in the $\tilde{w}$-space satisfies $(x^0, y^0, s^0) = (D\tilde{x}^0, \tilde{y}^0, D^{-1}\tilde{s}^0)$. For this reason, the PC-TR algorithm should have an iteration-complexity bound which does not depend on the scaled space where the sequence of iterates is generated. Indeed, the iteration-complexity bound stated in Theorem 2.9 is scaling-invariant since the condition number $\bar{\chi}_A^*$ is too. It is worth noting that the PC-TR algorithm is also scaling-invariant with respect to a more strict notion of scaling invariance described in Tunçel [24], which corresponds to choosing the set $\mathcal{D}$ in the above defintion as the full automorphism group of R_+^n. Note that the latter set is larger than the set of positive diagonal maps since it contains the permutation maps, and hence it leads to a stronger notion of scaling invariance.

We note also that, to prove Theorem 2.9, it suffices to show that the the number of iterations of the PC-TR algorithm applied to (1) and (2) is bounded by $O(n^{3.5}\log(\bar{\chi}_A + n + \varepsilon_0^{-1}))$. Indeed, since in the D-scaled space, the iterates can also be viewed as being generated by the PC-TR algorithm (started from a different point), then its complexity is also bounded by

$$O(n^{3.5}\log(\bar{\chi}_{AD} + n + \varepsilon_0^{-1})), \tag{27}$$

and hence by the infimum of (27) over all $D \in \mathcal{D}$, that is, by $O(n^{3.5}\log(\bar{\chi}_A^* + n + \varepsilon_0^{-1}))$.

Finally, Theorem 2.9 does not deal with the overall arithmetic complexity of the PC-TR algorithm. This issue will be dealt with in the next subsection and section 5, where we discuss the arithmetic complexity involved in the computation of a TR step for a suitable variant of the PC-TR algorithm. Roughly speaking, we will derive a bound on the number of arithmetic operations required to compute a TR step which depends on the ratio between the current duality gap and the initial duality gap. This implies that the overall arithmetic complexity obtained in this paper for the above variant of the PC-TR algorithm depends (weakly) on b and c, though its number of iterations just depends on A as shown in Theorem 2.13.

2.6. Computing the TR step. In this subsection, we present an algorithm to compute the TR step and derive the arithmetic complexity for the PC-TR algorithm.

For the sake of simplicity, we focus our discussion on the computation of the primal TR direction. We start by introducing a search direction that is closely related to the optimal solution of (22). Given a scalar $\lambda > 0$ and $w \in \mathcal{P}^{++} \times \mathcal{D}^{++}$, consider the following direction defined as

$$\Delta x(\lambda) := \operatorname{argmin}\left\{\|\delta_N(x_N + \Delta x_N)\|^2 + \lambda\|\delta_B \Delta x_B\|^2 : A\Delta x = 0\right\}, \tag{28}$$

where $\delta \equiv \delta(w)$ and $(B, N) \equiv (B(w), N(w))$. Note that this direction is well-defined in the sense that the above optimization problem has a unique optimal solution. Now, let $\psi_p : \Re_{++} \to \Re_+$ denote the mapping given by

$$\psi_p(\lambda) \equiv \frac{\|\delta_B \Delta x_B(\lambda)\|}{\sqrt{\mu}}, \tag{29}$$

where $\mu \equiv \mu(w)$.

The following technical result can be proved regarding the functions $\Delta x(\lambda)$ and $\psi_p(\lambda)$. Note that in the discussion below, we denote the derivatives of $\Delta x(\lambda)$ and $\psi_p(\lambda)$ as $\Delta x'(\lambda)$ and $\psi_p'(\lambda)$, respectively.

LEMMA 2.10. *The following statements hold:*

(a) *The limits* $\Delta x(0) \equiv \lim_{\lambda\to 0^+} \Delta x(\lambda)$ *and* $\psi_p(0) \equiv \lim_{\lambda\to 0^+} \psi_p(\lambda)$ *exist and are given by*

$$\Delta x_N(0) = \operatorname{argmin}\left\{\|\delta_N(x_N + \Delta x_N)\|^2 : A_N \Delta x_N \in \operatorname{Im}(A_B)\right\}, \tag{30}$$

$$\Delta x_B(0) = \operatorname{argmin}\left\{\|\delta_B \Delta x_B\|^2 : A_B \Delta x_B = -A_N \Delta x_N(0)\right\}, \tag{31}$$

$$\psi_p(0) = \|\delta_B \Delta x_B(0)\|/\sqrt{\mu}. \tag{32}$$

(b) *The limit* $\Delta x_B'(0) \equiv \lim_{\lambda\to 0^+} \Delta x_B'(\lambda)$ *exists. Moreover, if* $\psi_p(0) \neq 0$, *then* $\psi_p'(0) \equiv \lim_{\lambda\to 0^+} \psi_p'(\lambda)$ *also exists;*

(c) *If* $\psi_p(0) \neq 0$, *then the function* $\psi_p(\cdot)$ *is strictly convex, strictly decreasing, and* $\lim_{\lambda\to\infty} \psi_p(\lambda) = 0$; *otherwise, if* $\psi_p(0) = 0$, *then the function* $\psi_p(\cdot)$ *is identically zero.*

(d) *If* $0 < \lambda_1 \le \lambda_2$, *then* $\psi_p(\lambda_1)/\psi_p(\lambda_2) \le \lambda_2/\lambda_1$ *(with the convention that* $0/0 = 0$*).*

We note that, in view of Lemma 2.10, the functions $\Delta x(\lambda)$ and $\psi_p(\lambda)$ can be extended to $\lambda = 0$ and their extensions are continuously differentiable at $\lambda = 0$. The following result relates the direction $\Delta x(\lambda)$ above to the primal TR direction, i.e., the optimal solutions of (22).

LEMMA 2.11. *The following statements hold:*

(a) *For any* $\lambda > 0$, $\Delta x(\lambda)$ *is an optimal solution of* (22) *with* $\gamma_p = \psi_p(\lambda)$;

(b) $\Delta x(0)$ *is an optimal solution of* (22) *for any* $\gamma_p \ge \psi_p(0)$.

Proof. (a) Using the fact that $\Delta x(\lambda)$ satisfies the optimality conditions for (28), we easily see that it also satisfies the optimality conditions, and hence is an optimal solution, of (22) with $\gamma_p = \psi_p(\lambda)$.

(b) In view of Lemma 2.10, we can pass the optimality conditions of (28) to the limit as $\lambda \downarrow 0$ to conclude that $A\Delta x(0) = 0$ and $\delta_N^2(x_N + \Delta x_N(0)) \in \operatorname{Im}(A^T)$. This together with (29) and the assumption that $\gamma_p \ge \psi_p(0)$ imply that $\Delta x(0)$ satisfies the optimality conditions, and hence is an optimal solution, of (22). ∎

Using the above results, the primal TR direction required by the algorithm can be computed as follows. Recall that the goal is to find an optimal solution of (22) for some $\gamma_p \in [\gamma/2, 2\gamma]$. We start by computing $\Delta x(0)$ and then $\psi_p(0)$. If $\psi_p(0) \le 2\gamma$, then by Lemma 2.11(b), we conclude that $\Delta x(0)$ is an optimal solution of (22) with $\gamma_p = 2\gamma$, and hence can be chosen as the required TR direction. Otherwise, if $\psi_p(0) > 2\gamma$, we search for some $\lambda_p > 0$ such that

$$\gamma/2 \le \psi_p(\lambda_p) \le 2\gamma, \tag{33}$$

which always exists in view of Lemma 2.10(c) and the fact that $\psi_p(0) > 2\gamma$.

Now, to find some $\lambda_p > 0$ satisfying (33), it suffices to determine $0 < \lambda_l \le \lambda_u$ such that

$$\psi_p(\lambda_l) \ge \gamma \ge \psi_p(\lambda_u), \tag{34}$$

$$\lambda_u/\lambda_l \le 2. \tag{35}$$

In such a case, any scalar $\lambda_p \in [\lambda_l, \lambda_u]$ satisfies (33). Indeed, by Lemma 2.10(d), we have

$$\frac{\gamma}{2} \le \frac{\lambda_l}{\lambda_u}\gamma \le \frac{\lambda_l}{\lambda_p}\psi_p(\lambda_l) \le \psi_p(\lambda_p) \le \frac{\lambda_u}{\lambda_p}\psi_p(\lambda_u) \le \frac{\lambda_u}{\lambda_l}\gamma \le 2\gamma.$$

Assuming that initial λ_l and λ_u satisfying (34) are given, a standard bisection procedure on the $(\log\lambda)$-space can then be used to determine scalars λ_l and λ_u satisfying both (34) and (35). An iteration of this bisection scheme updates λ_l or λ_u as follows. First, compute $\tilde{\lambda}$ such that $\log\tilde{\lambda} = (\log\lambda_l + \log\lambda_u)/2$, that is $\tilde{\lambda} = (\lambda_l\lambda_u)^{1/2}$. Second, if $\psi_p(\tilde{\lambda}) > \gamma$, λ_l is updated to $\tilde{\lambda}$; otherwise λ_u is updated to $\tilde{\lambda}$. It is clear that each iteration of this bisection scheme always preserves condition (34) and halves the length of the interval $[\log\lambda_l, \log\lambda_u]$. Hence, it eventually finds a pair (λ_l, λ_u) satisfying (34) and (35) in $\mathcal{O}(\log(\log(\lambda_u/\lambda_l)))$ bisection iterations, where λ_l and λ_u are the initial values of these scalars at the start of the procedure.

It remains to describe how to choose initial scalars $0 < \lambda_l \le \lambda_u$ such that (34) holds. We first focus our attention on the description of λ_l. Since $\psi_p(\lambda)$ is convex, we have $\psi_p(\lambda) \ge \psi_p(0) + \psi_p'(0)\lambda$ for every $\lambda > 0$. Hence, choosing λ_l to be the root of the linear equation $\psi_p(0) + \psi_p'(0)\lambda = \gamma$, i.e.,

$$\lambda_l = \frac{\psi_p(0) - \gamma}{|\psi_p'(0)|},$$

we conclude that $\psi_p(\lambda_l) \ge \gamma$.

The following lemma provides the needed information to obtain a lower bound on λ_l. We observe that, in spite of the notation, the quantities $\psi_p(0)$ and $\psi_p'(0)$ depend on the point $w \in \mathcal{P}^{++} \times \mathcal{D}^{++}$.

LEMMA 2.12. *Let w_0 denote the initial iterate of the PC-TR algorithm and set $\mu_0 = \mu(w_0)$. Then, for any $w \in \mathcal{N}(\beta)$ such that $\mu := \mu(w) \le \mu_0$, we have*

$$\frac{\psi_p(0)}{|\psi_p'(0)|} \ge \frac{(1-\beta)^8\mu^2}{n^4(1+\beta)^4\mu_0^2\,\bar{\chi}^2_{A\delta_0^{-1}}}. \tag{36}$$

Using the above result, a lower bound on λ_l can be obtained by observing that, under the assumption that $\psi_p(0) \ge 2\gamma$, we have

$$\lambda_l = \frac{\psi_p(0) - \gamma}{|\psi_p'(0)|} \ge \frac{\psi_p(0)}{2\,|\psi_p'(0)|} \ge \frac{(1-\beta)^8\mu^2}{2\,n^4\,(1+\beta)^4\mu_0^2\,\bar{\chi}^2_{A\delta_0^{-1}}}. \tag{37}$$

We now discuss how to choose the initial scalar λ_u satisfying (34). In contrast to the choice of λ_l, there is no clear way of choosing λ_u for an arbitrary curve $\psi_p(\lambda)$. Fortunately, the PC-TR algorithm stated in the previous subsection can be slightly modified so as to compute the TR step only when the condition $\psi_p(1) \le \gamma$ (in addition to the previously required condition that $\varepsilon_2^{\mathrm{a}}(w) > \varepsilon_0$) is satisfied. Hence, we may always choose the initial λ_u to be 1. In view of our discussion above, we conclude that the computation of a TR step in this variant of the PC-TR algorithm requires $\mathcal{O}\left(\log\log\lambda_l^{-1}\right)$ bisection steps, which is bounded by

$$\mathcal{O}\left[\log\left(\log\bar{\chi}_{A\delta_0^{-1}} + \log\frac{\mu_0}{\mu}\right)\right] \tag{38}$$

bisection steps, in view of (37).

We will now precisely discuss the variant of the PC-TR algorithm mentioned in the previous paragraph. First, we mention that the whole discussion of this subsection up to this point also applies to the computation of the dual TR direction, with the dual auxiliary direction

$$\begin{aligned}(\Delta y(\lambda), \Delta s(\lambda)) := \operatorname{argmin}\{&\|\delta_B^{-1}(s_B + \Delta s_B)\|^2 \\ &+ \lambda\|\delta_N^{-1}\Delta s_N\|^2 : A^T\Delta y + \Delta s = 0\}\end{aligned} \tag{39}$$

replacing $\Delta x(\lambda)$ and the dual curve

$$\psi_d(\lambda) \equiv \frac{\|\delta_N^{-1}\Delta s_N(\lambda)\|}{\sqrt{\mu}} \tag{40}$$

replacing $\psi_p(\lambda)$. We then have the following convergence result about a certain variant of the PC-TR algorithm.

THEOREM 2.13. *Consider the variant of the PC-TR algorithm where step* 2) *is replaced by the following step:*

2') *If* $\varepsilon_2^{\mathrm{a}}(w) > \varepsilon_0$ *and* $\max\{\psi_p(1), \psi_d(1)\} > \gamma/18$, *then set* $w \leftarrow w + \alpha_a \Delta w^{\mathrm{a}}$, *where* α_a *is defined as in* (10) *and go to* 6);

Then, the conclusions of Theorem 2.9 *also hold for the resulting variant of the PC-TR algorithm.*

We will now briefly discuss the arithmetic complexity of the above variant. We will see later in section 4 that the bisection procedure to compute a TR step takes

$$T(\mu; w_0) \equiv \mathcal{O}\left[n^3 + n\,\log\left(\log \bar{\chi}_{A\delta_0^{-1}} + \log(\mu_0/\mu)\right)\right] \tag{41}$$

arithmetic operations, since the procedure requires $\mathcal{O}[\log(\log \bar{\chi}_{A\delta_0^{-1}} + \log(\mu_0/\mu))]$ evaluations of the curves $\psi_p(\lambda)$ and $\psi_d(\lambda)$ with the first evaluation of either curve taking $\mathcal{O}(n^3)$ arithmetic operations and subsequent ones taking only $\mathcal{O}(n)$ arithmetic operations.

The above observation together with Theorem 2.13 yields the following arithmetic complexity result for the above PC-TR variant.

THEOREM 2.14. *The number of arithmetic operations performed by the variant of the PC-TR algorithm stated in Theorem* 2.13 *to find an iterate* w *such that* $\mu(w) \leq \mu_f$ *is bounded by*

$$\mathcal{O}\left[\frac{n^3\,\log(\bar{\chi}_A^* + n + \varepsilon_0^{-1})}{\log \varepsilon_0^{-1}}\, T(\mu_f, w_0) + n^{6.5}\,\log(\bar{\chi}_A^* + n + \varepsilon_0^{-1})\right],$$

where $T(\cdot,\cdot)$ *is defined in* (41). *In particular, if* $\varepsilon_0^{-1} = \Theta((n + \bar{\chi}_A^*)^\kappa)$ *for some* $\kappa > 0$, *the above arithmetic complexity bound reduces to*

$$\mathcal{O}\left[n^{6.5}\log(\bar{\chi}_A^* + n) + n^4 \log\left(\log \bar{\chi}_{A\delta_0^{-1}} + \log(\mu_0/\mu_f)\right)\right].$$

3. Basic tools. In this section we introduce the basic tools that will be used in the proof of Theorem 2.9. The analysis heavily relies on the notion of layered least squares(LLS) directions and crossover events due to Vavasis and Ye [28]. Subsection 3.1 below gives the definition of a crossover event which is slightly different than the one used in [28] and discusses some of its properties. Subsection 3.2 defines the layered least squares directions that will be used in the complexity analysis and also states an approximation result that provides an estimation of the closeness between the LLS direction with respect to a partition J of $\{1, \ldots, n\}$ and the AS direction. Subsection 3.3 reviews from a different perspective an important result from [28], namely Lemma 17 of [28], that essentially guarantees the occurrence of crossover events. Since this result is stated in terms of the residual of an LLS step, the use of the approximation result of subsection 3.2 between the AS and LLS steps allows us to obtain a similar result stated in terms of the residual of the AS direction.

3.1. Crossover events. In this subsection we discuss the notion of crossover event which plays a fundamental role in our convergence analysis.

DEFINITION. *For two indices $i, j \in \{1, \ldots, n\}$ and a constant $\mathcal{C} \geq 1$, a $\mathcal{C}$-crossover event for the pair (i, j) is said to occur on the interval $(\nu', \nu]$ if*

$$\begin{aligned} &\text{there exists } \nu_0 \in (\nu', \nu] \text{ such that } \frac{s_j(\nu_0)}{s_i(\nu_0)} \leq \mathcal{C}, \\ &\text{and, } \frac{s_j(\tilde{\nu})}{s_i(\tilde{\nu})} > \mathcal{C} \text{ for all } \tilde{\nu} \leq \nu'. \end{aligned} \tag{42}$$

Moreover, the interval $(\nu', \nu]$ is said to contain a $\mathcal{C}$-crossover event if (42) holds for some pair (i, j).

Hence, the notion of a crossover event is independent of any algorithm and is a property of the central path only. Note that in view of (3), condition (42) can be reformulated into an equivalent condition involving only the primal variable. For our purposes, we will use only (42).

We have the following simple but crucial result about crossover events.

PROPOSITION 3.1. *Let $\mathcal{C} > 0$ be a given constant. There can be at most $n(n-1)/2$ disjoint intervals of the form $(\nu', \nu]$ containing $\mathcal{C}$-crossover events.*

The notion of $\mathcal{C}$-crossover events can be used to define the notion of $\mathcal{C}$-crossover events between two iterates of the PC-TR algorithm as follows. We say that a $\mathcal{C}$-crossover event occurs between two iterates w^k and w^l, $k < l$, generated by the PC-TR algorithm if the interval $(\mu(w^l), \mu(w^k)]$ contains a $\mathcal{C}$-crossover event. Note that in view of Proposition 3.1, there can be at most $n(n-1)/2$ intervals of this type. We will show in the remaining part of this paper that there exists a constant $\mathcal{C} > 0$ with the following property: for any index k, there exists an index $l > k$ such that $l - k = \mathcal{O}(n^{1.5} \log(\bar{\chi}_A + n + \varepsilon_0^{-1}))$ and a $\mathcal{C}$-crossover event occurs between the iterates w^k and w^l of the PC-TR algorithm. Proposition 3.1 and a simple argument then show that the PC-TR algorithm must terminate within $\mathcal{O}(n^{3.5} \log(\bar{\chi}_A + n + \varepsilon_0^{-1}))$ iterations.

3.2. The layered least squares step. In this subsection we describe another type of direction, namely the layered least squares (LLS) step, which is very important in the analysis of our algorithm. This step was first introduced by Vavasis and Ye in [28]. We also describe two ordered partitions of the index set $\{1, \ldots, n\}$ that are crucial in the definition of the LLS directions.

Let $w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ and a partition $(J_1, \ldots, J_p)$ of the index set $\{1, \ldots, n\}$ be given and define $\delta \equiv \delta(w)$. The primal LLS direction $\Delta x^{\rm ll} = (\Delta x^{\rm ll}_{J_1}, \ldots, \Delta x^{\rm ll}_{J_p})$ at w with the respect to J is defined recursively according to the order $\Delta x^{\rm ll}_{J_p}, \ldots, \Delta x^{\rm ll}_{J_1}$ as follows. Assume that the components $\Delta x^{\rm ll}_{J_p}, \ldots, \Delta x^{\rm ll}_{J_{k+1}}$ have been determined. Let $\Pi_{J_k} : \Re^n \to \Re^{J_k}$ denote the projection map defined as $\Pi_{J_k}(u) = u_{J_k}$ for all $u \in \Re^n$. Then $\Delta x^{\rm ll}_{J_k} \equiv \Pi_{J_k}(L^x_k)$ where L^x_k is given by

$$\begin{aligned} L^x_k &\equiv \operatorname{Argmin}_{u \in \Re^n} \left\{ \|\delta_{J_k}(x_{J_k} + u_{J_k})\|^2 : u \in L^x_{k-1} \right\} \\ &= \operatorname{Argmin}_{u \in \Re^n} \Big\{ \|\delta_{J_k}(x_{J_k} + u_{J_k})\|^2 : u \in \operatorname{Ker}(A), \\ &\qquad\qquad u_{J_i} = \Delta x^{\rm ll}_{J_i} \ \text{ for all } i = k+1, \ldots, p \Big\}, \end{aligned} \tag{43}$$

with the convention that $L^x_0 = \operatorname{Ker}(A)$. The slack component $\Delta s^{\rm ll} = (\Delta s^{\rm ll}_{J_1}, \ldots, \Delta s^{\rm ll}_{J_p})$ of the dual LLS direction $(\Delta y^{\rm ll}, \Delta s^{\rm ll})$ at w with the respect to J is defined recursively

as follows. Assume that the components $\Delta s^{\rm ll}_{J_1}, \ldots, \Delta s^{\rm ll}_{J_{k-1}}$ have been determined. Then $\Delta s^{\rm ll}_{J_k} \equiv \Pi_{J_k}(L^s_k)$ where L^s_k is given by

$$
\begin{aligned}
L^s_k &\equiv \mathrm{Argmin}_{v\in\Re^n} \left\{ \|\delta^{-1}_{J_k}(s_{J_k} + v_{J_k})\|^2 : v \in L^s_{k-1} \right\} \\
&= \mathrm{Argmin}_{v\in\Re^n} \Big\{ \|\delta^{-1}_{J_k}(s_{J_k} + v_{J_k})\|^2 : v \in \mathrm{Im}(A^T), \\
&\qquad\qquad v_{J_i} = \Delta s^{\rm ll}_{J_i} \ \text{ for all } i = 1, \ldots, k-1 \Big\},
\end{aligned} \tag{44}
$$

with the convention that $L^s_0 = \mathrm{Im}(A^T)$. Finally, once $\Delta s^{\rm ll}$ has been determined, the component $\Delta y^{\rm ll}$ is determined from the relation $A^T \Delta y^{\rm ll} + \Delta s^{\rm ll} = 0$.

Note that (11) and (12) imply that the AS direction is a special LLS direction, namely the one with respect to the only partition in which $p = 1$. Clearly, the LLS direction at a given $w \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ depends on the partition $J = (J_1, \ldots, J_p)$ used.

A partition $J = (J_1, \ldots, J_p)$ of $\{1, \ldots, n\}$ is said to be *ordered* with respect to a fixed vector $z \in \Re^n_{++}$ if $\max(z_{J_i}) \le \min(z_{J_{i+1}})$ for all $i = 1, \ldots, p-1$. In such a case, we define the gap of J with respect to z as

$$
\mathrm{gap}(z, J) \ := \ \min_{1\le i\le p-1} \left\{ \frac{\min(z_{J_{i+1}})}{\max(z_{J_i})} \right\} \ge 1,
$$

with the convention that $\mathrm{gap}(z, J) = \infty$ if $p = 1$. We say that a partition J is ordered at $w \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ if it is ordered with respect to $z = \delta(w)$, in which case we denote the quantity $\mathrm{gap}(\delta(w), J)$ simply by $\mathrm{gap}(w, J)$. For partition $J = (J_1, \ldots, J_p)$ and a point $w \in \mathcal{P}^{++} \times \mathcal{D}^{++}$, the spread of the layer J_k with respect to w is defined as

$$
\mathrm{spr}(w, J_k) \equiv \frac{\max(\delta_{J_k}(w))}{\min(\delta_{J_k}(w))}, \quad \forall k = 1, \ldots, p.
$$

We now state how the AS direction can be well approximated by suitable LLS steps. Lemma 3.2, whose proof can be found in [14], essentially states that the larger the gap of J is, the closer the AS direction and the LLS direction with respect to J will be to one another.

LEMMA 3.2. *Let $w = (x, y, s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ and an ordered partition $J = (J_1, \ldots, J_p)$ at w be given. Define $\delta \equiv \delta(w)$ and let $\Delta w^{\rm a} = (\Delta x^{\rm a}, \Delta y^{\rm a}, \Delta s^{\rm a})$ and $\Delta w^{\rm ll} = (\Delta x^{\rm ll}, \Delta y^{\rm ll}, \Delta s^{\rm ll})$ denote the AS direction at w and the LLS direction at w with respect to J, respectively. If $\mathrm{gap}(w, J) \ge 4\, p\, \bar{\chi}_A$, then*

$$
\max \left\{ \ \left\| Rx^{\rm a}(w) - Rx^{\rm ll}(w) \right\|_\infty , \ \left\| Rs^{\rm a}(w) - Rs^{\rm ll}(w) \right\|_\infty \ \right\} \ \le \ \frac{12\sqrt{n}\ \bar{\chi}_A}{\mathrm{gap}(w, J)},
$$

where $(Rx^{\rm a}(w), Rs^{\rm a}(w))$ and $(Rx^{\rm ll}(w), Rs^{\rm ll}(w))$ denote the residual pairs for the AS direction $\Delta w^{\rm a}$ and the LLS direction $\Delta w^{\rm ll}$, respectively.

In the remainder of this subsection, we describe the two important LLS directions in the analysis of our algorithm that differs in the definition of ordered partitions. The first ordered partition is due to Vavasis and Ye [28]. Given a point $w \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ and a parameter $\bar{g} \ge 1$, this partition, which we refer to as the *VY partition*, is defined as follows. Let $(i_1, \ldots, i_n)$ be an ordering of $\{1, \ldots, n\}$ such that $\delta_{i_1} \le \ldots \le \delta_{i_n}$, where $\delta = \delta(w)$. For $k = 2, \ldots, n$, let $r_k \equiv \delta_{i_k}/\delta_{i_{k-1}}$ and define $r_1 \equiv \infty$. Let $k_1 < \ldots < k_p$ be all the indices k such that $r_k > \bar{g}$. The VY $\bar{g}$-partition J is then defined as $J = (J_1, \ldots, J_p)$, where $J_q \equiv \{i_{k_q}, i_{k_q+1}, \ldots, i_{k_{q+1}-1}\}$ for all $q = 1, \ldots, p$. More generally, given a subset $I \subset \{1, \ldots, n\}$, we can similarly define the *VY $\bar{g}$-partition* of

I at w by taking an ordering $(i_1, \ldots, i_m)$ of I satisfying $\delta_{i_1} \leq \ldots \leq \delta_{i_m}$ where $m = |I|$, defining the ratios $r_1, \ldots, r_m$ as above, and proceeding exactly as in the construction above to obtain the partition $J = (J_1, \ldots, J_p)$ of I.

It is easy to see that the following result holds for the partition J described in the previous paragraph.

PROPOSITION 3.3. *Given a subset $I \subseteq \{1, \ldots, n\}$, a point $w \in \mathcal{P}^{++} \times \mathcal{D}^{++}$, and a constant $\bar{g} \geq 1$, the VY $\bar{g}$-partition $J = (J_1, \ldots, J_p)$ of I at w satisfies* $\mathrm{gap}(w, J) > \bar{g}$ *and* $\mathrm{spr}(w, J_q) \leq \bar{g}^{|J_q|} \leq \bar{g}^n$ *for all* $q = 1, \ldots, p$.

The second-ordered partition, which is used heavily in our analysis, was introduced by Monteiro and Tsuchiya [14]. Given a point $w \in \mathcal{P}^{++} \times \mathcal{D}^{++}$. First, we compute the bipartition (B, N) of $\{1, \ldots, n\}$ according to (20). Next, an order $(i_1, \ldots, i_n)$ of the index variables is chosen such that $\delta_{i_1} \leq \ldots \leq \delta_{i_n}$. Then, the first block of consecutive indices in the n-tuple $(i_1, \ldots, i_n)$ lying in the same set B or N are placed in the first layer $\mathcal{J}_1$, the next block of consecutive indices lying in the other set is placed in $\mathcal{J}_2$, and so on. As an example assume that $(i_1, i_2, i_3, i_4, i_5, i_6, i_7) \in B \times B \times N \times B \times B \times N \times N$. In this case, we have $\mathcal{J}_1 = \{i_1, i_2\}$, $\mathcal{J}_2 = \{i_3\}$, $\mathcal{J}_3 = \{i_4, i_5\}$, and $\mathcal{J}_4 = \{i_6, i_7\}$. A partition obtained according to the above construction is clearly ordered at w. We refer to it as an *ordered AS-partition*, and denote it by $\mathcal{J} = \mathcal{J}(w)$.

Note that an ordered AS-partition is not uniquely determined since there can be more than one n-tuple $(i_1, \ldots, i_n)$ satisfying $\delta_{i_1} \leq \ldots \leq \delta_{i_n}$. This situation happens exactly when there are two or more indices i with the same value for δ_i. If these tying indices do not all belong to the same set B or N, then there will be more than one way to generate an ordered AS-partition $\mathcal{J}$.

We say that the bipartition (B, N) is *regular* if there do not exist $i \in B$ and $j \in N$ such that $\delta_i = \delta_j$. Observe that there exists a unique ordered AS-partition if and only if (B, N) is regular. When (B, N) is not regular, our algorithm avoids the computation of an ordered AS-partition and hence of any LLS direction with respect to such a partition.

3.3. Relation between crossover events, the AS step, and the LLS step. In this subsection, we state some variants of Lemma 17 of Vavasis and Ye [28]. Specifically, we present two estimates on the number of iterations needed to guarantee the occurrence of a crossover event. While the first estimate essentially depends on the size of the residual of the LLS step and the step-size at the initial iterate, the second one depends only on the size of the residual of the AS direction at the initial iterate. Lemma 3.4 is a restatement of Lemma 17 of of Vavasis and Ye [28]. Its proof can be found in Lemma 3.4 of Monteiro and Tsuchiya [14].

LEMMA 3.4. *Let $w = (x, y, s) \in \mathcal{N}(\beta)$ for some $\beta \in (0, 1)$ and an ordered partition $J = (J_1, \ldots, J_p)$ at w be given. Let $\delta \equiv \delta(w)$, $\mu = \mu(w)$, and $(Rx^{\mathrm{ll}}(w), Rs^{\mathrm{ll}}(w))$ denote the residual of the LLS direction $(\Delta x^{\mathrm{ll}}, \Delta y^{\mathrm{ll}}, \Delta s^{\mathrm{ll}})$ at w with respect to J. Then, for any $q = 1, \ldots, p$ and any constant*

$$\mathcal{C}_q \geq (1 + \beta)\, \mathrm{spr}(w, J_q)/(1 - \beta)^2$$

and for any $\mu' \in (0, \mu)$ such that

$$\frac{\mu'}{\mu} \leq \frac{\|Rx^{\mathrm{ll}}_{J_q}(w)\|_\infty \|Rs^{\mathrm{ll}}_{J_q}(w)\|_\infty}{n^3 \mathcal{C}_q^2 \bar{\chi}_A^2},$$

the interval $(\mu', \mu]$ contains a $\mathcal{C}_q$-crossover event.

The following lemma is the immediate consequence of Lemma 3.4 and an adaption from Lemma 3.5 of Monteiro and Tsuchiya [14].

LEMMA 3.5. *Let $w = (x, y, s) \in \mathcal{N}(\beta)$ for some $\beta \in (0, 1/4]$ and an ordered partition $J = (J_1, \ldots, J_p)$ at w be given. Define $\delta \equiv \delta(w)$ and $\mu = \mu(w)$, and let $(Rx^{\rm ll}(w), Rs^{\rm ll}(w))$ denote the residual of the LLS direction $(\Delta x^{\rm ll}, \Delta y^{\rm ll}, \Delta s^{\rm ll})$ at w with respect to J. Then, for every $q \in \{1, \ldots, p\}$ and every*

$$\mathcal{C}_q \geq (1+\beta)\mathrm{spr}(w, J_q)/(1-\beta)^2, \tag{45}$$

the following statements hold:

(a) *the PC-TR algorithm (or its variant) started from the point w will generate an iterate $\hat{w}$ with a $\mathcal{C}_q$-crossover event occurring between w and $\hat{w}$ in $\mathcal{O}(\sqrt{n}\,\Phi)$ iterations, where*

$$\Phi \equiv \log(\bar{\chi}_A + n) + \log \mathcal{C}_q + \log\left(\frac{\mu_+/\mu}{\|Rx^{\rm ll}_{J_q}(w)\|_\infty \|Rs^{\rm ll}_{J_q}(w)\|_\infty}\right) \tag{46}$$

and μ_+ is the normalized duality gap attained immediately after the first iteration. Moreover, steps 3 through 5 of the PC-TR algorithm (or its variant), and hence computation of the TR step, is performed in only

$$\mathcal{O}(\Phi/\log(\varepsilon_0^{-1})) \tag{47}$$

of these iterations.

(b) *if, in addition,*

$$\mathrm{gap}(w, J) \geq \max\left\{4\, n\bar{\chi}_A\,,\ \frac{24\sqrt{n}\bar{\chi}_A}{\varepsilon^{\rm a}_{J_q}}\right\} \tag{48}$$

where $\varepsilon^{\rm a}_{J_q} \equiv \min\left\{\|Rx^{\rm a}_{J_q}(w)\|_\infty\,,\ \|Rs^{\rm a}_{J_q}(w)\|_\infty\right\}$, then

$$\Phi = \mathcal{O}\left(\log(\bar{\chi}_A + n) + \log \mathcal{C}_q + \log(\varepsilon^{\rm a}_{J_q})^{-1}\right). \tag{49}$$

Proof. The proofs of the first part of statement (a) and the whole statement (b) are given in Lemma 3.5 of [14]. It remains to prove the latter part of statement (a). We refer to an iteration of the PC-TR algorithm as a TR-iteration whenever the TR direction is computed. Let N_0 be the number of TR-iterations performed before reaching the first iterate $\hat{w}$ such that a $\mathcal{C}_q$-crossover event occurs between w and $\hat{w}$. We will show that N_0 is bounded by (47). Indeed, let $\tilde{w}$ denote the iterate obtained immediately after the (N_0-1)-th TR-iteration. Then, in view of Lemma 3.4, we have

$$\frac{\mu(\tilde{w})}{\mu(w)} > \frac{\|Rx^{\rm ll}_{J_q}(w)\|_\infty \|Rs^{\rm ll}_{J_q}(w)\|_\infty}{n^3 \mathcal{C}_q^2 \bar{\chi}_A^2}. \tag{50}$$

Since the duality gap is reduced by a factor of μ_+/μ in the first iteration, and by a factor of at least $\left((\sqrt{1+\beta}+\gamma)/(2\gamma)\right)\varepsilon_0$ in subsequent TR-iterations, due to relation (26), we conclude that

$$\begin{aligned}\log\left(\frac{\mu_+}{\mu}\right) + (N_0 - 2)\log\left(\frac{\sqrt{1+\beta}+\gamma}{2\gamma}\ \varepsilon_0\right) &\geq \log\frac{\mu(\tilde{w})}{\mu(w)} \\ &> \log\left[\frac{\|Rx^{\rm ll}_{J_q}(w)\|_\infty \|Rs^{\rm ll}_{J_q}(w)\|_\infty}{n^3 \mathcal{C}_q^2 \bar{\chi}_A^2}\right],\end{aligned}$$

which clearly implies that N_0 is bounded by (47). ∎

4. Convergence analysis of the PC-TR algorithm. In this section, we will provide the proofs of Theorems 2.9 and 2.13.

Lemma 3.5 gives a good idea of the effort that will be undertaken in this section, namely, to show that there exists a universal constant $\mathcal{C} = \mathcal{C}(\varepsilon_0) > 0$ with the property that, for each iterate w of the PC-TR algorithm, or its variant, there exists an ordered partition $J = (J_1, \ldots, J_p)$ and an index $q = 1, \ldots, p$ such that $\mathcal{C} \geq (1+\beta)\mathrm{spr}(w, J_q)/(1-\beta)^2$ and the quantity Φ defined in (46) with $\mathcal{C}_q = \mathcal{C}$ is bounded by $\mathcal{O}(n \log(\bar{\chi}_A + n + \varepsilon_0^{-1}))$. In view of Lemma 3.5(a), we would then conclude that a $\mathcal{C}$-crossover event occurs every time $\mathcal{O}(n^{1.5} \log(\bar{\chi}_A + n + \varepsilon_0^{-1}))$ iterations of the PC-TR algorithm is performed. Proposition 3.1 together with the previous fact would then imply that the PC-TR algorithm, or its variant, terminates in $\mathcal{O}(n^{3.5} \log(\bar{\chi}_A + n + \varepsilon_0^{-1}))$ iterations.

We start by introducing the aforementioned constant $\mathcal{C} = \mathcal{C}(\varepsilon_0)$ and another global constant used in this section. Let

$$\bar{g}(\varepsilon_0) \equiv \frac{24\, n \bar{\chi}_A}{\varepsilon_0}, \qquad \mathcal{C}(\varepsilon_0) \equiv \frac{(1+\beta)}{(1-\beta)^2} [\bar{g}(\varepsilon_0)]^n. \tag{51}$$

The proof of the above claim will be broken into three cases, namely: (i) $\varepsilon_2^{\mathrm{a}}(w) \geq \varepsilon_0$; (ii) $\mathrm{gap}(w, \mathcal{J}) \leq \bar{g}(\varepsilon_0)$; and (iii) $\mathrm{gap}(w, \mathcal{J}) \geq \bar{g}(\varepsilon_0)$ and $\varepsilon_2^{\mathrm{a}}(w) \leq \varepsilon_0$, where $\varepsilon_2^{\mathrm{a}}(w)$ is given by (21), $\mathcal{J}$ is the AS-partition at w, and $\mathrm{gap}(w, \mathcal{J})$ is defined in subsection 3.2. The first result below considers the case (i).

LEMMA 4.1. *Suppose that $w \in \mathcal{N}(\beta)$ for some $\beta \in (0, 1/4]$ and that $\varepsilon_2^{\mathrm{a}}(w) \geq \varepsilon_0$ for some constant $\varepsilon_0 > 0$. Then PC-TR algorithm, or its variant, started from the point w will generate an iterate $\hat{w}$ with a $\mathcal{C}(\varepsilon_0)$-crossover event occurring between w and $\hat{w}$ in $\mathcal{O}(n^{1.5} \log(\bar{\chi}_A + n + \varepsilon_0^{-1}))$ iterations, of which $\mathcal{O}(n \log(\bar{\chi}_A + n + \varepsilon_0^{-1}) / \log \varepsilon_0^{-1})$ are TR-iterations.*

Proof. The assumption that $\varepsilon_2^{\mathrm{a}}(w) \geq \varepsilon_0$ implies $\varepsilon_\infty^{\mathrm{a}}(w) \geq \varepsilon_0/\sqrt{n}$, and hence, in view of definition (16), there exists an index $i = 1, \ldots, n$ such that $\min\{|Rx_i^{\mathrm{a}}(w)|, |Rs_i^{\mathrm{a}}(w)|\} \geq \varepsilon_0/\sqrt{n}$. Now let $J = (J_1, \ldots, J_p)$ be a VY $\bar{g}(\varepsilon_0)$-partition at w and let J_q be the layer containing the index i above. Clearly, we have

$$\varepsilon_{J_q}^{\mathrm{a}} \equiv \min\left\{\|Rx_{J_q}^{\mathrm{a}}(w)\|_\infty, \|Rs_{J_q}^{\mathrm{a}}(w)\|_\infty\right\} \geq \varepsilon_0/\sqrt{n}. \tag{52}$$

Using the above inequality, the fact that $\mathrm{gap}(w, J) \geq \bar{g}(\varepsilon_0)$ and (51), we easily see that (48) holds. Since by Proposition 3.3 the spread of every layer of a VY $\bar{g}(\varepsilon_0)$-partition at w is bounded above by $\bar{g}(\varepsilon_0)^n$, we conclude that $\mathrm{spr}(w, J_q) \leq \bar{g}^n$, and hence that the constant $\mathcal{C}(\varepsilon_0)$ defined in (51) satisfies (45) with $\mathcal{C}_q = \mathcal{C}(\varepsilon_0)$. We then conclude from Lemma 3.5(b) that $\Phi = \mathcal{O}(n^{1.5} \log(\bar{\chi}_A + n + \varepsilon_0^{-1}))$ in view of (52), (51), and the fact that $\log(\mathcal{C}_q) = \mathcal{O}(n \log \bar{g}(\varepsilon_0)) = \mathcal{O}(n \log(\bar{\chi}_A + n + \varepsilon_0^{-1}))$. The conclusion of the lemma now follows from the previous observation and Lemma 3.5(a). ∎

The next result takes care of case (ii), namely the case in which $\mathrm{gap}(w, \mathcal{J}) \leq \bar{g}(\varepsilon_0)$.

LEMMA 4.2. *Suppose that $w \in \mathcal{N}(\beta)$ for some $\beta \in (0, 1/4]$. Let $\bar{g}(\varepsilon_0)$ and $\mathcal{C}(\varepsilon_0)$ be the constants defined in (51). Let $\mathcal{J} = (\mathcal{J}_1, \ldots, \mathcal{J}_r)$ be an ordered AS-partition at w and assume that $\mathrm{gap}(w, \mathcal{J}) \leq \bar{g}(\varepsilon_0)$. Then, the PC-TR algorithm, or its variant, started from the point w will generate an iterate $\hat{w}$ with a $\mathcal{C}(\varepsilon_0)$-crossover event occurring between w and $\hat{w}$ in $\mathcal{O}(n^{1.5} \log(\bar{\chi}_A + n + \varepsilon_0^{-1}))$ iterations, of which $\mathcal{O}(n \log(\bar{\chi}_A + n + \varepsilon_0^{-1}) / \log \varepsilon_0^{-1})$ are TR-iterations.*

Proof. Assume that $\mathrm{gap}(w, \mathcal{J}) \leq \bar{g}(\varepsilon_0)$ and let $J = (J_1, \ldots, J_p)$ be a VY $\bar{g}(\varepsilon_0)$-partition at w. Using the assumption that $\mathrm{gap}(w, \mathcal{J}) \leq \bar{g}(\varepsilon_0)$, it is easy to see that

there exist two indices i, j of different types, say $i \in B(w)$ and $j \in N(w)$, both lying in some layer J_q of J. By Lemma 2.6 and the definition of $(B(w), N(w))$ given in (20), it follows that $|Rx_i^{\mathrm{a}}(w)| \geq 1/4$ and $|Rs_j^{\mathrm{a}}(w)| \geq 1/4$, and hence that

$$\varepsilon_{J_q}^{\mathrm{a}} \equiv \min\left\{\|Rx_{J_q}^{\mathrm{a}}(w)\|_\infty\,,\ \|Rs_{J_q}^{\mathrm{a}}(w)\|_\infty\right\} \geq \frac{1}{4}. \tag{53}$$

Using this inequality and the fact that $\mathrm{gap}(w, J) \geq \bar{g}(\varepsilon_0) \geq 96\bar{\chi}_A n$, where the last inequality is due to (51) and (25), we easily see that (48) holds. Since by Proposition 3.3 the spread of every layer of a VY $\bar{g}(\varepsilon_0)$–partition at w is bounded above by $\bar{g}(\varepsilon_0)^n$, we conclude that $\mathrm{spr}(w, J_q) \leq \bar{g}^n$, and hence that (45) holds with $\mathcal{C}_q = \mathcal{C}(\varepsilon_0)$ in view of (51). The result now follows from Lemma 3.5 by noting that the quantity Φ in (49) with $\mathcal{C}_q = \mathcal{C}(\varepsilon_0)$ is bounded by $\mathcal{O}(n^{1.5}\log(\bar{\chi}_A + n + \varepsilon_0^{-1}))$ in view of (51) and (53). □

From now on, we consider case (iii), namely the case in which $\mathrm{gap}(w, \mathcal{J}) \geq \bar{g}(\varepsilon_0)$ and $\varepsilon_2^{\mathrm{a}}(w) \leq \varepsilon_0$.

We start by stating a technical result whose proof is given in Lemma 4.3 of [14] and holds for any $\bar{g}(\varepsilon_0) \geq 96n\bar{\chi}_A$, hence for our specific choice of $\bar{g}(\varepsilon_0)$ given in (51), in view of (25).

LEMMA 4.3. *Suppose that $w \in \mathcal{N}(\beta)$ for some $\beta \in (0, 1/4]$. Let $\bar{g}(\varepsilon_0)$ and $\mathcal{C}(\varepsilon_0)$ be the constants defined in* (51). *Let $\mathcal{J} = (\mathcal{J}_1, \ldots, \mathcal{J}_r)$ denote the AS-partition at w and assume that $\mathrm{gap}(w, \mathcal{J}) \geq \bar{g}(\varepsilon_0)$. Let $(Rx^{\mathrm{l}}(w), Rs^{\mathrm{l}}(w))$ denote the residual of the LLS direction at w with respect to $\mathcal{J}$. Let*

$$\hat{\Phi} \equiv n\log\left(\bar{\chi}_A + n + \varepsilon_0^{-1}\right) + \log\left(\frac{\mu_+/\mu}{\varepsilon_\infty^{\mathrm{l}}(w)}\right), \tag{54}$$

μ_+ is the normalized duality gap attained immediately after the first iteration,

$$\varepsilon_\infty^{\mathrm{l}}(w) \equiv \max\left\{\left\|Rx_N^{\mathrm{l}}(w)\right\|_\infty, \left\|Rs_B^{\mathrm{l}}(w)\right\|_\infty\right\} \tag{55}$$

and $(B, N) \equiv (B(w), N(w))$. Then, the PC-TR algorithm started from the point w will generate an iterate $\hat{w}$ with a $\mathcal{C}(\varepsilon_0)$-crossover event occurring between w and $\hat{w}$ in $\mathcal{O}(\sqrt{n}\,\hat{\Phi})$ iterations, of which $\mathcal{O}(\hat{\Phi}/\log\varepsilon_0^{-1})$ are TR-iterations.

Our goal now will be to estimate, under the conditions of case (iii), the second logarithm that appears in the iteration-complexity bound (54). In this estimation procedure, it is important to show that the first iteration from w is a TR-iteration. This will always be the case for the PC-TR algorithm since a TR-iteration occurs in this algorithm if and only if $\varepsilon_2^{\mathrm{a}}(w) \leq \varepsilon_0$ and case (iii) assumes this condition. On the other hand, for the variant, TR-iteration occurs if and only if, in addition to $\varepsilon_2^{\mathrm{a}}(w) \leq \varepsilon_0$, we also have $\max\{\psi_p(1), \psi_d(1)\} \leq \gamma$, where the curves $\psi_p(\cdot)$ and $\psi_d(\cdot)$ are defined in (29) and (40). The next two results show that the latter condition also holds under case (iii).

Given $F \in \Re^{m\times n}$, $h \in \Re^m$, and a scaling vector $z \in \Re^n_{++}$, consider the projection $p^0 \in \Re^n$ given by

$$p^0 \equiv \mathrm{argmin}_{p\in\Re^n}\left\{\|h - p\|^2 : FZp = 0\right\}, \tag{56}$$

where $Z \equiv \mathrm{Diag}(z)$. For a given ordered partition $J = (J_1, \ldots, J_l)$, Lemma 4.4 shows that if $\mathrm{gap}(z, J)$ is large, then the projection matrix onto $\mathrm{Ker}(F\,\mathrm{Diag}(z))$ can be well approximated by a block diagonal matrix where each block is a projection matrix

associated with some layer J_k of J. This fact was first established in [23] and an alternative proof can be found in [13]. The proof of a slightly stronger version of the variant stated below can be found in [16].

LEMMA 4.4. *Let $F \in \Re^{m\times n}$, $h \in \Re^m$, $z \in \Re^n_{++}$, and an ordered partition $J = (J_1, \ldots, J_l)$ of $\{1, \ldots, n\}$ with respect to z be given. Define $p^0 \in \Re^n$ as in* (56) *and $\tilde{p}^0 \in \Re^n$ as*

$$\tilde{p}^0_{J_k} \equiv \operatorname{argmin}_{\tilde{p}_{J_k}\in\Re^{J_k}} \left\{ \|\tilde{p}_{J_k} - h_{J_k}\|^2 : F_{J_k} Z_{J_k} \tilde{p}_{J_k} \in \operatorname{Im}(F_{\bar{J}_k}) \right\}, \tag{57}$$

for every $k = 1, \ldots, l$, where $\bar{J}_k \equiv J_{k+1} \cup \ldots \cup J_l$. Then,

$$\|p^0 - \tilde{p}^0\|_\infty \le K(3 + 2K)\|h\|, \tag{58}$$

where $K \equiv \bar{\chi}_F/\operatorname{gap}(z, J)$.

Using this approximation result, we are now able to prove the result mentioned just after Lemma 4.3.

LEMMA 4.5. *Assume that $w \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ and that $\operatorname{gap}(w, \mathcal{J}) \ge \bar{g}(\varepsilon_0)$ for some $\varepsilon_0 \in (0, 12n]$, where $\mathcal{J} = (\mathcal{J}_1, \ldots, \mathcal{J}_r)$ denotes the ordered AS-partition at w. Then, the curves $\psi_p(\cdot)$ and $\psi_d(\cdot)$ defined in* (29) *and* (40)*, respectively, satisfy $\max\{\psi_p(1), \psi_d(1)\} \le \varepsilon_0/6$.*

Proof. We will show only the inequality $\psi_p(1) \le \varepsilon_0/6$. The proof of the inequality $\psi_d(1) \le \varepsilon_0/6$ is similar. Consider the projections p^0 and $\tilde{p}^0$ defined in Lemma 4.4 with $F = A$, $h = (h_B, h_N) \equiv (0, \delta_N x_N)$, $z = \delta^{-1}$, and $J = (\mathcal{J}_r, \ldots, \mathcal{J}_1)$, where $\delta \equiv \delta(w)$. It is easy to see that the constant K of Lemma 4.4 is exactly equal to $\bar{\chi}_A/\operatorname{gap}(w, \mathcal{J})$. It then follows from relation (51) and the assumptions $\operatorname{gap}(w, \mathcal{J}) \ge \bar{g}(\varepsilon_0)$ and $\varepsilon_0 \le 12n$ that $K \le \varepsilon_0/(24n) \le 1/2$. Using these two inequalities, the conclusion of Lemma 4.4 and the fact that $\|h\| \le \|\delta x\| = \sqrt{x^T s} = \sqrt{n\mu}$, we then obtain

$$\frac{1}{\sqrt{\mu}}\|p^0_B - \tilde{p}^0_B\| \le \frac{\sqrt{n}}{\sqrt{\mu}}\|p^0_B - \tilde{p}^0_B\|_\infty \le nK(3 + 2K) \le 4nK \le \varepsilon_0/6. \tag{59}$$

Moreover, definition (28) clearly implies that $p^0 = \delta\Delta x(1)$, where we recall that $\Delta x(1)$ is the optimal solution of (28) with $\lambda_p = 1$. Using the fact that $h_{J_k} = 0$ for every $J_k \subset B$ and the definition (57), we easily see that $\tilde{p}^0_{J_k} = 0$ for every $J_k \subset B$ and hence that $\tilde{p}^0_B = 0$. The last two observations together with (59) and (29) then imply that

$$\psi_p(1) = \frac{1}{\sqrt{\mu}}\|\delta_B \Delta x_B(1)\| = \frac{1}{\sqrt{\mu}}\|p^0_B\| = \frac{1}{\sqrt{\mu}}\|p^0_B - \tilde{p}^0_B\| \le \varepsilon_0/6. \qquad \square$$

The following result follows an immediate consequence of Lemma 4.5.

LEMMA 4.6. *Assume that w is an iterate of the PC-TR variant such that $\varepsilon^{\mathrm{a}}_2(w) \le \varepsilon_0$ and $\operatorname{gap}(w, \mathcal{J}) \ge \bar{g}(\varepsilon_0)$. Then, the iteration of the PC-TR variant from w is a TR-iteration.*

Proof. In view of Lemma 4.5 and the assumptions that $\operatorname{gap}(w, \mathcal{J}) \ge \bar{g}(\varepsilon_0)$ and $\varepsilon_0 \le \gamma/3$, we conclude that $\max\{\psi_p(1), \psi_d(1)\} \le \gamma/18$. This inequality, together with the assumption $\varepsilon^{\mathrm{a}}_2(w) \le \varepsilon_0$, implies that the iteration of the PC-TR variant from w is a TR-iteration (see the statement of Theorem 2.13). $\square$

When a TR-iteration is performed, it follows from relation (26) that the duality gap is reduced by a factor bounded by $\mathcal{O}(\varepsilon^\tau_2(w; \gamma_p, \gamma_d))$. The following result shows that this factor is indeed $\mathcal{O}(\sqrt{n}\varepsilon^{\mathrm{l}}_\infty(w))$, where $\varepsilon^{\mathrm{l}}_\infty(w)$ is defined in (55), thereby giving the necessary means to bound the second logarithm which appears in (54).

LEMMA 4.7. *Suppose that* $w \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ *is such that* $\varepsilon_2^{\mathrm{a}}(w) \leq \varepsilon_0$. *Let* $\Delta w^l = (\Delta x^l, \Delta y^l, \Delta s^l)$ *denote the LLS direction at* w *with respect to the AS-partition* $\mathcal{J} = (\mathcal{J}_1, \ldots, \mathcal{J}_r)$ *and assume that* $\mathrm{gap}(\mathcal{J}) \geq \bar{g}(\varepsilon_0)$. *Then, we have*

$$\max\left\{\frac{\|\delta_B \Delta x_B^l\|}{\sqrt{\mu}}, \frac{\|\delta_N^{-1} \Delta s_N^l\|}{\sqrt{\mu}}\right\} \leq \frac{3\varepsilon_0}{2}. \tag{60}$$

Moreover, if in addition $\varepsilon_0 \leq \gamma/3$, *then*

$$\varepsilon_2^{\tau}(w; \gamma_p, \gamma_d) \leq \sqrt{n}\varepsilon_{\infty}^{\mathrm{l}}(w) \tag{61}$$

for any $\gamma_p, \gamma_d \geq \gamma/2$, *where* $\varepsilon_{\infty}^{\mathrm{l}}(w)$ *and* $\varepsilon_2^{\tau}(w; \gamma_p, \gamma_d)$ *are defined in* (55) *and* (24), *respectively.*

Proof. Clearly, by definitions (9) and (43) we have $A\Delta x^l = 0$. Moreover, from the triangle inequality for norms, Theorem 3.2, relations (14), (21) and (51) and the assumptions that $\mathrm{gap}(w, \mathcal{J}) \geq \bar{g}(\varepsilon_0)$ and $\varepsilon_2^{\mathrm{a}}(w) \leq \varepsilon_0$, we conclude that

$$\begin{aligned}\frac{\|\delta_B \Delta x_B^l\|}{\sqrt{\mu}} \leq \frac{\|\delta_B \Delta x_B^{\mathrm{a}}\|}{\sqrt{\mu}} + \frac{\|\delta_B(\Delta x_B^l - \Delta x_B^{\mathrm{a}})\|}{\sqrt{\mu}} &\leq \|Rs_B^a\| + \sqrt{n}\,\|Rx^l - Rx^a\|_{\infty} \\ &\leq \varepsilon_2^{\mathrm{a}}(w) + \frac{12n\bar{\chi}_A}{\mathrm{gap}(w, \mathcal{J})} \leq \varepsilon_0 + \frac{12n\bar{\chi}_A}{\bar{g}(\varepsilon_0)} \leq \varepsilon_0 + \frac{\varepsilon_0}{2} \leq \frac{3\varepsilon_0}{2}.\end{aligned}$$

In a similar manner, we can also show that $\|\delta_N^{-1}\Delta s_N^l\|/\sqrt{\mu} \leq 3\varepsilon_0/2$, showing that (60) holds.

Assume now that $\varepsilon_0 \leq \gamma/3$ also holds. In view of (60), it follows that Δx_B^l and Δs_N^l are feasible for subproblems (22) and (23), respectively, whenever $\gamma_p, \gamma_d \geq \gamma/2$. Hence, we conclude that $\|Rx_N^{\tau}\| \leq \|Rx_N^l\|$ and $\|Rs_B^{\tau}\| \leq \|Rs_B^l\|$, from which it follows that

$$\varepsilon_2^{\tau}(w; \gamma_p, \gamma_d) := \max\{\|Rx_N^{\tau}\|, \|Rs_B^{\tau}\|\} \leq \sqrt{n}\max\{\|Rx_N^l\|_{\infty}, \|Rs_B^l\|_{\infty}\} = \sqrt{n}\varepsilon_{\infty}^{\mathrm{l}}(w). \quad \square$$

We are now ready to prove Theorems 2.9 and 2.13.

Proof of Theorems 2.9 *and* 2.13. Let $\mathcal{C}$ and $\bar{g}(\varepsilon_0)$ be the constant defined in (51). We claim that the PC-TR algorithm started from any $w \in \mathcal{N}(\beta)$ generates an iterate $\hat{w}$ with a $\mathcal{C}(\varepsilon_0)$-crossover event occurring between w and $\hat{w}$ in $\mathcal{O}(n^{1.5}\log(\bar{\chi}_A + n + \varepsilon_0^{-1}))$ iteration, of which $\mathcal{O}(n\log(\bar{\chi}_A + n + \varepsilon_0^{-1})/\log\varepsilon_0^{-1})$ are TR-iterations. Since by Proposition 3.1 there can be at most $n(n+1)/2$ $\mathcal{C}(\varepsilon_0)$-crossover events, we conclude that the PC-TR algorithm must ultimately terminate in $\mathcal{O}(n^{3.5}\log(\bar{\chi}_A + n + \varepsilon_0^{-1}))$ iterations, of which $\mathcal{O}(n^3\log(\bar{\chi}_A + n + \varepsilon_0^{-1})/\log\varepsilon_0^{-1})$ are TR-iterations. Let $\mathcal{J} = (\mathcal{J}_1, \ldots, \mathcal{J}_r)$ denote an AS-partition at w. We split the proof into one of the following three possible cases: (1) $\varepsilon_2^{\mathrm{a}}(w) \geq \varepsilon_0$; (2) $\mathrm{gap}(\mathcal{J}) \leq \bar{g}(\varepsilon_0)$; and (3) $\varepsilon_2^{\mathrm{a}}(w) \leq \varepsilon_0$ and $\mathrm{gap}(\mathcal{J}) \geq \bar{g}(\varepsilon_0)$. The claim clearly holds for the first two cases due to Lemmas 4.1 and 4.2. Moreover, Lemma 4.3 implies that the claim also holds in the third case as long as we can show that the quantity $(\mu_+/\mu)/\varepsilon_{\infty}^{\mathrm{l}}(w)$ appearing in (54) is $\mathcal{O}(\sqrt{n})$. Indeed, assume that $\varepsilon_{\infty}^{\mathrm{a}}(w) \leq \varepsilon_0$ and $\mathrm{gap}(\mathcal{J}) \geq \bar{g}(\varepsilon_0)$. Then, the iteration from w for both the PC-TR algorithm and its variant is a TR-iteration in view of Lemma 4.6. Then, it follows from Proposition 2.8 and Lemma 4.7 that $(\mu_+/\mu)/\varepsilon_{\infty}^{\mathrm{l}}(w) = \mathcal{O}(\sqrt{n})$. $\square$

5. Arithmetic complexity for the PC-TR variant. In this section, we will provide the details of the several claims made on subsection 2.6 and prove the main result stated in that subsection, namely Theorem 2.14.

We start by noting that the results stated in subsection 2.6 remain invariant if elementary row operations are applied to the rows of A. Indeed, the condition that $A\Delta x(\lambda) = 0$ can be replaced by the condition that $\Delta x(\lambda)$ is in the null space of A, which remains invariant when the elementary row operations are performed on A. In this section, we will therefore freely perform elementary row operations to bring A to a more convenient form.

Let $w \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ be given. By placing the columns with indices in $B \equiv B(w)$ before the ones with indices in $N \equiv N(w)$, it is easy to see that there exists a sequence of elementary row operations which brings A to a matrix of the form

$$\begin{pmatrix} B & E \\ 0 & N \end{pmatrix}, \tag{62}$$

where $B \in \Re^{r_b \times |B|}$ and $N \in \Re^{r_n \times |N|}$ are full row rank matrices with $r_b \equiv \mathrm{rank}(A_B)$ and $r_n \equiv m - r_b$. By performing further elementary row operations, we may assume that A contains an $m \times m$ identity matrix, or equivalently, after permuting columns of A if necessary, the matrices B, N, and E have the form

$$B = [\ \tilde{B} \quad I\ ], \quad N = [\ I \quad \tilde{N}\ ], \quad E = [\ 0 \quad \tilde{E}\ ], \tag{63}$$

and hence, A has the form

$$A = \begin{pmatrix} \tilde{B} & I & 0 & \tilde{E} \\ 0 & 0 & I & \tilde{N} \end{pmatrix}, \tag{64}$$

where $\tilde{B} \in \Re^{r_b \times (|B| - r_b)}$, $\tilde{N} \in \Re^{r_n \times (|N| - r_n)}$, and $\tilde{E} \in \Re^{r_b \times (|N| - r_n)}$. Note that, by abuse of notation, we still denote the above matrix by A.

The following result, which is only used in the proof of Lemma 2.12, strongly uses the fact that A has the form (64). We observe however that the weaker form (62) of A is sufficient to establish the other results of subsection 2.6.

LEMMA 5.1. *For any positive diagonal $n \times n$ matrix D, there exists a matrix $W \in \Re^{|B| \times |N|}$ such that $E = BW$ and $\|D_B^{-1} W D_N\| \leq \bar{\chi}_{AD}$, where B and E are given by* (63).

Proof. We first prove the result for $D = I$. In this case, we choose W as

$$W = \begin{pmatrix} 0 & 0 \\ 0 & \tilde{E} \end{pmatrix}.$$

It is easy to see that $E = BW$ and that $\|W\| = \|\tilde{E}\| \leq \bar{\chi}_A$, where the last inequality follows as a consequence of Proposition 2.3(c).

Assume now that D is an arbitrary $n \times n$ positive diagonal matrix and let D_I denote the diagonal submatrix of D corresponding to the identity matrix of A. Also, let D_b and D_n denote the diagonal submatrices of D_I corresponding to the first $|B|$ columns and the last $|N|$ columns of the identity matrix of A, respectively. Then, the matrix given by

$$\hat{A} \equiv D_I^{-1} A D = \begin{pmatrix} D_b^{-1} B D_B & D_b^{-1} E D_N \\ 0 & D_n^{-1} N D_N \end{pmatrix}$$

also contains an $m \times m$ identity matrix. Applying the result shown in the first paragraph of this proof to the matrix $\hat{A}$, we conclude that there exists a matrix

$\hat{W}$ such that $\|\hat{W}\| \le \bar{\chi}_{\hat{A}} = \bar{\chi}_{AD}$ and $(D_b^{-1}BD_B)\hat{W} = D_b^{-1}ED_N$, or equivalently $B(D_B\hat{W}D_N^{-1}) = E$. The result now follows by letting $W = D_B\hat{W}D_N^{-1}$. ∎

The following lemma establishes some technical results about the direction $\Delta x(\lambda)$ defined in (28). Its proof is based on techniques developed in [23].

LEMMA 5.2. *Let $\Delta x(\cdot)$ be the curve defined as in* (28) *and let $\Delta x'(\cdot)$ denote its derivative. Then:*

$$\Delta x_B(0) \equiv \lim_{\lambda\to 0^+} \Delta x_B(\lambda) = \Delta_B^{-2}B^T(B\Delta_B^{-2}B^T)^{-1}H\Delta_N x_N; \tag{65}$$

$$\Delta x_N(0) \equiv \lim_{\lambda\to 0^+} \Delta x_N(\lambda) = -\Delta_N^{-1}P_{N\Delta_N^{-1}}x_N; \tag{66}$$

$$\Delta x'_B(0) \equiv \lim_{\lambda\to 0^+} \Delta x'_B(\lambda) = \Delta_B^{-2}B^T(B\Delta_B^{-2}B^T)^{-1}HH^T(B\Delta_B^{-2}B^T)^{-1}H\Delta_N x_N; \tag{67}$$

where $\Delta_B \equiv \mathrm{Diag}\{\delta_B\}$, $\Delta_N \equiv \mathrm{Diag}\{\delta_N\}$, $H \equiv E\Delta_N^{-1}P_{N\Delta_N^{-1}}$, and $P_{N\Delta_N^{-1}}$ denote the projection matrix onto the null space of $N\Delta_N^{-1}$.

Proof. Defining $D_\lambda \equiv \mathrm{Diag}\{\sqrt{\lambda}\delta_B, \delta_N\}$, we can easily see from (28) that $D_\lambda\Delta x(\lambda)$ is the projection of the vector $(0, \delta_N x_N)$ onto the null space of AD_λ^{-1}. Hence, for any $\lambda > 0$ we have

$$\Delta x_B(\lambda) = \lambda^{-1}\Delta_B^{-2}A_B^T(AD_\lambda^{-2}A^T)^{-1}A_N x_N, \tag{68}$$

$$\Delta x_N(\lambda) = \Delta_N^{-2}A_N^T(AD_\lambda^{-2}A^T)^{-1}A_N x_N - x_N. \tag{69}$$

Using (62) and the definition of D_λ, we have

$$\begin{aligned} AD_\lambda^{-2}A^T &= \begin{pmatrix} B & E \\ 0 & N \end{pmatrix}\begin{pmatrix} \lambda^{-1}\Delta_B^{-2} & 0 \\ 0 & \Delta_N^{-2} \end{pmatrix}\begin{pmatrix} B^T & 0 \\ E^T & N^T \end{pmatrix} \\ &= \begin{pmatrix} \lambda^{-1}B\Delta_B^{-2}B^T + E\Delta_N^{-2}E^T & E\Delta_N^{-2}N^T \\ N\Delta_N^{-2}E^T & N\Delta_N^{-2}N^T \end{pmatrix} \\ &= \begin{pmatrix} \lambda^{-1}R_{BB} + R_{EE} & R_{EN} \\ R_{EN}^T & R_{NN} \end{pmatrix}, \end{aligned}$$

where

$$R_{BB} \equiv B\Delta_B^{-2}B^T,\ \ R_{NN} \equiv N\Delta_N^{-2}N^T,\ \ R_{EE} \equiv E\Delta_N^{-2}E^T,\ \ R_{EN} \equiv E\Delta_N^{-2}N^T. \tag{70}$$

Using the standard way to compute the inverse of a 2×2 block matrix (see, for example, page 71–72 of [1]), it is easy to verify that

$$\left(AD_\lambda^{-2}A^T\right)^{-1} = \begin{pmatrix} U_\lambda & V_\lambda \\ V_\lambda^T & Z_\lambda \end{pmatrix},$$

where

$$U_\lambda = (\lambda^{-1}R_{BB} + R_{EE} - R_{EN}R_{NN}^{-1}R_{EN}^T)^{-1}, \tag{71}$$

$$V_\lambda = -U_\lambda R_{EN}R_{NN}^{-1}, \tag{72}$$

$$Z_\lambda = (R_{NN} - R_{EN}^T(\lambda^{-1}R_{BB} + R_{EE})^{-1}R_{EN})^{-1}. \tag{73}$$

Note that, by (70), we have

$$\begin{aligned} R_{EE} - R_{EN}R_{NN}^{-1}R_{EN}^T &= E\Delta_N^{-1}(I - \Delta_N^{-1}N^T(N\Delta_N^{-2}N^T)^{-1}N\Delta_N^{-1})\Delta_N^{-1}E^T \\ &= E\Delta_N^{-1}P_{N\Delta_N^{-1}}\Delta_N^{-1}E^T = HH^T. \end{aligned}$$

where we recall $H = E\Delta_N^{-1}P_{N\Delta_N^{-1}}$ and $P_{N\Delta_N^{-1}}$ denotes the projection matrix onto the null space of $N\Delta_N^{-1}$. Hence, by (71), we have

$$U_\lambda = (\lambda^{-1}R_{BB} + HH^T)^{-1} = \lambda(R_{BB} + \lambda HH^T)^{-1}. \tag{74}$$

Also, by (62) and (72), we have

$$\left(AD_\lambda^{-2}A^T\right)^{-1}A_N = \begin{pmatrix} U_\lambda E + V_\lambda N \\ V_\lambda^T E + Z_\lambda N \end{pmatrix} = \begin{pmatrix} U_\lambda(E - R_{EN}R_{NN}^{-1}N) \\ -R_{NN}^{-1}R_{EN}^T U_\lambda^T E + Z_\lambda N \end{pmatrix}. \tag{75}$$

Hence, using relations (68), (70), (74), and (75) and the definition of H, we obtain

$$\begin{aligned}
\Delta x_B(\lambda) &= \lambda^{-1}\Delta_B^{-2}B^T U_\lambda(E - R_{EN}R_{NN}^{-1}N)x_N \\
&= \lambda^{-1}\Delta_B^{-2}B^T U_\lambda E\Delta_N^{-1}\left(I - \Delta_N^{-1}N^T(N\Delta_N^{-2}N^T)^{-1}N\Delta_N^{-1}\right)\Delta_N x_N \\
&= \lambda^{-1}\Delta_B^{-2}B^T U_\lambda E\Delta_N^{-1}P_{ND_N^{-1}}\Delta_N x_N \;=\; \Delta_B^{-2}B^T(R_{BB} + \lambda HH^T)^{-1}H\Delta_N x_N,
\end{aligned} \tag{76}$$

from which we can easily see that (65) and (67) hold. Now, using relations (74) and (73), we easily see that

$$\lim_{\lambda\to 0^+} U_\lambda = 0, \quad \lim_{\lambda\to 0^+} Z_\lambda = R_{NN}^{-1}.$$

Relation (66) now follows from the last conclusion and relations (69) and (75). ∎

We need one more technical result before giving the proofs of the results of subsection 2.6.

LEMMA 5.3. *Let $G \in \Re^{p\times q}$ and $g \in \Re^q$ be given. Then, $(GG^T)Gg = 0$ if and only if $Gg = 0$.*

Proof. The assumption $(GG^T)Gg = 0$ clearly implies that $\|G^TGg\|^2 = 0$, and hence that $G^TGg = 0$. Also, the latter condition implies that $\|Gg\|^2 = 0$, or equivalently, $Gg = 0$. ∎

We are now ready to prove Lemma 2.10, Lemma 2.12, and Theorem 2.14 stated in subsection 2.6.

Proof of Lemma 2.10. We first prove statements (a) and (b). The existence and characterizations of the two limits $\Delta x(0) \equiv \lim_{\lambda\to 0^+}\Delta x(\lambda)$ and $\Delta x'_B(0) \equiv \lim_{\lambda\to 0^+}\Delta x'_B(\lambda)$ were established in Lemma 5.2. The alternative characterization given by (30) and (31) of the limit $\Delta x(0) \equiv \lim_{\lambda\to 0^+}\Delta x(\lambda)$ can be easily proved by showing that $\Delta x_B(0)$ and $\Delta x_N(0)$ satisfy the optimality conditions, and hence are optimal solutions of (30) and (31), respectively. Now, relation (32) follows immediately from (29). Moreover, differentiating (29) with respect to λ, we conclude that

$$\psi'_p(\lambda) = \frac{[\delta_B\Delta x_B(\lambda)]^T[\delta_B\Delta x'_B(\lambda)]}{\sqrt{\mu}\|\delta_B\Delta x_B(\lambda)\|} = \frac{[\delta_B\Delta x_B(\lambda)]^T[\delta_B\Delta x'_B(\lambda)]}{\mu\psi_p(\lambda)}. \tag{77}$$

Hence, under the the condition that $\psi_p(0) \neq 0$, the limit $\psi'_p(0) \equiv \lim_{\lambda\to 0^+}\psi'_p(\lambda)$ exists and is equal to the right-hand side of (77) with $\lambda = 0$.

We now outline the proof of statement (c). Using definition (29) and relation (76) we conclude that

$$\begin{aligned}
\mu\psi_p(\lambda)^2 &= x_N^T\Delta_N H^T(R_{BB} + \lambda HH^T)^{-1}R_{BB}(R_{BB} + \lambda HH^T)^{-1}H\Delta_N x_N \\
&= x_N^T\Delta_N H^T R_{BB}^{-1/2}(I + \lambda R_{BB}^{-1/2}HH^TR_{BB}^{-1/2})^{-2}R_{BB}^{-1/2}H\Delta_N x_N, \\
&= \|(I + \lambda\tilde{H})^{-1}\tilde{g}\|^2,
\end{aligned} \tag{78}$$

where $\mu \equiv \mu(w)$, $\tilde{H} \equiv R_{BB}^{-1/2} H H^T R_{BB}^{-1/2}$, and $\tilde{g} \equiv R_{BB}^{-1/2} H \Delta_N x_N$. The above formula for $\psi_p(\cdot)$ allows us to express it in terms of the eigenvalues and eigenvector of the positive semidefinite matrix $\tilde{H}$, and the resulting expression easily reveals that (i) $\tilde{H}\tilde{g} = 0$ if and only if $\psi_p(\cdot)$ is identically constant, and (ii) $\tilde{H}\tilde{g} \neq 0$ if and only if $\psi_p(\cdot)$ is strictly decreasing and strictly convex over the interval $(0, \infty)$. Moreover, if case (i) occurs, it follows from Lemma 5.3 with $G = R_{BB}^{-1/2} H$ and $g = \Delta_N x_N$ that $0 = Gg = R_{BB}^{-1/2} H \Delta_N x_N$, and hence that $H \Delta_N x_N = 0$. In view of (65), this implies that $\Delta x_B(0) = 0$, and hence that $\psi_p(0) = 0$. We have thus shown that $\psi_p(\cdot)$ is indeed identically zero when case (i) occurs.

We now show statement d). Let $0 < \lambda_1 \leq \lambda_2$ be given. By (78), we have

$$\frac{\psi_p^2(\lambda_1)}{\psi_p^2(\lambda_2)} = \frac{\tilde{g}^T(\lambda_1 \tilde{H} + I)^{-2}\tilde{g}}{\tilde{g}^T(\lambda_2 \tilde{H} + I)^{-2}\tilde{g}} = \frac{u^T \tilde{M} u}{u^T u},$$

where $u \equiv (\lambda_2 \tilde{H} + I)^{-1}\tilde{g}$ and $\tilde{M} = (\lambda_2 \tilde{H} + I)(\lambda_1 \tilde{H} + I)^{-2}(\lambda_2 \tilde{H} + I)$. Using the fact that $0 < \lambda_1 \leq \lambda_2$, we easily see that the largest eigenvalue of $\tilde{M}$, and hence $\psi_p^2(\lambda_1)/\psi_p^2(\lambda_2)$ is bounded by $(\lambda_2/\lambda_1)^2$. ∎

Proof of Lemma 2.12. First observe that, by (29), (77), and the Cauchy–Schwarz inequality, we have

$$\frac{|\psi_p'(0)|}{\psi_p(0)} = \frac{\left|[\delta_B \Delta x_B(0)]^T [\delta_B \Delta x_B'(0)]\right|}{\|\delta_B \Delta x_B(0)\|^2} \leq \frac{\|\delta_B \Delta x_B'(0)\|}{\|\delta_B \Delta x_B(0)\|}. \tag{79}$$

We will now use the formulas developed in Lemma 5.2 to bound the above ratio from above. Letting $\Delta_B \equiv \mathrm{Diag}(\delta_B)$ and $\Delta_N \equiv \mathrm{Diag}(\delta_N)$, we have that the matrix H defined in Lemma 5.2 can be written as

$$H \equiv E \Delta_N^{-1} P_{N\Delta_N^{-1}} = B \Delta_B^{-1} (\Delta_B W \Delta_N^{-1} P_{N\Delta_N^{-1}}) = B \Delta_B^{-1} M,$$

where $M \equiv \Delta_B W \Delta_N^{-1} P_{N\Delta_N^{-1}}$ and W is a matrix as in Lemma 5.1. Using this expression for H and relations (65) and (67), we then obtain

$$\begin{aligned}
\delta_B \Delta x_B'(0) &= \Delta_B^{-1} B^T (B \Delta_B^{-2} B^T)^{-1} H H^T (B \Delta_B^{-2} B^T)^{-1} H \Delta_N x_N \\
&= \Delta_B^{-1} B^T (B \Delta_B^{-2} B^T)^{-1} \left(B \Delta_B^{-1} M\right) \left(M^T \Delta_B^{-1} B^T\right) (B \Delta_B^{-2} B^T)^{-1} H \Delta_N x_N \\
&= [\Delta_B^{-1} B^T (B \Delta_B^{-2} B^T)^{-1} B \Delta_B^{-1}] (M M^T) [\delta_B \Delta x_B(0)].
\end{aligned}$$

Using the fact that the matrix inside the first bracket in the right-hand side of the above inequality is a projection matrix and Lemma 5.1, we then conclude that

$$\begin{aligned}
\frac{\|\delta_B \Delta x_B'(0)\|}{\|\delta_B \Delta x_B(0)\|} &\leq \|M\|^2 = \|\Delta_B W \Delta_N^{-1} P_{N D_N^{-1}}\|^2 \leq \|\Delta_B W \Delta_N^{-1}\|^2 \\
&\leq \|\delta_B (\delta_0)_B^{-1}\|_\infty^2 \|(\Delta_0)_B W (\Delta_0)_N^{-1}\|^2 \|\delta_N^{-1}(\delta_0)_N\|_\infty^2 \\
&\leq [\bar{\chi}_{A\Delta_0^{-1}}]^2 \|\delta_B (\delta_0)_B^{-1}\|_\infty^2 \|\delta_N^{-1}(\delta_0)_N\|_\infty^2, \qquad (80)
\end{aligned}$$

where $\delta_0 \equiv \delta(w_0)$ and $\Delta_0 \equiv \mathrm{Diag}(\delta_0)$. Moreover, using the fact that $\mu \leq \mu_0$ together with Propositions 2.1 and 2.2, we conclude that

$$\|\delta_B (\delta_0)_B^{-1}\|_\infty \leq \frac{1+\beta}{(1-\beta)^2} \sqrt{\frac{\mu_0}{\mu}} \left\| \frac{s(\mu)}{s(\mu_0)} \right\|_\infty \leq \frac{(1+\beta)n}{(1-\beta)^2} \sqrt{\frac{\mu_0}{\mu}}, \tag{81}$$

$$\|\delta_N^{-1}(\delta_0)_N\|_\infty \le \frac{1+\beta}{(1-\beta)^2}\sqrt{\frac{\mu}{\mu_0}}\left\|\frac{s_N(\mu_0)}{s_N(\mu)}\right\|_\infty$$

$$\le \frac{1+\beta}{(1-\beta)^2}\sqrt{\frac{\mu_0}{\mu}}\left\|\frac{x_N(\mu)}{x_N(\mu_0)}\right\|_\infty \le \frac{(1+\beta)n}{(1-\beta)^2}\sqrt{\frac{\mu_0}{\mu}}. \tag{82}$$

Inequality (36) now follows by combining the estimates (79), (80), (81), and (82). ∎

Proof of Theorem 2.14. Recall that our goal is to prove that the arithmetic complexity of computing a TR direction during a TR-iteration is bounded by (41). It suffices to examine just the computation of the primal TR direction since the argument for the dual TR direction is analogous. We have seen in the proof of Lemma 2.10 that $\psi_p(\lambda)$ can be expressed as $\psi_p(\lambda) = \|(I+\lambda\tilde{H})^{-1}\tilde{g}\|/\sqrt{\mu}$, where $\tilde{H} \in \Re^{|B|\times|B|}$ and $\tilde{g} \in \Re^{|B|}$ can be computed in $\mathcal{O}(n^3)$ arithmetic operations. It is well known that we can compute an orthogonal matrix Q such that the matrix $T \equiv Q^T\tilde{H}Q$ is tridiagonal in $\mathcal{O}(n^3)$ arithmetic operations. Moreover, using the fact that orthogonal matrices preserve vector lengths, we easily see that $\psi_p(\lambda) = \|(I+\lambda T)^{-1}Q^T\tilde{g}\|/\sqrt{\mu}$. Hence, for any fixed $\lambda > 0$, the fact that T is tridiagonal implies that $\psi_p(\lambda)$ can be computed in $\mathcal{O}(n)$ arithmetic operations. We have thus shown that the arithmetic complexity to compute a TR direction during a TR iteration is bounded by (41). ∎

6. Conclusion. In this paper, we have developed a predictor-corrector, trust-region algorithm for linear programming whose iteration-complexity just depends on $\bar{\chi}_A^*$. The overall arithmetic complexity of the algorithm is not independent of b and/or c, due to work involved in the computation of the trust region steps. An interesting and challenging open question is whether the arithmetic complexity of the PC-TR algorithm, or a variant of it, has an arithmetic complexity that does not depend on b and c.

Appendix. The objective of this section is to provide a proof of Lemma 2.7.

First, we state a technical result whose proof is given in Lemma 4.4 of Monteiro and Tsuchiya [14].

LEMMA A.1. *Let $w = (x,y,s) \in \mathcal{P}^{++} \times \mathcal{D}^{++}$ be given and assume that $\|xs - \nu e\| \le \tau\nu$ for some constants $\tau \in (0,1)$ and $\nu > 0$. Then, $(1-\tau/\sqrt{n})\nu \le \mu(w) \le (1+\tau/\sqrt{n})\nu$ and $w \in \mathcal{N}(\tau/(1-\tau))$.*

We are now ready to prove Lemma 2.7.

Proof of Lemma 2.7. Define $v(\alpha) \equiv (x+\alpha\Delta x)(s+\alpha\Delta s)$ for all $\alpha \in \Re$. We claim that

$$\|v(\alpha) - (1-\alpha)\mu e\| \le \frac{2\beta}{1+2\beta}(1-\alpha)\mu \quad \forall\, 0 \le \alpha \le 1-\bar{\alpha}, \tag{83}$$

where $\mu \equiv \mu(w)$,

$$\varepsilon_2(w) \equiv \max\{\|Rx_N\|, \|Rs_B\|\}, \qquad \bar{\alpha} \equiv \frac{\sqrt{1+\beta}+\gamma}{4\gamma}\varepsilon_2(w). \tag{84}$$

Using this claim, the result can be proved as follows. By Lemma A.1 with $w = w + \alpha\Delta w$, $\nu = (1-\alpha)\mu$ and $\tau = 2\beta/(1+2\beta)$, we conclude from the claim that for any $0 \le \alpha \le 1-\bar{\alpha}$, we have $w + \alpha\Delta w \in \mathcal{N}(2\beta)$ and

$$\mu(w+\alpha\Delta w) \le \left(1 + \frac{2\beta}{\sqrt{n}(1+2\beta)}\right)(1-\alpha)\mu \le 2(1-\alpha)\mu. \tag{85}$$

By the definition of α_τ, we then conclude that $\alpha_\tau \ge 1-\bar{\alpha}$. Setting $\alpha = 1-\bar{\alpha}$ in (85) and using the fact that $\alpha_\tau \ge 1-\bar{\alpha}$ and $\mu(w+\alpha\Delta w)$ is a decreasing (affine) function

of α, we obtain

$$\frac{\mu(w+\alpha_\tau \Delta w)}{\mu(w)} \leq \frac{\sqrt{1+\beta}+\gamma}{2\gamma}\varepsilon_2(w),$$

that is, the result holds.

In the remaining part of the proof, we show that (83) holds. It is easy to see that

$$\begin{aligned} v(\alpha)-(1-\alpha)\mu e &= (x+\alpha\Delta x)(s+\alpha\Delta s)-(1-\alpha)\mu e \\ &= (1-\alpha)(xs-\mu e)+\alpha h^1+\alpha(1-\alpha)h^2+\alpha^2 h^3, \end{aligned} \tag{86}$$

where h^1, h^2, and h^3 are vectors in $\Re^n$ defined as

$$\begin{pmatrix} h^1_B \\ h^1_N \end{pmatrix} \equiv \begin{pmatrix} x_B(s_B+\Delta s_B) \\ s_N(x_N+\Delta x_N) \end{pmatrix} = \mu \begin{pmatrix} w_B p_B \\ w_N p_N \end{pmatrix}, \tag{87}$$

$$\begin{pmatrix} h^2_B \\ h^2_N \end{pmatrix} \equiv \begin{pmatrix} s_B \Delta x_B \\ x_N \Delta s_N \end{pmatrix} = \mu \begin{pmatrix} w_B q_B \\ w_N q_N \end{pmatrix}, \tag{88}$$

$$\begin{pmatrix} h^3_B \\ h^3_N \end{pmatrix} \equiv \begin{pmatrix} \Delta x_B(s_B+\Delta s_B) \\ \Delta s_N(x_N+\Delta x_N) \end{pmatrix} = \mu \begin{pmatrix} p_B q_B \\ p_N q_N \end{pmatrix}. \tag{89}$$

Here, the vectors p, q, and w appearing in the second alternative expressions for h^1, h^2, and h^3 are defined as

$$\begin{pmatrix} p_B \\ p_N \end{pmatrix} \equiv \begin{pmatrix} Rs_B(w) \\ Rx_N(w) \end{pmatrix}, \qquad \begin{pmatrix} q_B \\ q_N \end{pmatrix} \equiv \begin{pmatrix} \Delta_B \Delta x_B/\sqrt{\mu} \\ \Delta_N^{-1}\Delta s_N/\sqrt{\mu} \end{pmatrix}, \qquad w \equiv \frac{x^{1/2}s^{1/2}}{\sqrt{\mu}}.$$

Clearly, in view of (84), (18), and the fact that $w \in \mathcal{N}(\beta)$, we have

$$\|p\|_\infty \leq \varepsilon_2(w), \ \ \|p\| \leq \sqrt{2}\varepsilon_2(w), \quad \|q\| \leq \sqrt{2}\gamma, \ \ \|w\|_\infty \leq \sqrt{1+\beta}, \ \ \|w\| = \sqrt{n}. \tag{90}$$

Using (87), (88), (89), and (90), we obtain

$$\begin{aligned} \|h^1\| \leq \mu\|w\|_\infty \|p\| &\leq \mu\sqrt{2(1+\beta)}\varepsilon_2(w), \\ \|h^2\| \leq \mu\|w\|_\infty \|q\| &\leq \mu\sqrt{2(1+\beta)}\gamma, \\ \|h^3\| \leq \mu\|p\|_\infty \|q\| &\leq \mu\sqrt{2}\gamma\varepsilon_2(w). \end{aligned}$$

Using (86), the triangle inequality for norms, the three estimates above and relations (84) and (17), we then obtain

$$\begin{aligned} \|v(\alpha)-(1-\alpha)\mu e\| &\leq (1-\alpha)\|xs-\mu e\|+\alpha\|h^1\|+\alpha(1-\alpha)\|h^2\|+\alpha^2\|h^3\| \\ &\leq (1-\alpha)\left(\|xs-\mu e\|+\|h^2\|\right)+\|h^1\|+\|h^3\| \\ &\leq \left[(1-\alpha)\left(\beta+\sqrt{2(1+\beta)}\,\gamma\right)+\left(\sqrt{2(1+\beta)}+\sqrt{2}\,\gamma\right)\varepsilon_2(w)\right]\mu \\ &\leq \left[(1-\alpha)\left(\beta+\sqrt{2(1+\beta)}\,\gamma\right)+4\sqrt{2}\gamma\bar{\alpha}\right]\mu \\ &\leq \left[\left(\beta+\sqrt{2(1+\beta)}\,\gamma\right)+4\sqrt{2}\gamma\right](1-\alpha)\mu \\ &\leq \frac{2\beta}{1+2\beta}(1-\alpha)\mu, \end{aligned}$$

for all $0 \leq \alpha \leq 1-\bar{\alpha}$. ∎

REFERENCES

[1] S. Barnett, *Matrices: Methods and Applications*, Oxford Univerisy Press, Oxford, 1990.

[2] I. I. Dikin and V. I. Zorkalcev, *Iterative Solution of Mathematical Programming Problems: Algorithms for the Method of Interior Points* (in Russian), Nauka, Novosibirsk, USSR, 1980.

[3] C. C. Gonzaga and H. J. Lara, *A note on properties of condition numbers*, Linear Algebra Appl., 261 (1997), pp. 269–273.

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

[5] M. Kojima, S. Mizuno, and A. Yoshise, *A primal-dual interior-point algorithm for linear programming*, in Progress in Mathematical Programming, Interior-point and Related Methods, N. Megiddo, ed., Springer-Verlag, New York, 1989, pp. 29–47.

[6] M. Kojima, S. Mizuno, and A. Yoshise, *A polynomial-time algorithm for a class of linear complementarity problems*, Math. Program., 44 (1989), pp. 1–26.

[7] N. Megiddo, *Pathways to the optimal set in linear programming*, Progress in Mathematical Programming, Pacific Grove, CA, 1987, pp. 131–158, Springer, New York-Berlin, 1989.

[8] N. Megiddo, S. Mizuno, and T. Tsuchiya, *A modified layered-step interior-point algorithm for linear programming*, Math. Program., 82 (1998), pp. 339–355.

[9] S. Mizuno, M. J. Todd, and Y. Ye, *On adaptive-step primal-dual interior-point algorithms for linear programming*, Math. Oper. Res., 18 (1993), pp. 964–981.

[10] S. Mizuno, N. Megiddo, and T. Tsuchiya, *A linear programming instance with many crossover events*, J. Complexity, 12 (1996), pp. 474–479.

[11] R. D. C. Monteiro and I. Adler, *Interior path-following primal-dual algorithms. Part* I: *Linear programming*, Math. Program., 44 (1989), pp. 27–41.

[12] R. D. C. Monteiro and I. Adler, *Interior path-following primal-dual algorithms. Part* II: *Convex quadratic programming*, Math. Program., 44 (1989), pp. 43–66.

[13] R. D. C. Monteiro and T. Tsuchiya, *Global convergence of the affine scaling algorithm for convex quadratic programming*, SIAM J. Optim., 8 (1998), pp. 26–58.

[14] R. D. C. Monteiro and T. Tsuchiya, *A variant of the Vavasis-Ye layered-step interior-point algorithm for linear programming*, SIAM J. Optim., 13 (2003), pp. 1054–1079.

[15] R. D. C. Monteiro and T. Tsuchiya, *A new iteration-complexity bound for the MTY predictor-corrector algorithm*, SIAM J. Optim., 15 (2004), pp. 319–347.

[16] R. D. C. Monteiro and T. Tsuchiya, *A strong bound on the integral of the central path curvature and its relationship with the iteration complexity of primal-dual path-following LP algorithms*, working paper, School of ISyE, Georgia Tech, USA, September 2005 (accepted in Mathematical Programming).

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

[18] G. W. Stewart, *On Scaled projections and pseudo-inverses*, Linear Algebra Appl., 112 (1989), pp. 189–193.

[19] K. Tanabe, *Centered Newton method for mathematical programming*, System Modeling and Optimization, M. Iri and K. Yajima, eds., Springer-Verlag, Berlin, 1988, pp. 197–206

[20] É. Tardos, *A strongly polynomial algorithm to solve combinatorial linear programs*, Oper. Res., 34 (1986), pp. 250–256.

[21] M. J. Todd, *A Dantzig-Wolfe-like variant of Karmarkar's interior-point linear programming algorithm*, Oper. Res., 38 (1990), pp. 1006–1018.

[22] M. J. Todd, L. Tunçel, and Y. Ye, *Characterizations, bounds, and probabilistic analysis of two complexity measures for linear programming problems*, Math. Program., 90 (2001), pp. 59–70.

[23] T. Tsuchiya, *Global convergence property of the affine scaling methods for primal degenerate linear programming problems*, Math. Oper. Res., 17 (1992), pp. 527–557.

[24] L. Tunçel, *Primal-dual symmetry and scale invariance of interior-point algorithms for convex optimization*, Math. Oper. Res., 23 (1998), pp. 708–718.

[25] L. Tunçel, *Approximating the complexity measure of Vavasis-Ye algorithm is NP-hard*, Math. Program., 86 (1999), pp. 219–223.

[26] L. Tunçel, *On the condition numbers for polyhedra in Karmarkar's form*, Oper. Res. Lett., 24 (1999), pp. 149–155.

[27] R. J. Vanderbei and J. C. Lagarias, *I. I. Dikin's convergence result for the affine-scaling algorithm, Vol.* 17, Contemp. Math., 114 (1990), pp. 109–119.

[28] S. Vavasis and Y. Ye, *A primal-dual accelerated interior-point method whose running time depends only on A*, Math. Program., 74 (1996), pp. 79–120.

SIAM J. OPTIM.
Vol. 19, No. 4, pp. 1947–1969

A NEW CLASS OF MINIMUM NORM DUALITY THEOREMS*

ACHIYA DAX†

Abstract. In this paper we derive new duality results on the width and the length of symmetrical convex bodies. Let K be a symmetrical convex body in R^m and let $\|\cdot\|$ be some (arbitrary) norm on R^m. The width of K is obtained by searching for the smallest "sandwich" (or "slab") that contains K. It is shown that the dual problem has the following form: Find the largest diameter of a norm ball that is contained in K. Indeed, the diameter of a maximal norm ball equals the width of the smallest sandwich. Moreover, the solutions of the two problems obey certain alignment relations. The length of K is found by searching the largest "sandwich" that contains K. The last problem is closely related to the "maximal chord problem" whose optimal value is called the "diameter" of K. In this case the dual problem is to find the smallest norm ball that contains K. It is proved that the diameter of the smallest norm ball equals the diameter (the length) of K, and that primal and dual solutions satisfy certain alignment relations. Part of the results remain valid for more general convex sets.

Key words. minimum norm duality theorems, dual norm, alignment relations, norm balls, symmetrical convex bodies, depth, width, length, diameter, nearest supporting hyperplane, the smallest sandwich problem, the largest sandwich problem

AMS subject classifications. 49K35, 49N15, 52B12, 52B15, 52B55, 65K05, 65K10, 90C46, 90C47

DOI. 10.1137/070706161

1. Introduction.

The minimum norm duality (MND) theorem considers the distance between a point $\mathbf{z}$ and a closed convex set Y. It says that the shortest distance from $\mathbf{z}$ to Y is equal to the maximum of the distances from $\mathbf{z}$ to any hyperplane separating $\mathbf{z}$ and Y. (See Figure 1.) This fundamental observation gives rise to several useful duality relations in best approximation problems, linear least norm problems, and theorems of the alternative. As far as we know, the first statement of the MND theorem is due to Nirenberg [21], who established this assertion in any normed linear space by applying the Hahn–Banach theorem. The name "MND theorem" was coined by Luenberger [19], who also derived the "alignment" relation between primal and dual solutions.

Recently Dax [8] extended the MND theorem to consider the distance between two convex sets. Roughly speaking the new theorem says that the shortest distance between the two sets is equal to the maximal "separation" between the sets, where the term "separation" refers to the distance between a pair of parallel hyperplanes that separates the two sets. (See Figure 2.)

Another extension of the MND theorem is proposed by Briec [5], who considers the case when $\mathbf{z}$ is an interior point of Y. This theorem says that the shortest distance from $\mathbf{z}$ to any supporting hyperplane of Y is equal to the "depth" of $\mathbf{z}$. (See Figure 4.) The depth of $\mathbf{z}$ is defined as the shortest distance from $\mathbf{z}$ to any boundary point of Y. In other words, the depth of $\mathbf{z}$ is equal to the radius of the largest norm ball that is centered at $\mathbf{z}$ and contained in Y.

In this paper we establish a new class of duality theorems, one that considers the "width" and the "length" of a convex body. Let K be a convex body in R^m. That

*Received by the editors October 23, 2007; accepted for publication (in revised form) November 10, 2008; published electronically February 27, 2009.
http://www.siam.org/journals/siopt/19-4/70616.html
†Hydrological Service, P.O.B. 36118, Jerusalem 91360, Israel (dax20@water.gov.il).

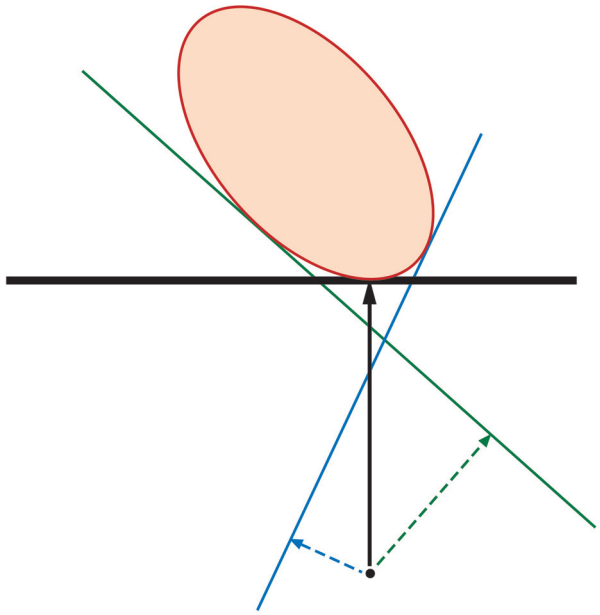


FIG. 1. *The Nirenberg–Luenberger MND theorem.*

is, K is a closed bounded convex set that contains interior points. Let $\|\cdot\|$ be some (arbitrary) norm on R^m and let $\|\cdot\|'$ denote the corresponding dual norm. Then for any nonzero vector $\mathbf{a} \in R^m$ there exists a pair of points, $\mathbf{y}_1 \in K$ and $\mathbf{y}_2 \in K$, such that

$$\mathbf{a}^T\mathbf{y}_1 = \inf_{\mathbf{x}\in K} \mathbf{a}^T\mathbf{x} \quad \text{and} \quad \mathbf{a}^T\mathbf{y}_2 = \sup_{\mathbf{x}\in K} \mathbf{a}^T\mathbf{x}. \tag{1.1}$$

Consequently K is "sandwiched" between the parallel hyperplanes

$$H_1 = \left\{ \mathbf{x} \,\middle|\, \mathbf{a}^T\mathbf{x} = \mathbf{a}^T\mathbf{y}_1 \right\} \quad \text{and} \quad H_2 = \left\{ \mathbf{x} \,\middle|\, \mathbf{a}^T\mathbf{x} = \mathbf{a}^T\mathbf{y}_2 \right\}. \tag{1.2}$$

The distance between a pair of parallel hyperplanes of this form is defined as

$$\inf \left\{ \|\mathbf{x}_1 - \mathbf{x}_2\| \,\middle|\, \mathbf{x}_1 \in H_1,\, \mathbf{x}_2 \in H_2 \right\}.$$

Using this definition one can show that the width function

$$\omega(\mathbf{a}) = (\mathbf{a}^T\mathbf{y}_2 - \mathbf{a}^T\mathbf{y}_1)/\|\mathbf{a}\|' \tag{1.3}$$

measures the distance between the two hyperplanes in (1.2). See section 2. The set

$$S(\mathbf{a}, K) = \left\{ \mathbf{x} \,\middle|\, \inf_{\mathbf{x}\in K} \mathbf{a}^T\mathbf{x} \le \mathbf{a}^T\mathbf{x} \le \sup_{\mathbf{x}\in K} \mathbf{a}^T\mathbf{x} \right\}$$

is said to be a "sandwich" (or "slab") of K, and $\omega(\mathbf{a})$ provides the "width" of this sandwich. The smallest sandwich of K is constructed, therefore, by a vector $\mathbf{a}$ for which $\omega(\mathbf{a})$ attains the smallest possible value. Similarly, the largest sandwich of K is constructed by a vector $\mathbf{a}$ for which $\omega(\mathbf{a})$ attains the largest possible value. The

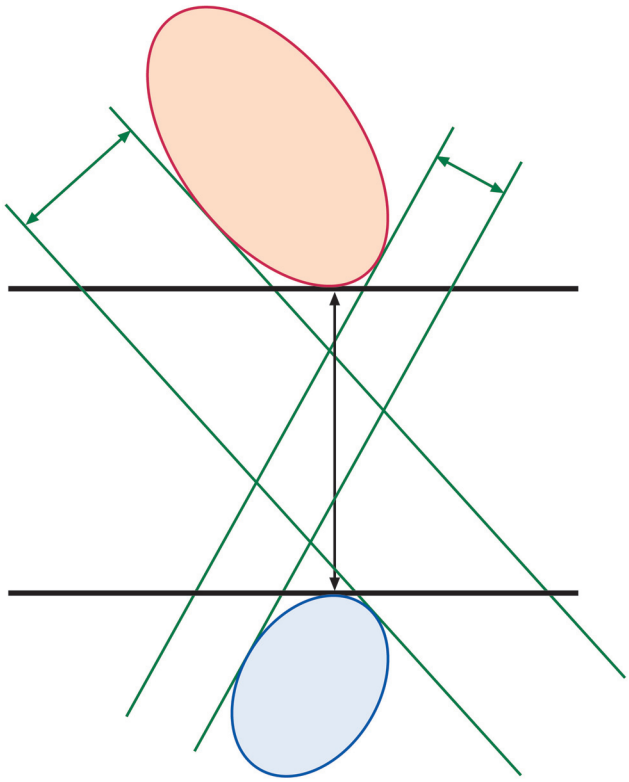


FIG. 2. *Dax MND theorem.*

smallest value of $\omega(\mathbf{a})$ is called the "width" of K. The largest value of $\omega(\mathbf{a})$ is called the "length" of K.

The paper explores the duality properties that characterize the smallest and largest sandwiches of K. The strongest duality results require K to be a symmetrical convex body, for which there exists a point $\mathbf{c} \in K$ with the following property: For any $\mathbf{x} \in K$ the point $\mathbf{y} = 2\mathbf{c} - \mathbf{x}$ belongs to K. The point $\mathbf{c}$ is called the "center of symmetry" of K, or simply the "center" of K. The point $\mathbf{y} = 2\mathbf{c} - \mathbf{x}$ is called the "symmetrical image" of $\mathbf{x}$.

The plan of our paper is as follows. It starts with a brief overview of the basic alignment relations on a norm ball. (See Figure 3.) These relations provide the tools for proving our duality results. The third section considers the nearest hyperplane problem, giving a new simple proof to the Briec theorem [5]. The "opposite" problem is to find a supporting hyperplane of K which is the farthest away from a given point of K. (See Figures 5 and 6.) This problem is discussed in section 4. The duality features that characterize the smallest possible sandwich are derived in section 5. It is shown there that the dual problem has the following form: Find the largest norm ball that is contained in K. (See Figure 7.) The largest possible sandwich is considered in section 6. In this case the dual problem is to find the smallest norm ball that contains K. (See Figure 8.) The duality results that we prove require K to be a symmetrical convex body. Yet, as section 7 shows, part of the results remain valid without symmetry.

2. Alignment relations on a norm ball. Here and henceforth $\|\cdot\|$ denotes some (arbitrary) norm on R^m and $\|\cdot\|'$ denotes the corresponding dual norm. Recall that the dual norm is defined by the rule

$$\|\mathbf{v}\|' = \sup_{\mathbf{u}\in U} \mathbf{u}^T\mathbf{v}, \tag{2.1}$$

where

$$U = \{\, \mathbf{x} \mid \|\mathbf{x}\| \le 1 \,\} \tag{2.2}$$

denotes the unit norm ball in R^m. Recall also that two nonzero vectors $\mathbf{u} \in R^m$ and $\mathbf{v} \in R^m$ are aligned if they satisfy

$$\mathbf{u}^T\mathbf{v} = \|\mathbf{u}\| \cdot \|\mathbf{v}\|'. \tag{2.3}$$

In this case the supremum in (2.1) is obtained for the unit vector $\mathbf{u}^* = \mathbf{u}/\|\mathbf{u}\|$. Moreover, since U is a compact set, there exists a vector $\mathbf{u}^* \in U$ that solves the problem

$$\begin{aligned} &\text{maximize} && \mathbf{u}^T\mathbf{v} \\ &\text{subject to} && \mathbf{u} \in U, \end{aligned}$$

and this vector satisfies $\|\mathbf{u}^*\| = 1$ and $(\mathbf{u}^*)^T\mathbf{v} = \|\mathbf{v}\|'$. That is, the vectors $\mathbf{u}^*$ and $\mathbf{v}$ are aligned and the hyperplane

$$H = \{\, \mathbf{x} \mid \mathbf{v}^T\mathbf{x} = \mathbf{v}^T\mathbf{u}^* \,\}$$

supports U at $\mathbf{u}^*$. See Figure 3.

Similar alignment relations exist on any norm ball in R^m. Let

$$B = \{\, \mathbf{x} \mid \|\mathbf{x} - \mathbf{z}\| \le r \,\}$$

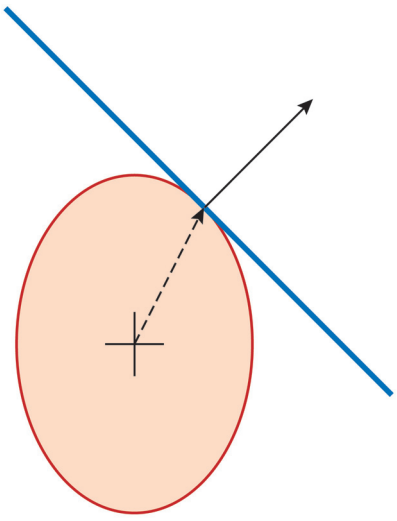


FIG. 3. *Alignment relations on a norm ball.*

be a norm ball of radius r that is centered at some point $\mathbf{z} \in R^m$. Then for any vector $\mathbf{a} \in R^m$, $\mathbf{a} \neq \mathbf{0}$, there exists a boundary point of B, say $\mathbf{y}_1$, such that

$$\mathbf{a}^T\mathbf{y}_1 = \sup_{\mathbf{x}\in B} \mathbf{a}^T\mathbf{x}, \tag{2.4}$$

and

$$H_1 = \{\, \mathbf{x} \mid \mathbf{a}^T\mathbf{x} = \mathbf{a}^T\mathbf{y}_1 \,\}$$

is a supporting hyperplane of B at $\mathbf{y}_1$. A further consequence of (2.4) is that the vectors $\mathbf{u} = \mathbf{y}_1 - \mathbf{z}$ and $\mathbf{a}$ are aligned, and that $\|\mathbf{u}\| = r$. In other words,

$$\mathbf{u}^T\mathbf{a} = \|\mathbf{u}\| \cdot \|\mathbf{a}\|' = r\|\mathbf{a}\|'. \tag{2.5}$$

Recall that the distance between $\mathbf{z}$ and H_1 is defined as

$$d(\mathbf{z}, H_1) = \inf\{\, \|\mathbf{x} - \mathbf{z}\| \mid \mathbf{x} \in H_1 \,\}.$$

Therefore, since H_1 is a supporting hyperplane of B, this distance equals r and satisfies

$$\operatorname{dist}(\mathbf{z}, H_1) = r = \mathbf{a}^T\mathbf{u}/\|\mathbf{a}\|' = (\mathbf{a}^T\mathbf{y}_1 - \mathbf{a}^T\mathbf{z})/\|\mathbf{a}\|'.$$

The last observation can be generalized in two ways. Let

$$H_\alpha = \{\, \mathbf{x} \mid \mathbf{a}^T\mathbf{x} = \alpha \,\}$$

be a hyperplane in R^m. Then the distance between $\mathbf{z}$ and H_α satisfies the following rule:

$$\operatorname{dist}(\mathbf{z}, H_\alpha) = (\alpha - \mathbf{a}^T\mathbf{z})/\|\mathbf{a}\|' \quad \text{when} \quad \mathbf{a}^T\mathbf{z} \leq \alpha,$$

and

$$\operatorname{dist}(\mathbf{z}, H_\alpha) = (\mathbf{a}^T\mathbf{z} - \alpha)/\|\mathbf{a}\|' \quad \text{when} \quad \mathbf{a}^T\mathbf{z} \geq \alpha.$$

Moreover, let

$$H_\beta = \{\, \mathbf{x} \mid \mathbf{a}^T\mathbf{x} = \beta \,\}$$

be some other hyperplane that is parallel to H_α and satisfies $\beta > \alpha$. Then for any point $\mathbf{x} \in H_\beta$

$$\operatorname{dist}(\mathbf{x}, H_\alpha) = (\mathbf{a}^T\mathbf{x} - \alpha)/\|\mathbf{a}\|' = (\beta - \alpha)/\|\mathbf{a}\|'.$$

Consequently

$$\operatorname{dist}(H_\beta, H_\alpha) = (\beta - \alpha)/\|\mathbf{a}\|', \tag{2.6}$$

which proves (1.3).

Another consequence that stems from (2.5) is related to the width of the sandwich $S(\mathbf{a}, B)$. Using the symmetry of B with respect to $\mathbf{z}$, one can verify that the point

$$\mathbf{y}_2 = \mathbf{z} - \mathbf{u} = 2\mathbf{z} - \mathbf{y}_1,$$

which is the "symmetrical image" of $\mathbf{y}_1$, satisfies

$$(-\mathbf{a})^T\mathbf{y}_2 = \sup_{\mathbf{x}\in B}(-\mathbf{a})^T\mathbf{x}.$$

The last equality means that the hyperplane

$$H_2 = \left\{\, \mathbf{x} \;\middle|\; (-\mathbf{a})^T\mathbf{x} = (-\mathbf{a})^T\mathbf{y}_2 \,\right\} = \left\{\, \mathbf{x} \;\middle|\; \mathbf{a}^T\mathbf{x} = \mathbf{a}^T\mathbf{y}_2 \,\right\}$$

supports B at $\mathbf{y}_2$, and that

$$\mathbf{a}^T\mathbf{y}_2 = \inf_{\mathbf{x}\in B}\mathbf{a}^T\mathbf{x}.$$

Therefore, since H_1 and H_2 are parallel hyperplanes,

$$\operatorname{dist}(H_1, H_2) = (\mathbf{a}^T\mathbf{y}_1 - \mathbf{a}^T\mathbf{y}_2)/\|\mathbf{a}\|' = 2\mathbf{u}^T\mathbf{a}/\|\mathbf{a}\|' = 2r,$$

where the last equality comes from (2.5). In other words, for any norm ball, B, and any vector $\mathbf{a} \in R^m$, $\mathbf{a} \neq \mathbf{0}$, the width of the sandwich $S(\mathbf{a}, B)$ equals the diameter of B.

3. The nearest supporting hyperplane. Let X be a closed convex set in R^m with nonempty interior and nonempty boundary. A hyperplane $H = \{\mathbf{x} \mid \mathbf{a}^T\mathbf{x} = \alpha\}$ is said to be a supporting hyperplane of X if $\mathbf{a} \neq \mathbf{0}$ and $\alpha = \sup_{\mathbf{x}\in X}\mathbf{a}^T\mathbf{x}$ attains a finite value. If, in addition, there exists a boundary point $\mathbf{y} \in X$ such that

$$\mathbf{a}^T\mathbf{y} = \alpha = \sup_{\mathbf{x}\in X}\mathbf{a}^T\mathbf{x},$$

then H is said to be a supporting hyperplane of X at $\mathbf{y}$. Let $\mathbf{z}$ be a given interior point of X. Then, as we have seen, the distance between $\mathbf{z}$ and H equals $(\alpha - \mathbf{a}^T\mathbf{z})/\|\mathbf{a}\|'$. The nearest supporting hyperplane is defined, therefore, by a vector $\mathbf{a} \in R^m$, $\mathbf{a} \neq \mathbf{0}$, that solves the minimization problem

$$\text{minimize} \quad \nu(\mathbf{a}) = \left(\sup_{\mathbf{x}\in X}\mathbf{a}^T\mathbf{x} - \mathbf{a}^T\mathbf{z}\right)/\|\mathbf{a}\|',$$

or

$$\begin{aligned} &\text{minimize} && \xi(\mathbf{a}) = \sup_{\mathbf{x}\in X}\mathbf{a}^T\mathbf{x} - \mathbf{a}^T\mathbf{z} \\ &\text{subject to} && \|\mathbf{a}\|' = 1. \end{aligned} \tag{3.1}$$

Below we will show that (3.1) is closely related to the problem of calculating the "depth" of $\mathbf{z}$, which is the radius of the largest norm ball that is centered in $\mathbf{z}$ and contained in X.

The formal definition of the "depth" function is

$$\operatorname{depth}(\mathbf{z}) = \inf\left\{\, \|\mathbf{y} - \mathbf{z}\| \;\middle|\; \mathbf{y} \in \tilde{X} \,\right\},$$

where $\tilde{X}$ denotes the boundary of X. Since X is a closed set, $\tilde{X} \subseteq X$ and there exists a point $\mathbf{y}^* \in \tilde{X}$ such that

$$\operatorname{depth}(\mathbf{z}) = \|\mathbf{y}^* - \mathbf{z}\|.$$

In other words, $\mathbf{y}^*$ solves the depth problem

$$\begin{aligned} &\text{minimize} && \rho(\mathbf{y}) = \|\mathbf{y} - \mathbf{z}\| \\ &\text{subject to} && \mathbf{y} \in \tilde{X}, \end{aligned} \tag{3.2}$$

and

$$B_{\mathbf{z}}^* = \left\{ \mathbf{x} \;\middle|\; \|\mathbf{x} - \mathbf{z}\| \leq \text{depth}(\mathbf{z}) \right\}$$

is the largest norm ball that is centered in $\mathbf{z}$ and contained in X.

Recall that for any boundary point $\mathbf{y} \in \tilde{X}$ one can find a hyperplane H that supports X at $\mathbf{y}$; see, e.g., [8]. Thus, in particular, there exists a vector $\mathbf{a}^* \in R^m$, $\|\mathbf{a}^*\|' = 1$, for which the hyperplane

$$H^* = \left\{ \mathbf{x} \;\middle|\; (\mathbf{a}^*)^T \mathbf{x} = (\mathbf{a}^*)^T \mathbf{y}^* \right\}$$

supports X at $\mathbf{y}^*$. That is,

$$(\mathbf{a}^*)^T \mathbf{y}^* = \sup_{\mathbf{x} \in X} (\mathbf{a}^*)^T \mathbf{x}. \tag{3.3}$$

Observe also that H^* is a supporting hyperplane of $B_{\mathbf{z}}^*$ at $\mathbf{y}^*$. One consequence of this observation is that the vectors $\mathbf{u} = \mathbf{y}^* - \mathbf{z}$ and $\mathbf{a}^*$ are aligned. That is,

$$\mathbf{u}^T \mathbf{a}^* = \|\mathbf{u}\| \, \|\mathbf{a}^*\|'. \tag{3.4}$$

A second consequence is that the distance between $\mathbf{z}$ and H^* equals depth($\mathbf{z}$).

On the other hand, the distance between $\mathbf{z}$ and any other supporting hyperplane of X exceeds depth($\mathbf{z}$), since $B_{\mathbf{z}}^*$ is contained in the "negative half-space" of that hyperplane. The last observation means that H^* solves the nearest hyperplane problem. Moreover, any other supporting hyperplane of X that solves this problem is forced to be a supporting hyperplane of $B_{\mathbf{z}}^*$ at some boundary point of $B_{\mathbf{z}}^*$. The next statement summarizes our findings.

THEOREM 1. *The nearest hyperplane problem* (3.1) *and the depth problem* (3.2) *are solvable and share the same optimal value, which is depth*($\mathbf{z}$). *Let* $\mathbf{y}^*$ *solve* (3.2)*; then there exists a vector* $\mathbf{a}^* \in R^m$ *for which the corresponding supporting hyperplane of* X *satisfies* (3.3)–(3.4) *and solves* (3.1). *Conversely, let* $\mathbf{a}^*$ *define a supporting hyperplane* H^* *of* X *that solves* (3.1). *Then* H^* *is also a supporting hyperplane of* $B_{\mathbf{z}}^*$ *at some boundary point,* $\mathbf{y}^*$*, of* $B_{\mathbf{z}}^*$. *Moreover, in both cases the vectors* $\mathbf{u} = \mathbf{y}^* - \mathbf{z}$ *and* $\mathbf{a}^*$ *are aligned.*

Theorem 1 is essentially the Briec theorem in R^m. The original proof in [5] uses the Edelheit separation theorem and the Nirenberg–Luenberger MND theorem. The current proof replaces these theorems with simple geometric arguments.

An important consequence of Theorem 1 is that searching for a point which maximizes depth($\mathbf{z}$) is equivalent to searching for a point which is farthest from any supporting hyperplane of X. Note also that both (3.1) and (3.2) may have more than one solution. Yet, as we have seen, any solution of one problem is related to a certain solution of the other problem. Methods for solving the nearest hyperplane problem are discussed in [5], [9], and [27]. In the next section we derive analogous results for the farthest supporting hyperplane.

4. The farthest supporting hyperplane. Here X denotes a convex body in R^m and $\mathbf{z}$ is a given interior point of X. Let $\mathbf{a}$ be some nonzero vector in R^m, and

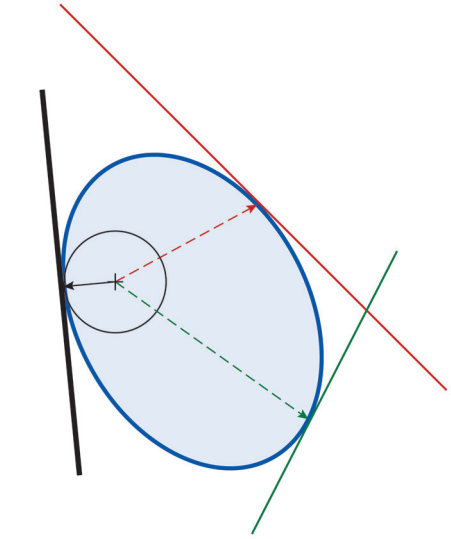


FIG. 4. *The nearest supporting hyperplane (Briec theorem).*

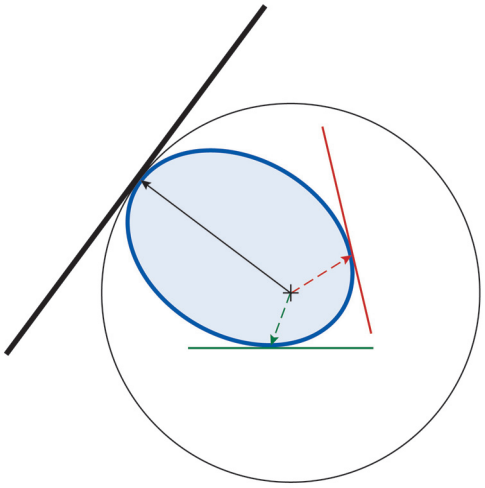


FIG. 5. *The farthest supporting hyperplane and the farthest point.*

let $H = \{\mathbf{x} \mid \mathbf{a}^T\mathbf{x} = \alpha\}$ denote the corresponding supporting hyperplane of X. Then, as we have seen,

$$\alpha = \sup_{\mathbf{x}\in X} \mathbf{a}^T\mathbf{x},$$

and the distance between $\mathbf{z}$ and H equals $(\alpha - \mathbf{a}^T\mathbf{z})/\|\mathbf{a}\|'$. The problem of calculating a supporting hyperplane of X whose distance from $\mathbf{z}$ is maximal can be posed, therefore, as

$$\text{maximize} \quad \varphi(\mathbf{a}) = \Big(\sup_{\mathbf{x}\in X} \mathbf{a}^T\mathbf{x} - \mathbf{a}^T\mathbf{z}\Big)/\|\mathbf{a}\|'$$

or

$$\begin{aligned} &\text{maximize} \quad \psi(\mathbf{a}) = \sup_{\mathbf{x}\in X} \mathbf{a}^T\mathbf{x} - \mathbf{a}^T\mathbf{z} \\ &\text{subject to} \quad \|\mathbf{a}\|' = 1. \end{aligned} \tag{4.1}$$

Below we will show that this problem is related to the "radius" of $\mathbf{z}$, which is the radius of the smallest norm ball that is centered at $\mathbf{z}$ and contains X.

The formal definition of the "radius" function is

$$\text{radius}(\mathbf{z}) = \sup\left\{ \|\mathbf{y} - \mathbf{z}\| \;\middle|\; \mathbf{y} \in X \right\}.$$

Since X is a convex body, there exists a point, $\hat{\mathbf{y}} \in X$, such that

$$\text{radius}(\mathbf{z}) = \|\hat{\mathbf{y}} - \mathbf{z}\|.$$

In other words, $\hat{\mathbf{y}}$ solves the radius problem

$$\begin{aligned} &\text{maximize} \quad \rho(\mathbf{y}) = \|\mathbf{y} - \mathbf{z}\| \\ &\text{subject to} \quad \mathbf{y} \in X. \end{aligned} \tag{4.2}$$

It is also easy to verify that $\hat{\mathbf{y}}$ is a boundary point of X, and that

$$\hat{B}_{\mathbf{z}} = \left\{ \mathbf{x} \;\middle|\; \|\mathbf{x} - \mathbf{z}\| \le \text{radius}(\mathbf{z}) \right\}$$

is the smallest norm ball that is centered at $\mathbf{z}$ and contains X. Consequently there exists a vector $\hat{\mathbf{a}} \in R^m$, $\|\hat{\mathbf{a}}\|' = 1$, for which the hyperplane

$$\hat{H} = \left\{ \mathbf{x} \;\middle|\; \hat{\mathbf{a}}^T\mathbf{x} = \hat{\mathbf{a}}^T\hat{\mathbf{y}} \right\}$$

supports $\hat{B}_{\mathbf{z}}$ at $\hat{\mathbf{y}}$. Therefore, since $\hat{B}_{\mathbf{z}}$ contains X,

$$\hat{\mathbf{a}}^T\hat{\mathbf{y}} = \sup_{\mathbf{x}\in\hat{B}_{\mathbf{z}}} \hat{\mathbf{a}}^T\mathbf{x} = \sup_{\mathbf{x}\in X} \hat{\mathbf{a}}^T\mathbf{x}, \tag{4.3}$$

which means that $\hat{H}$ is also a supporting hyperplane of X at $\hat{\mathbf{y}}$. The distance between $\mathbf{z}$ and $\hat{H}$ is, clearly, radius($\mathbf{z}$). On the other hand, let $\tilde{H}$ be another hyperplane that supports X at some point $\tilde{\mathbf{y}} \in X$. Then, since $\tilde{\mathbf{y}} \in \hat{B}_{\mathbf{z}}$, the distance between $\mathbf{z}$ and $\tilde{H}$ does not exceed radius($\mathbf{z}$). These observations bring us to the following conclusions.

THEOREM 2. *The farthest hyperplane problem* (4.1) *and the radius problem* (4.2) *are solvable and share the same optimal value, which is radius*($\mathbf{z}$). *Let* $\hat{\mathbf{y}}$ *solve* (4.2)*; then there exists a vector* $\hat{\mathbf{a}} \in R^m$ *that solves* (4.1) *and satisfies* (4.3). *Conversely, let* $\hat{\mathbf{a}} \in R^m$ *solve* (4.1)*; then there exists a boundary point,* $\hat{\mathbf{y}} \in X$*, which is also a boundary point of* $\hat{B}_{\mathbf{z}}$*, that solves* (4.2) *and satisfies* (4.3). *Moreover, in both cases the vectors* $\mathbf{u} = \hat{\mathbf{y}} - \mathbf{z}$ *and* $\hat{\mathbf{a}}$ *are aligned. That is,*

$$\mathbf{u}^T\hat{\mathbf{a}} = \|\mathbf{u}\| \cdot \|\hat{\mathbf{a}}\|' = \|\hat{\mathbf{y}} - \mathbf{z}\|.$$

Methods for solving the farthest hyperplane problem are discussed in [9]. One consequence of Theorem 2 is that searching for a point which minimizes radius($\mathbf{z}$) is equivalent to searching for a point that has the smallest distance to its farthest supporting hyperplane. Note also that Theorem 2 remains valid when X is an arbitrary compact set in R^m and $\mathbf{z}$ is an arbitrary point of R^m. Yet in this case the objective function of (4.1) may attain negative values and the terms "supporting hyperplane" or "farthest hyperplane problem" are not quite appropriate.

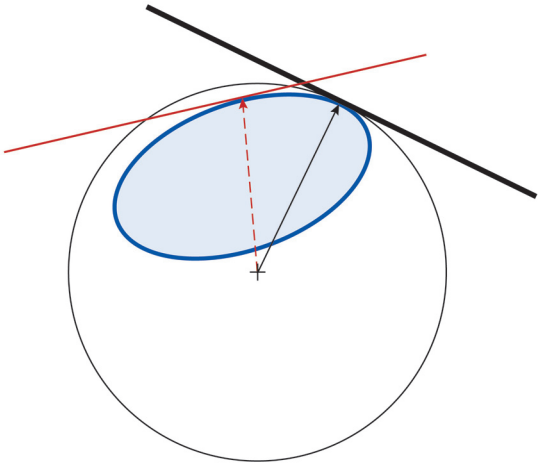


FIG. 6. *The farthest supporting hyperplane and the farthest point.*

5. The smallest sandwich problem. Let K denote a symmetrical convex body in R^m. As mentioned in the introduction, the width function (1.3) measures the distance between a pair of parallel supporting hyperplanes that "sandwich" K. The smallest sandwich problem is, therefore, to find a vector $\mathbf{a} \in R^m$, $\mathbf{a} \neq \mathbf{0}$, that solves the problem

$$\text{minimize} \quad w(\mathbf{a}) = \Big(\sup_{\mathbf{x}\in K} \mathbf{a}^T\mathbf{x} - \inf_{\mathbf{x}\in K} \mathbf{a}^T\mathbf{x} \Big) / \|\mathbf{a}\|'$$

or

$$\begin{aligned} &\text{minimize} && \sigma(\mathbf{a}) = \sup_{\mathbf{x}\in K} \mathbf{a}^T\mathbf{x} - \inf_{\mathbf{x}\in K} \mathbf{a}^T\mathbf{x} \\ &\text{subject to} && \|\mathbf{a}\|' = 1. \end{aligned} \tag{5.1}$$

The aim of this section is to show that the dual problem of (5.1) has the form

$$\begin{aligned} &\text{maximize} && \eta(\mathbf{z}) = 2\,\text{depth}(\mathbf{z}) \\ &\text{subject to} && \mathbf{z} \in K. \end{aligned} \tag{5.2}$$

Recall that

$$\text{depth}(\mathbf{z}) = \inf_{\mathbf{x}\in \tilde{K}} \|\mathbf{x} - \mathbf{z}\|, \tag{5.3}$$

where $\tilde{K}$ denotes the boundary of K. Thus $\eta(\mathbf{z})$ is the diameter of the largest norm ball that is centered at $\mathbf{z}$ and contained in K. Maximizing $\eta(\mathbf{z})$ over K means, therefore, that we search for the largest possible diameter of a norm ball that is contained in K. (See Figure 7.)

Let $\mathbf{z}$ be some interior point of K. Then the width function of the norm ball

$$B_{\mathbf{z}} = \big\{ \mathbf{x} \;\big|\; \|\mathbf{x} - \mathbf{z}\| \leq \text{depth}(\mathbf{z}) \big\}$$

has a constant value, which is $\eta(\mathbf{z})$. In other words, for any vector $\mathbf{a} \in R^m$, $\|\mathbf{a}\|' = 1$, the width of the sandwich $S(\mathbf{a}, B_{\mathbf{z}})$ equals $\eta(\mathbf{z})$. On the other hand, since $B_{\mathbf{z}}$ is contained in K, the width of $S(\mathbf{a}, B_{\mathbf{z}})$ is smaller than the width of $S(\mathbf{a}, K)$. That is,

$$\eta(\mathbf{z}) \leq \sigma(\mathbf{a}). \tag{5.4}$$

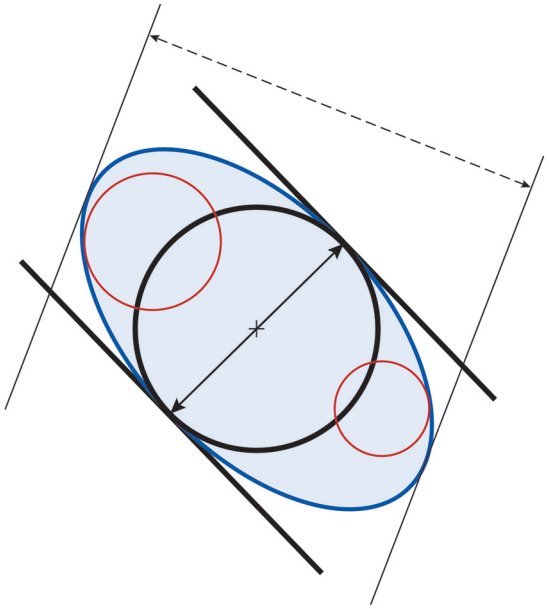


FIG. 7. *The width of the smallest sandwich equals the diameter of the largest norm ball.*

The symmetry of K implies the existence of a "center point," $\mathbf{c}$, that has the following property: For any $\mathbf{x} \in K$ the point $2\mathbf{c} - \mathbf{x}$ (the "symmetrical image" of $\mathbf{x}$) belongs to K. Thus, in particular, for any interior point $\mathbf{z}$ the symmetrical image of $B_{\mathbf{z}}$ is contained in K. Now the convexity of K implies that the norm ball

$$\{\, \mathbf{x} \mid \|\mathbf{x} - \mathbf{c}\| \leq \operatorname{depth}(\mathbf{z}) \,\}$$

is contained in K, so

$$\eta(\mathbf{z}) \leq \eta(\mathbf{c}). \tag{5.5}$$

The last inequality means that $\mathbf{c}$ solves (5.2). That is, the norm ball

$$B_{\mathbf{c}} = \{\, \mathbf{x} \mid \|\mathbf{x} - \mathbf{c}\| \leq \operatorname{depth}(\mathbf{c}) \,\}$$

has the largest possible radius. Yet, as Figure 9 shows, $\mathbf{c}$ is not necessarily the only solution of (5.2).

Let $\tilde{B}_{\mathbf{c}}$ denote the boundary of $B_{\mathbf{c}}$. Then, as we have seen in section 3, there exist a boundary point $\mathbf{y}^* \in \tilde{B}_{\mathbf{c}} \cap \tilde{K}$ and a vector $\mathbf{a}^*$, $\|\mathbf{a}^*\|' = 1$, such that

$$(\mathbf{a}^*)^T \mathbf{y}^* = \sup_{\mathbf{x} \in K} (\mathbf{a}^*)^T \mathbf{x},$$

the hyperplane $H^* = \{\mathbf{x} \mid (\mathbf{a}^*)^T \mathbf{x} = (\mathbf{a}^*)^T \mathbf{y}^*\}$ supports both K and $B_{\mathbf{c}}$ at $\mathbf{y}^*$, and the vectors $\mathbf{u} = \mathbf{y}^* - \mathbf{c}$ and $\mathbf{a}^*$ are aligned. Let the vector $\mathbf{v}^* = 2\mathbf{c} - \mathbf{y}^*$ denote the symmetrical image of $\mathbf{y}^*$. Then, clearly, $\mathbf{v}^* \in \tilde{B}_{\mathbf{c}} \cap \tilde{K}$ and

$$(\mathbf{a}^*)^T \mathbf{v}^* = \inf_{\mathbf{x} \in K} (\mathbf{a}^*)^T \mathbf{x}.$$

These observations mean that $S(\mathbf{a}^*, K) = S(\mathbf{a}^*, B_{\mathbf{c}})$ and $\sigma(\mathbf{a}^*) = \eta(\mathbf{c})$, while (5.4) implies that $\mathbf{a}^*$ solves (5.1).

Finally we extend the above relations to any pair of primal-dual solutions. Let $\hat{\mathbf{a}} \in R^m$ solve (5.1) and let $\hat{\mathbf{z}} \in R^m$ solve (5.2). Then, as we have seen,

$$\eta(\hat{\mathbf{z}}) = \eta(\mathbf{c}) = \sigma(\hat{\mathbf{a}}).$$

Furthermore, since the norm ball

$$B_{\hat{\mathbf{z}}} = \left\{ \mathbf{x} \;\middle|\; \|\mathbf{x} - \hat{\mathbf{z}}\| \le \operatorname{depth}(\hat{\mathbf{z}}) \right\}$$

is contained in K, the sandwich $S(\hat{\mathbf{a}}, B_{\hat{\mathbf{z}}})$ is contained in $S(\hat{\mathbf{a}}, K)$. Yet the optimality of $\hat{\mathbf{a}}$ and $\hat{\mathbf{z}}$ indicates that the two sandwiches have the same width and, therefore, coincide. That is, $S(\hat{\mathbf{a}}, B_{\hat{\mathbf{z}}}) = S(\hat{\mathbf{a}}, K)$. The last equality implies the existence of a common boundary point, $\hat{\mathbf{y}} \in \tilde{B}_{\hat{\mathbf{z}}} \cap \tilde{K}$, such that

$$\hat{\mathbf{a}}^T \hat{\mathbf{y}} = \sup_{\mathbf{x} \in B_{\hat{\mathbf{z}}}} \hat{\mathbf{a}}^T \mathbf{x} = \sup_{\mathbf{x} \in K} \hat{\mathbf{a}}^T \mathbf{x}, \tag{5.6}$$

and the vectors $\hat{\mathbf{u}} = \hat{\mathbf{y}} - \hat{\mathbf{z}}$ and $\hat{\mathbf{a}}$ are aligned. Moreover, since $\hat{\mathbf{z}}$ serves as a "center of symmetry" for $B_{\hat{\mathbf{z}}}$, the point $\hat{\mathbf{v}} = 2\hat{\mathbf{z}} - \hat{\mathbf{y}}$ belongs to $\tilde{B}_{\hat{\mathbf{z}}} \cap \tilde{K}$ and satisfies

$$\hat{\mathbf{a}}^T \hat{\mathbf{v}} = \inf_{\mathbf{x} \in B_{\hat{\mathbf{z}}}} \hat{\mathbf{a}}^T \mathbf{x} = \inf_{\mathbf{x} \in K} \hat{\mathbf{a}}^T \mathbf{x}. \tag{5.7}$$

The next theorem summarizes our conclusions.

THEOREM 3. *The dual of the smallest sandwich problem* (5.1) *is the maximal norm ball problem* (5.2)*, and both problems are solvable. In particular, the center of symmetry,* $\mathbf{c}$*, solves* (5.2)*. Let* $\hat{\mathbf{a}}$ *and* $\hat{\mathbf{z}}$ *be any pair of primal-dual solutions. Then* $\eta(\hat{\mathbf{z}}) = \eta(\mathbf{c}) = \sigma(\hat{\mathbf{a}})$ *and* $S(\hat{\mathbf{a}}, B_{\hat{\mathbf{z}}}) = S(\hat{\mathbf{a}}, K)$*. Furthermore, there exist common boundary points,* $\hat{\mathbf{y}}$ *and* $\hat{\mathbf{v}} = 2\hat{\mathbf{z}} - \hat{\mathbf{y}}$*, that belong to* $\tilde{B}_{\hat{\mathbf{z}}} \cap \tilde{K}$ *and satisfy* (5.6) *and* (5.7)*, respectively. That is, the two points lie on opposite sides of* $S(\hat{\mathbf{a}}, K)$*. The vectors* $\hat{\mathbf{u}} = \hat{\mathbf{z}} - \hat{\mathbf{y}}$ *and* $\hat{\mathbf{a}}$ *are aligned. That is,*

$$\hat{\mathbf{u}}^T \hat{\mathbf{a}} = \|\hat{\mathbf{u}}\| \cdot \|\hat{\mathbf{a}}\|' = \operatorname{depth}(\mathbf{c}). \tag{5.8}$$

The proof of Theorem 3 does not rely on the assumption that K is a bounded set. This brings us to the following conclusion.

COROLLARY 4. *Let* X *be a symmetrical closed convex set in* R^m*, with nonempty interior and nonempty boundary. Then the claims of Theorem* 3 *remain valid when* X *replaces* K*. (Note that* X *may have more than one center of symmetry.)*

6. The largest sandwich problem. As before K denotes a symmetrical convex body in R^m, and $\mathbf{z}$ denotes an arbitrary point in R^m. The radius function of K is defined as

$$\operatorname{radius}(\mathbf{z}) = \sup_{\mathbf{x} \in K} \|\mathbf{x} - \mathbf{z}\|,$$

and

$$B_{\mathbf{z}} = \left\{ \mathbf{x} \;\middle|\; \|\mathbf{x} - \mathbf{z}\| \le \operatorname{radius}(\mathbf{z}) \right\}$$

is the smallest norm ball that is centered in $\mathbf{z}$ and contains K. The smallest norm ball that contains K can be found, therefore, by solving the problem

$$\text{minimize} \quad \tau(\mathbf{z}) = 2 \operatorname{radius}(\mathbf{z}), \tag{6.1}$$

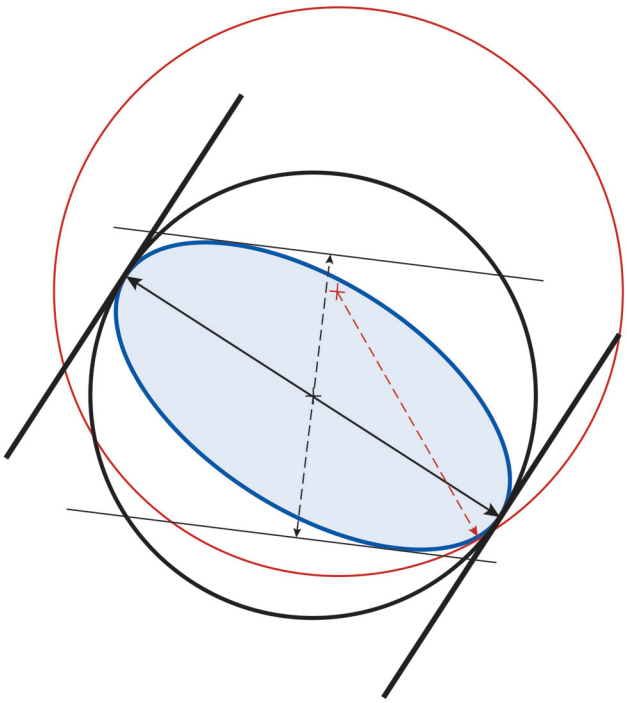


FIG. 8. *The width of the largest sandwich equals the diameter of the smallest norm ball.*

whose optimal value provides the smallest diameter of a norm ball that contains K. In this section we prove that the dual problem of (6.1) is the largest sandwich problem

$$\begin{aligned} &\text{maximize} && \sigma(\mathbf{a}) = \sup_{\mathbf{x}\in K} \mathbf{a}^T\mathbf{x} - \inf_{\mathbf{x}\in K} \mathbf{a}^T\mathbf{x} \\ &\text{subject to} && \|\mathbf{a}\|' = 1. \end{aligned} \tag{6.2}$$

Recall that $\sigma(\mathbf{a})$ equals the width of the sandwich $S(\mathbf{a}, K)$, while $\tau(\mathbf{z})$ equals the width of the sandwich $S(\mathbf{a}, B_{\mathbf{z}})$. Therefore, since K is contained in $B_{\mathbf{z}}$, the inequality

$$\sigma(\mathbf{a}) \le \tau(\mathbf{z}) \tag{6.3}$$

holds for all $\mathbf{a} \in R^m$, $\|\mathbf{a}\|' = 1$, and $\mathbf{z} \in R^m$. Thus to prove duality we need to find a pair of points, $\mathbf{a}^*$ and $\mathbf{z}^*$ say, such that $\|\mathbf{a}^*\|' = 1$ and

$$\sigma(\mathbf{a}^*) = \tau(\mathbf{z}^*). \tag{6.4}$$

For this purpose we consider the norm ball

$$B_{\mathbf{c}} = \left\{ \mathbf{x} \;\middle|\; \|\mathbf{x} - \mathbf{c}\| \le \text{radius}(\mathbf{c}) \right\}, \tag{6.5}$$

where $\mathbf{c}$ denotes the center of K. Then $B_{\mathbf{c}}$ is the smallest norm ball that is centered in $\mathbf{c}$ and contains K. Also, since K is a compact set, there exists a point $\tilde{\mathbf{y}} \in K$ such that

$$\|\tilde{\mathbf{y}} - \mathbf{c}\| = \text{radius}(\mathbf{c}).$$

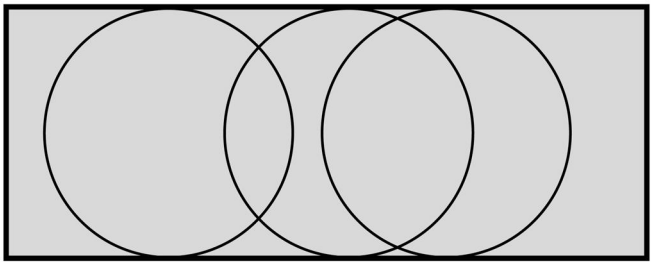


FIG. 9. *Nonuniqueness of the largest norm ball inside a convex body.*

Let $\tilde{K}$ and $\tilde{B}_{\mathbf{c}}$ denote the boundaries of K and $B_{\mathbf{c}}$, respectively. Then, clearly, $\tilde{\mathbf{y}} \in \tilde{K} \cap \tilde{B}_{\mathbf{c}}$. In other words, $\tilde{\mathbf{y}}$ is a boundary point of both K and $B_{\mathbf{c}}$. Hence there exists a vector $\tilde{\mathbf{a}} \in R^m$, $\|\tilde{\mathbf{a}}\|' = 1$, such that

$$\tilde{\mathbf{a}}^T \tilde{\mathbf{y}} = \sup_{\mathbf{x} \in B_{\mathbf{c}}} \tilde{\mathbf{a}}^T \mathbf{x} = \sup_{\mathbf{x} \in K} \tilde{\mathbf{a}}^T \mathbf{x}$$

and

$$\tilde{H} = \left\{ \mathbf{x} \;\middle|\; \tilde{\mathbf{a}}^T \mathbf{x} = \tilde{\mathbf{a}}^T \tilde{\mathbf{y}} \right\}$$

is a supporting hyperplane of both $B_{\mathbf{c}}$ and K at $\tilde{\mathbf{y}}$. Moreover, define

$$\tilde{\mathbf{u}} = \tilde{\mathbf{y}} - \mathbf{c};$$

then the vectors $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{a}}$ are aligned. Let the point

$$\hat{\mathbf{y}} = \mathbf{c} - \tilde{\mathbf{u}} = 2\mathbf{c} - \tilde{\mathbf{y}}$$

denote the "symmetrical image" of $\tilde{\mathbf{y}}$. Then, clearly, $\hat{\mathbf{y}} \in \tilde{B}_{\mathbf{c}} \cap \tilde{K}$. That is, $\hat{\mathbf{y}}$ is a boundary point of both $B_{\mathbf{c}}$ and K. A further consequence of the symmetry of K and $B_{\mathbf{c}}$ with respect to $\mathbf{c}$ is that the hyperplane

$$\hat{H} = \left\{ \mathbf{x} \;\middle|\; (-\tilde{\mathbf{a}})^T \mathbf{x} = (-\tilde{\mathbf{a}})^T \hat{\mathbf{y}} \right\}$$

supports both $B_{\mathbf{c}}$ and K at $\hat{\mathbf{y}}$. The last observation implies the relations

$$\tilde{\mathbf{a}}^T \hat{\mathbf{y}} = \inf_{\mathbf{x} \in B_{\mathbf{c}}} \tilde{\mathbf{a}}^T \mathbf{x} = \inf_{\mathbf{x} \in K} \tilde{\mathbf{a}}^T \mathbf{x},$$

$$S(\tilde{\mathbf{a}}, K) = S(\tilde{\mathbf{a}}, B_{\mathbf{c}}),$$

and

$$\sigma(\tilde{\mathbf{a}}) = \tau(\mathbf{c}),$$

which proves (6.4). The geometric interpretation of this result is quite simple: The width of the largest sandwich equals the diameter of the smallest norm ball that contains K. See Figure 8. Note that the chord vector $\tilde{\mathbf{v}} = \tilde{\mathbf{y}} - \hat{\mathbf{y}} = 2\tilde{u}$ and $\tilde{a}$ are aligned. Below we will show that any solution of (6.2) satisfies similar relations with $B_{\mathbf{c}}$. The nonuniqueness of the solutions is illustrated in Figures 9–11.

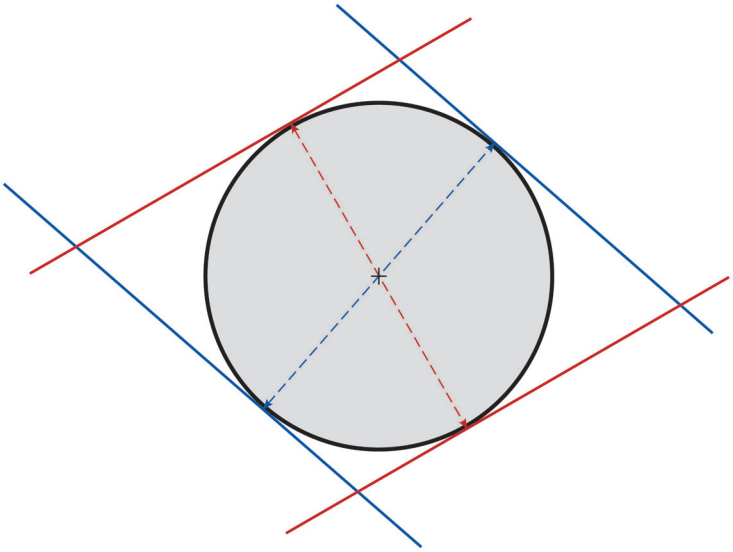


FIG. 10. *Nonuniqueness of the largest/smallest sandwich.*

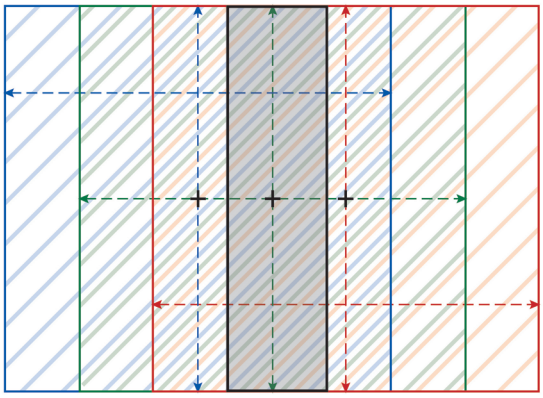


FIG. 11. *Nonuniqueness of the smallest norm ball that contains a convex body.*

Let $\mathbf{a}^* \in R^m$, $\|\mathbf{a}^*\|' = 1$, be some other solution of (6.2). Then, clearly, $S(\mathbf{a}^*, K)$ is contained in $S(\mathbf{a}^*, B_{\mathbf{c}})$. On the other hand, since $\mathbf{a}^*$ solves (6.2), $\sigma(\mathbf{a}^*) = \tau(\mathbf{c})$, so the two sandwiches have the same width and, therefore, coincide. That is,

$$S(\mathbf{a}^*, K) = S(\mathbf{a}^*, B_{\mathbf{c}}). \tag{6.6}$$

The last equality implies the existence of a common boundary point, $\mathbf{y}^* \in \tilde{K} \cap \tilde{B}_{\mathbf{c}}$, such that

$$(\mathbf{a}^*)^T \mathbf{y}^* = \sup_{\mathbf{x} \in B_{\mathbf{c}}} (\mathbf{a}^*)^T \mathbf{x} = \sup_{\mathbf{x} \in K} (\mathbf{a}^*)^T \mathbf{x}. \tag{6.7}$$

In other words, the hyperplane

$$H^* = \{\, \mathbf{x} \mid (\mathbf{a}^*)^T\mathbf{x} = (\mathbf{a}^*)^T\mathbf{y}^* \,\}$$

supports both $B_{\mathbf{c}}$ and K at $\mathbf{y}^*$. Define

$$\mathbf{u}^* = \mathbf{y}^* - \mathbf{c},$$

and let

$$\mathbf{y}^{**} = \mathbf{c} - \mathbf{u}^* = 2\mathbf{c} - \mathbf{y}^*$$

denote the symmetrical image of $\mathbf{y}^*$ with respect to $\mathbf{c}$. Then, as we have seen, the vectors $\mathbf{u}^*$ and $\mathbf{a}^*$ are aligned, $\mathbf{y}^{**} \in \tilde{K} \cap \tilde{B}_{\mathbf{c}}$, and the hyperplane

$$H^{**} = \{\, \mathbf{x} \mid (-\mathbf{a}^*)^T\mathbf{x} = (-\mathbf{a}^*)^T\mathbf{y}^{**} \,\}$$

supports both K and $B_{\mathbf{c}}$ at $\mathbf{y}^{**}$. That is,

$$(\mathbf{a}^*)^T\mathbf{y}^{**} = \inf_{\mathbf{x}\in B_{\mathbf{c}}} (\mathbf{a}^*)^T\mathbf{x} = \inf_{\mathbf{x}\in K} (\mathbf{a}^*)^T\mathbf{x}. \tag{6.8}$$

The next theorem summarizes our conclusions.

THEOREM 5. *The dual of the smallest norm ball problem* (6.1) *is the largest sandwich problem* (6.2)*, and both problems are solvable. In particular, the center of symmetry,* $\mathbf{c}$*, solves* (6.1) *and* $B_{\mathbf{c}}$ *has the smallest possible radius of all the norm balls that contain* K*. Moreover, let* $\mathbf{a}^* \in R^m$*,* $\|\mathbf{a}^*\|' = 1$*, be any solution of* (6.2)*. Then*

$$\sigma(\mathbf{a}^*) = \tau(\mathbf{c}), \tag{6.9}$$

$$S(\mathbf{a}^*, B_{\mathbf{c}}) = S(\mathbf{a}^*, K), \tag{6.10}$$

and there exists a pair of boundary points $\mathbf{y}^*$ *and* $\mathbf{y}^{**} = 2\mathbf{c} - \mathbf{y}^*$ *that belong to* $\tilde{B}_{\mathbf{c}} \cap \tilde{K}$*, but lie on opposite sides of* $S(\mathbf{a}^*, K)$*, and satisfy* (6.7)–(6.8)*. The vectors* $\mathbf{u}^* = \mathbf{y}^* - \mathbf{c}$ *and* $\mathbf{a}^*$ *are aligned. That is,*

$$(\mathbf{u}^*)^T\mathbf{a}^* = \|\mathbf{u}^*\| \cdot \|\mathbf{a}^*\|'. \tag{6.11}$$

COROLLARY 6. *The points* $\mathbf{y}^*$ *and* $\mathbf{y}^{**}$ *solve the maximal chord problem*

$$\begin{aligned} &\textit{maximize} \quad d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| \\ &\textit{subject to} \quad \mathbf{x} \in K \quad \textit{and} \quad \mathbf{y} \in K. \end{aligned} \tag{6.12}$$

Conversely, let the points $\mathbf{x}^*$ *and* $\mathbf{y}^*$ *solve* (6.12)*. Then there exists a vector* $\mathbf{a}^* \in R^m$*,* $\|\mathbf{a}^*\|' = 1$*, that solves* (6.2)*, for which the points* $\mathbf{y}^*$ *and* $\mathbf{y}^{**} = 2\mathbf{c} - \mathbf{y}^*$ *satisfy the relations mentioned in Theorem* 5*. The optimal value of* (6.12) *is called the diameter of* K*. So the diameter of* K *equals the diameter of* $B_{\mathbf{c}}$*.*

7. Nonsymmetrical convex bodies. The duality properties revealed in Theorems 3 and 5 rely on the symmetry of K. The question raised in this section is which properties hold without the symmetry assumption. Let G be an arbitrary convex body in R^m which is not symmetrical. Then in this section G replaces K. The discussion below is divided into two parts. The first one considers the largest sandwich that contains G. The second part considers the smallest sandwich that contains G.

7.1. The nonsymmetrical largest sandwich problem. We shall start by considering the links between the smallest diameter problem

$$\begin{aligned} &\text{minimize} && \tau(\mathbf{z}) = 2\,\text{radius}(\mathbf{z}) \\ &\text{subject to} && \mathbf{z} \in G, \end{aligned} \tag{7.1}$$

the largest sandwich problem

$$\begin{aligned} &\text{maximize} && \sigma(\mathbf{a}) = \sup_{\mathbf{x}\in G} \mathbf{a}^T\mathbf{x} - \inf_{\mathbf{x}\in G} \mathbf{a}^T\mathbf{x} \\ &\text{subject to} && \|\mathbf{a}\|' = 1, \end{aligned} \tag{7.2}$$

and the longest chord problem

$$\begin{aligned} &\text{maximize} && d(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\| \\ &\text{subject to} && \mathbf{x} \in G \ \text{ and } \ \mathbf{y} \in G. \end{aligned} \tag{7.3}$$

As before, for any $\mathbf{z} \in R^m$

$$\text{radius}(\mathbf{z}) = \sup_{\mathbf{x}\in G} \|\mathbf{x}-\mathbf{z}\|, \tag{7.4}$$

and

$$B_{\mathbf{z}} = \left\{ \mathbf{x} \;\middle|\; \|\mathbf{x}-\mathbf{z}\| \le \text{radius}(\mathbf{z}) \right\} \tag{7.5}$$

is the smallest norm ball that is centered in $\mathbf{z}$ and contains G. Using the compactness of G and the continuity of the objective functions it is easy to verify that the above three problems are solvable. Below we derive some useful relations between the solutions of these problems.

Recall that for any $\mathbf{a} \in R^m$, $\|\mathbf{a}\|' = 1$, and any $\mathbf{z} \in R^m$, the width of the sandwich $S(\mathbf{a}, B_{\mathbf{z}})$ equals $\tau(\mathbf{z})$. Therefore, since $S(\mathbf{a}, G)$ is contained in $S(\mathbf{a}, B_{\mathbf{z}})$,

$$\sigma(\mathbf{a}) \le \tau(\mathbf{z}). \tag{7.6}$$

However, unlike the symmetrical case, here the optimal value of (7.1) can be strictly larger than the optimal value of (7.2). Take, for example, an equilateral triangle. This type of relation is sometimes called "weak duality."

Since G is a compact set, there exists a pair of points in G, $\mathbf{y}_1$ and $\mathbf{y}_2$ say, such that

$$\inf_{\mathbf{x}\in G} \mathbf{a}^T\mathbf{x} = \mathbf{a}^T\mathbf{y}_1 \quad \text{and} \quad \sup_{\mathbf{x}\in G} \mathbf{a}^T\mathbf{x} = \mathbf{a}^T\mathbf{y}_2, \tag{7.7}$$

and the sandwich $S(\mathbf{a}, G)$ is constructed by the hyperplanes

$$H_1 = \left\{ \mathbf{x} \;\middle|\; \mathbf{a}^T\mathbf{x} = \mathbf{a}^T\mathbf{y}_1 \right\} \quad \text{and} \quad H_2 = \left\{ \mathbf{x} \;\middle|\; \mathbf{a}^T\mathbf{x} = \mathbf{a}^T\mathbf{y}_2 \right\}. \tag{7.8}$$

Note also that the width of this sandwich satisfies

$$\sigma(\mathbf{a}) = \inf \left\{ \|\mathbf{x}_1 - \mathbf{x}_2\| \;\middle|\; \mathbf{x}_1 \in H_1,\, \mathbf{x}_2 \in H_2 \right\} \le \|\mathbf{y}_1 - \mathbf{y}_2\|. \tag{7.9}$$

Let the points $\mathbf{y}_1^*$ and $\mathbf{y}_2^*$ solve the longest chord problem (7.3), and let

$$\delta = \|\mathbf{y}_2^* - \mathbf{y}_1^*\| \tag{7.10}$$

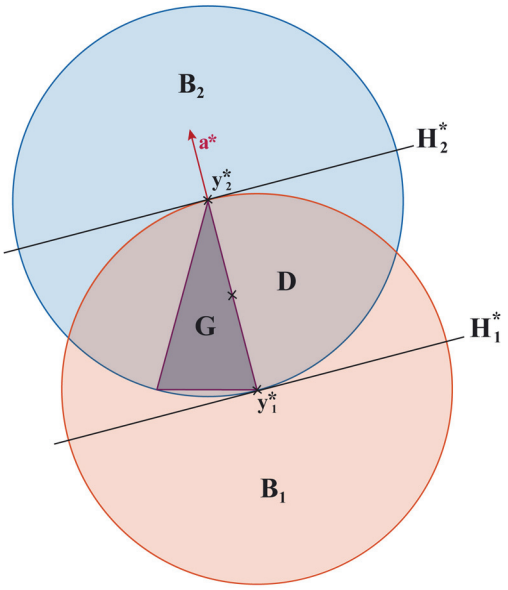


FIG. 12. *A longest chord touches a maximal sandwich.*

denote the optimal value of this problem. Then, clearly,

$$\sigma(\mathbf{a}) \le \|\mathbf{y}_1 - \mathbf{y}_2\| \le \delta.$$

Thus any vector $\mathbf{a}^* \in R^m$, $\|\mathbf{a}^*\|' = 1$, that satisfies

$$\sigma(\mathbf{a}^*) = \delta \tag{7.11}$$

solves the largest sandwich problem (7.2). Furthermore, as we are about to show, there exists a certain solution vector, $\mathbf{a}^*$, that is aligned to the optimal chord

$$\mathbf{u}^* = \mathbf{y}_2^* - \mathbf{y}_1^* \tag{7.12}$$

and satisfies (7.11).

To prove these assertions we introduce the sets

$$B_1 = \left\{ \mathbf{x} \;\middle|\; \|\mathbf{x} - \mathbf{y}_1^*\| \le \delta \right\}, \quad B_2 = \left\{ \mathbf{x} \;\middle|\; \|\mathbf{x} - \mathbf{y}_2^*\| \le \delta \right\}, \quad \text{and} \quad D = B_1 \cap B_2.$$

See Figure 12. Then D is a symmetrical convex body whose center of symmetry lies at the point $(\mathbf{y}_1^* + \mathbf{y}_2^*)/2$. Note also that D contains G, and that $\mathbf{y}_2^*$ is a boundary point of G, D, and B_1. Consequently there exists a vector $\mathbf{a}^* \in R^m$, $\|\mathbf{a}^*\|' = 1$, such that

$$(\mathbf{a}^*)^T\mathbf{y}_2^* = \sup_{\mathbf{x}\in B_1} (\mathbf{a}^*)^T\mathbf{x} = \sup_{\mathbf{x}\in D} (\mathbf{a}^*)^T\mathbf{x} = \sup_{\mathbf{x}\in G} (\mathbf{a}^*)^T\mathbf{x}, \tag{7.13}$$

and

$$H_2^* = \left\{ \mathbf{x} \;\middle|\; (\mathbf{a}^*)^T\mathbf{x} = (\mathbf{a}^*)^T\mathbf{y}_2^* \right\}$$

is a supporting hyperplane of the sets B_1, D, and G, at the point $\mathbf{y}_2^*$. Observe also that the vectors $\mathbf{u}^*$ and $\mathbf{a}^*$ are aligned:

$$(\mathbf{u}^*)^T(\mathbf{a}^*) = \|\mathbf{u}^*\| \cdot \|\mathbf{a}^*\|' = \|\mathbf{u}^*\|. \tag{7.14}$$

Similarly $\mathbf{y}_1^*$ is a boundary point of the sets B_2, D, and G. Hence the symmetry of D and the fact that $\mathbf{y}_1^*$ is the symmetrical image of $\mathbf{y}_2^*$ imply that

$$H_1^* = \left\{ \mathbf{x} \;\middle|\; (-\mathbf{a}^*)^T\mathbf{x} = (-\mathbf{a}^*)^T\mathbf{y}_1^* \right\}$$

is a supporting hyperplane of the sets B_2, D, and G, at the point $\mathbf{y}_1^*$. That is,

$$(-\mathbf{a}^*)^T\mathbf{y}_1^* = \sup_{\mathbf{x}\in B_2} (-\mathbf{a}^*)^T\mathbf{x} = \sup_{\mathbf{x}\in D} (-\mathbf{a}^*)^T\mathbf{x} = \sup_{\mathbf{x}\in G} (-\mathbf{a}^*)^T\mathbf{x} \tag{7.15}$$

and

$$(\mathbf{a}^*)^T\mathbf{y}_1^* = \inf_{\mathbf{x}\in G} (\mathbf{a}^*)^T\mathbf{x}. \tag{7.16}$$

Combining (7.12)–(7.16) gives

$$\begin{aligned}\sigma(\mathbf{a}^*) &= \sup_{\mathbf{x}\in G} (\mathbf{a}^*)^T\mathbf{x} - \inf_{\mathbf{x}\in G} (\mathbf{a}^*)^T\mathbf{x} = (\mathbf{a}^*)^T\mathbf{y}_2^* - (\mathbf{a}^*)^T\mathbf{y}_1 \\ &= (\mathbf{a}^*)^T(\mathbf{y}_2^* - \mathbf{y}_1^*) = (\mathbf{a}^*)^T\mathbf{u}^* = \|\mathbf{u}^*\|,\end{aligned}$$

which proves (7.11). The next theorem summarizes our findings. See Figure 12.

THEOREM 7. *The dual of the smallest norm ball problem* (7.1) *is the largest sandwich problem* (7.2)*, and both problems are solvable. Yet, unlike the symmetrical case, here the optimal primal value may exceed the optimal dual value. The longest chord problem* (7.3) *is also solvable and its optimal value equals that of* (7.2)*. In other words, the "diameter" of* G *equals the "length" of* G*. More precisely, let the points* $\mathbf{y}_1^*$ *and* $\mathbf{y}_2^*$ *solve* (7.3)*; then there exists a vector* $\mathbf{a}^* \in R^m$*,* $\|\mathbf{a}^*\|' = 1$*, that solves* (7.2) *and satisfies* (7.11)–(7.16)*. Thus, in particular, the vectors* $\mathbf{u}^* = \mathbf{y}_2^* - \mathbf{y}_1^*$ *and* $\mathbf{a}^*$ *are aligned.*

Theorem 7 raises the question of whether the converse claim is also true. Let $\hat{\mathbf{a}} \in R^m$, $\|\hat{\mathbf{a}}\|' = 1$, be some other solution of (7.2). Then, as we have seen, there exists a pair of boundary points of G, $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$ say, such that

$$\hat{\mathbf{a}}^T\hat{\mathbf{y}}_1 = \inf_{\mathbf{x}\in G} \hat{\mathbf{a}}^T\mathbf{x} \quad \text{and} \quad \hat{\mathbf{a}}^T\hat{\mathbf{y}}_2 = \sup_{\mathbf{x}\in G} \hat{\mathbf{a}}^T\mathbf{x}. \tag{7.17}$$

Hence the sandwich $S(\hat{\mathbf{a}}, G)$ is constructed by the hyperplanes

$$\hat{H}_1 = \left\{ \mathbf{x} \;\middle|\; \hat{\mathbf{a}}^T\mathbf{x} = \hat{\mathbf{a}}^T\hat{\mathbf{y}}_1 \right\} \quad \text{and} \quad \hat{H}_2 = \left\{ \mathbf{x} \;\middle|\; \hat{\mathbf{a}}^T\mathbf{x} = \hat{\mathbf{a}}^T\hat{\mathbf{y}}_2 \right\}, \tag{7.18}$$

and the width of this sandwich satisfies

$$\sigma(\hat{\mathbf{a}}) = \inf \left\{ \|\mathbf{x}_1 - \mathbf{x}_2\| \;\middle|\; \mathbf{x}_1 \in \hat{H}_1,\, \mathbf{x}_2 \in \hat{H}_2 \right\} \le \|\hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2\| \le \delta. \tag{7.19}$$

See Figure 13. On the other hand, since $\hat{\mathbf{a}}$ solves (7.2), $\sigma(\hat{\mathbf{a}}) = \delta$. Hence the inequalities in (7.19) imply that

$$\|\hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2\| = \delta. \tag{7.20}$$

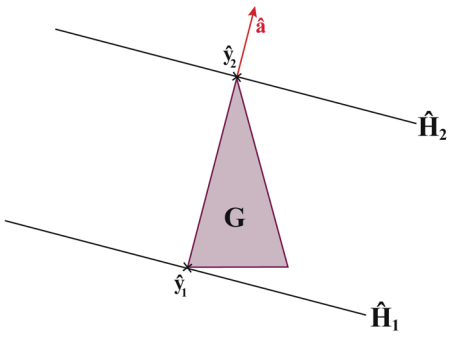


FIG. 13. *A maximal sandwich touches a longest chord.*

In other words, the contact points $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$ solve (7.3). Moreover, since the hyperplane $\hat{H}_2$ supports the norm ball $\hat{B}_1 = \{\mathbf{x} \mid \|\mathbf{x} - \hat{\mathbf{y}}_1\| \le \delta\}$ at $\hat{\mathbf{y}}_2$, the vectors $\hat{\mathbf{u}} = \hat{\mathbf{y}}_2 - \hat{\mathbf{y}}_1$ and $\hat{\mathbf{a}}$ are aligned:

$$\hat{\mathbf{u}}^T \hat{\mathbf{a}} = \|\hat{\mathbf{u}}\| \cdot \|\hat{\mathbf{a}}\|' = \|\hat{\mathbf{u}}\|. \tag{7.21}$$

This proves the following results. See Figure 13.

THEOREM 8. *Let the vector* $\hat{\mathbf{a}} \in R^m$, $\|\hat{\mathbf{a}}\|' = 1$, *solve* (7.2). *Then there exists a pair of points,* $\hat{\mathbf{y}}_1$ *and* $\hat{\mathbf{y}}_2$ *say, that satisfy* (7.17)–(7.21) *and solve the longest chord problem* (7.3). *Moreover, the vectors* $\hat{\mathbf{u}} = \hat{\mathbf{y}}_2 - \hat{\mathbf{y}}_1$ *and* $\hat{\mathbf{a}}$ *are aligned.*

It is also interesting to compare (7.1) with the unconstrained problem

$$\text{minimize} \quad \tau(\mathbf{z}) = 2\,\text{radius}(\mathbf{z}), \tag{7.22}$$

whose solution points may reside outside G. The last feature is easily seen by considering the smallest ℓ_∞ norm ball. See Figure 11. On the other hand, when using the Euclidean norm any solution of (7.22) lies in G. These observations raise the question of whether any solution of (7.1) also solves (7.22). Yet the answer is left beyond the scope of this paper.

7.2. The nonsymmetrical smallest sandwich problem. The smallest sandwich that contains G is obtained by solving the problem

$$\begin{aligned} &\text{minimize} && \sigma(\mathbf{a}) = \sup_{\mathbf{x}\in G} \mathbf{a}^T\mathbf{x} - \inf_{\mathbf{x}\in G} \mathbf{a}^T\mathbf{x} \\ &\text{subject to} && \|\mathbf{a}\|' = 1. \end{aligned} \tag{7.23}$$

The compactness of the sets G and $\{\mathbf{a} \mid \|\mathbf{a}\|' = 1\}$ and the continuity of the objective function ensure the existence of a vector $\mathbf{a}^* \in R^m$, $\|\mathbf{a}^*\|' = 1$, that solves (7.23). However, as Figure 10 shows, the solution is not necessarily unique. Let $\tilde{G}$ and $\mathring{G}$ denote the boundary and the interior of G, respectively. The discussion in section 5 suggests that the dual of (7.23) is the maximal norm ball problem

$$\begin{aligned} &\text{maximize} && \eta(\mathbf{z}) = 2\,\text{depth}(\mathbf{z}) \\ &\text{subject to} && \mathbf{z} \in G, \end{aligned} \tag{7.24}$$

where

$$\text{depth}(\mathbf{z}) = \inf_{\mathbf{x}\in\tilde{G}} \|\mathbf{x} - \mathbf{z}\|, \tag{7.25}$$

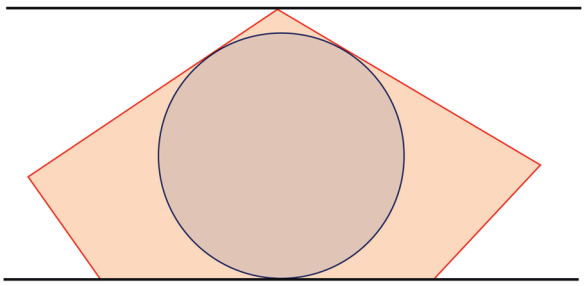


FIG. 14. *The touching conjecture: A maximal norm ball touches a minimal sandwich and vice versa.*

and

$$B_{\mathbf{z}} = \{\, \mathbf{x} \mid \|\mathbf{x} - \mathbf{z}\| \leq \text{depth}(\mathbf{z}) \,\} \tag{7.26}$$

is the largest norm ball that is centered in $\mathbf{z}$ and contained in G. As before, the compactness of G and the continuity of the objective function imply the existence of an interior point, $\mathbf{z}^*$ say, that solves (7.24). Yet, as Figure 9 shows, there may be many solutions.

Recall that for any pair of vectors, $\mathbf{z} \in G$ and $\mathbf{a} \in R^m$, $\|\mathbf{a}\|' = 1$, the width of the sandwich $S(\mathbf{a}, B_{\mathbf{z}})$ equals $\eta(\mathbf{z})$. On the other hand, since $B_{\mathbf{z}}$ is contained in G, the width of $S(\mathbf{a}, B_{\mathbf{z}})$ is smaller than the width of $S(\mathbf{a}, G)$. That is,

$$\eta(\mathbf{z}) \leq \sigma(\mathbf{a}). \tag{7.27}$$

The last inequality invites the question of whether the optimal values of (7.23) and (7.24) are always equal. However, there are many examples of convex sets for which the solution points satisfy a strict inequality of the form

$$\eta(\mathbf{z}^*) < \sigma(\mathbf{a}^*). \tag{7.28}$$

Take, for example, a triangle. Thus again we see that the lack of symmetry may result in "weak duality" relations.

It is also easy to verify that any interior point $\mathbf{z}^* \in \mathring{G}$ that solves the maximal norm ball problem (7.24) has the following property: The maximal norm ball

$$B_{\mathbf{z}^*} = \{\, \mathbf{x} \mid \|\mathbf{x} - \mathbf{z}^*\| \leq \text{depth}(\mathbf{z}^*) \,\} \tag{7.29}$$

"touches" $\widetilde{G}$ in more that one point. Let $\tilde{\mathbf{y}}$ be one of the "touching points." Then $\tilde{\mathbf{y}}$ is a boundary point of both $B_{\mathbf{z}^*}$ and G. Moreover, as we have seen, there exists a vector $\tilde{\mathbf{a}} \in R^m$, $\|\tilde{\mathbf{a}}\|' = 1$, such that

$$\tilde{H} = \{\, \mathbf{x} \mid \tilde{\mathbf{a}}^T\mathbf{x} = \tilde{\mathbf{a}}^T\tilde{\mathbf{y}} \,\}$$

is a supporting hyperplane of both G and $B_{\mathbf{z}^*}$ at the point $\tilde{\mathbf{y}}$, and the vectors $\tilde{\mathbf{y}} - \mathbf{z}^*$ and $\tilde{\mathbf{a}}$ are aligned. In other words, any "touching point," $\tilde{\mathbf{y}}$, is related to a certain "touching sandwich," $S(\tilde{\mathbf{a}}, G)$. There are many examples of convex sets in R^2 and R^3 in which one of the "touching sandwiches," $S(\tilde{\mathbf{a}}, G)$, has the smallest possible width. See Figure 14. These empirical observations bring us to make the following "touching conjecture": A maximal norm ball (inside G) always touches a minimal sandwich. Conversely, any minimal sandwich of G touches a maximal norm ball. However, the validity of these assertions remains an open question.

8. Concluding remarks. Recall that any unit norm ball of the form (2.2) defines an "equilibrated" convex body. That is, a symmetrical convex body with center at the origin. The converse is also true: Any equilibrated convex body defines a norm; see [16]. Thus any symmetrical convex body is essentially a "shifted" unit norm ball of some norm. Hence, from this point of view, Theorems 3 and 5 characterize the duality relations between two arbitrary norms on R^m.

The new MND theorems, together with the old ones, constitute an elegant collection of geometric problems that visually illustrate the basic principles of duality. See Figures 1–8. Being MND theorems, the new theorems have some analogy with the old ones. However, the geometry of the new problems is quite different, and there is no direct way to conclude the new theorems from the old ones. So the new theorems contribute a substantial extension to the range of problems that can be handled via the MND methodology.

The appeal of the MND methodology stems from a number of reasons. First, as noted above, it has a simple geometric interpretation that visually illustrates the links between the primal problem and the dual one. Second, it allows the freedom to use any norm in R^m. Third, the alignment relations add important insight into the nature of primal-dual problems. In particular, when using a smooth strictly convex norm, there are explicit rules for retrieving a primal solution from a dual one, and vice versa, e.g., [8], [10], [19]. Fourth, in spite of their simplicity, the MND theorems apply to a large family of problems. See [8] for a recent survey of such problems.

A recent discussion in [9] considers methods for solving the related duality problems, regarding two types of polyhedral convex sets. The first one is a polyhedron, which is defined as the intersection of n given half-spaces in R^m. The second set is a polytope, which is defined as the convex hull of ℓ given points in R^m. Then, as shown in [9], the nearest hyperplane problem is easy to solve on a polyhedron and difficult to solve on a polytope. Yet in the farthest hyperplane problem the situation is reversed. This problem is easy to solve on a polytope and difficult to solve on a polyhedron. A similar situation characterizes the solutions of the smallest sandwich problem and the largest sandwich problem. The need for calculating a norm ball, usually an ellipsoid, that lies inside (or contains) a given polyhedral convex set arises in several applications; see, e.g., [2], [4].

The new theorems are valid in any finite dimensional real Hilbert space H, with inner product $\langle \mathbf{x}, \mathbf{y} \rangle$. The modification of the current proof to handle this setting is rather simple: R^m is replaced by H, the Euclidean inner product $\mathbf{x}^T\mathbf{y}$ is replaced by $\langle \mathbf{x}, \mathbf{y} \rangle$, the Euclidean norm is replaced by $\|\mathbf{x}\| = (\langle \mathbf{x}, \mathbf{x} \rangle)^{1/2}$, and so forth. Using the geometric Hahn–Banach theorem one can prove the Nirenberg–Luenberger MND theorem in any real (or complex) normed linear space; see, e.g., [11], [12], [18], [19], [21]. This observation suggests that the new MND theorems remain valid in such setting. However, we have good reasons to keep the paper in the R^m setting. First, the restriction to R^m simplifies the presentation and helps to focus on the main ideas. Second, there is vast literature on properties of convex bodies, including closely related issues such as "width," "length," and "diameter" of convex bodies, and most of the discussions are carried out in R^m. See, for example, [1], [2], [4], [6], [13], [14], [17], [20], [22], [24], [26], [27], [28], [29], and the references therein.

REFERENCES

[1] R. V. Benson, *Euclidean Geometry and Convexity*, McGraw-Hill, New York, 1966.

[2] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*, SIAM, Philadelphia, 2001.

[3] J. M. Borwein and A. S. Lewis, *Convex Analysis and Nonlinear Optimization. Theory and Examples*, Springer-Verlag, New York, 2000.

[4] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.

[5] W. Briec, *Minimum distance to the complement of a convex set: A duality result*, J. Optim. Theory Appl., 93 (1997), pp. 301–319.

[6] G. D. Chakerian and H. Groemer, *Convex bodies of constant width*, in Convexity and Its Applications, P. M. Gruber and J. M. Wills, eds., Birkhäuser, Basel, 1983, pp. 49–96.

[7] A. Dax, *On minimum norm solutions*, J. Optim. Theory Appl., 76 (1993), pp. 183–193.

[8] A. Dax, *The distance between two convex sets*, Linear Algebra Appl., 416 (2006), pp. 184–213.

[9] A. Dax, *Dual Sandwich Theorems: Examples of Applications*, Technical report, Hydrological Service of Israel, Jerusalem, Israel, 2008.

[10] A. Dax and V. P. Sreedharan, *Theorems of the alternative and duality*, J. Optim. Theory Appl., 94 (1997), pp. 561–590.

[11] F. Deutsch, *Best Approximation in Inner Product Spaces*, Springer-Verlag, New York, 2001.

[12] F. R. Deutsch and P. H. Maserick, *Applications of the Hahn-Banach theorem in approximation theory*, SIAM Rev., 9 (1967), pp. 516–530.

[13] P. Gritzmann and V. Klee, *Mathematical programming and convex geometry*, in Handbook of Convex Geometry, Vol. A, P. M. Gruber and J. M. Wills, eds., North–Holland, Amsterdam, 1993, pp. 627–674.

[14] P. M. Gruber, *Approximation of convex bodies*, in Convexity and Its Applications, P. M. Gruber and J. M. Wills, eds., Birkhäuser, Basel, 1983, pp. 131–162.

[15] J.-B. Hiriart-Urruty and C. Lemarechal, *Convex Analysis and Minimization Algorithms* I. *Fundamentals*, Springer-Verlag, Berlin, 1993.

[16] A. S. Householder, *The Theory of Matrices in Numerical Analysis*, Blaisdell, New York, 1964.

[17] P. J. Kelly and M. L. Weiss, *Geometry and Convexity: A Study in Mathematical Methods*, John Wiley and Sons, New York, 1979.

[18] P. D. Lax, *Functional Analysis*, John Wiley and Sons, New York, 2002.

[19] D. G. Luenberger, *Optimization by Vector Space Methods*, John Wiley and Sons, New York, 1969.

[20] O. L. Mangasarian, *Arbitrary-norm separating plane*, Oper. Res. Lett., 24 (1999), pp. 15–23.

[21] L. Nirenberg, *Functional Analysis*, lectures given in 1960–61, notes by L. Sibner, New York University, New York, 1961.

[22] M. J. Panik, *Fundamentals of Convex Analysis*, Kluwer Academic Publishers, Dordrecht, 1993.

[23] A. W. Roberts and D. E. Varberg, *Convex Functions*, Academic Press, New York, 1973.

[24] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[25] I. Singer, *Duality for Nonconvex Approximation and Optimization*, Springer-Verlag, New York, 2006.

[26] J. van Tiel, *Convex Analysis. An Introductory Text*, John Wiley and Sons, New York, 1984.

[27] H. J. H. Tuenter, *Minimum L_1-distance projection onto the boundary of a convex set: Simple characterization*, J. Optim. Theory Appl., 112 (2002), pp. 441–445.

[28] R. Webster, *Convexity*, Oxford University Press, New York, 1994.

[29] I. M. Yaglom and V. G. Boltyanskii, *Convex Figures*, Holt, Rinehart and Winston, New York, 1961.

SIAM J. OPTIM.
Vol. 19, No. 4, pp. 1970–1994

PARETO SUBDIFFERENTIAL CALCULUS FOR CONVEX VECTOR MAPPINGS AND APPLICATIONS TO VECTOR OPTIMIZATION*

MOUNIR EL MAGHRI† AND MOHAMED LAGHDIR†

Abstract. This paper deals with the subdifferential of convex analysis defined in the Pareto sense, from a point of view of nonvacuity, characterizations, and calculus rules and their applications to the vector optimization, the convex maps being vector-valued in a finite- or infinite-dimensional ordered vector space. Subdifferentiability is characterized under conditions of Attouch–Brézis type. Formulations by derivatives, when they exist, are provided. Concerning the calculus rules, the first main result gives the gap between the Pareto subdifferential and the ordinary one, allowing thus the computation of the one from the other. Next, as central results, Pareto subdifferentials of the sum and/or composition of two convex vector mappings are developed. The formulas are obtained under Moreau–Rockafellar or Attouch–Brézis-type conditions, revealing, strangely, the presence of the ordinary subdifferential. These formulas actually allow the extension of the indicator function technique to the vector case, so that Pareto optimality (efficiency) conditions are easily derived and weakened with qualification conditions of the Attouch–Brézis kind. Finally, the gap between efficient and optimal sets is also deduced.

Key words. vector mappings, convex analysis, subdifferential, vector optimization, efficiency

AMS subject classifications. 49K27, 90C25, 90C29

DOI. 10.1137/070704046

1. Introduction.

The notion of the subdifferential defined in the Pareto sense is important for dealing with vector optimization problems (VOPs). There are globally two kinds of solutions for these problems: usual optimum and Pareto optimum, also called efficient point. It is well known that the ordinary optimum for vector problems does not often exist, so it is said to be a rather strongly efficient solution in Pareto language. Thus, for vector mappings, the subdifferential defined in the Pareto sense is naturally more required in the context of vector optimization than a subdifferential defined in the usual sense. We agree to call them, respectively, Pareto subdifferential and strong subdifferential. To be sometimes more precise, since among the Pareto concepts there is the proper or weak notion, we also talk about proper or weak subdifferential of vector mappings.

In this paper, we study several properties of the Pareto subdifferential of convex analysis that have not been considered previously in the literature and give their first direct applications to the constrained convex vector optimization. Only the proper and weak concepts of Pareto are considered; the other concepts remain open issues for future research. For a reason of enlarging the scope of this work, we assume the convex vector mappings taking values in (partially) ordered topological vector spaces. All the obtained results are important also in the finite-dimensional case, and the reader which is not familiar with the infinite-dimensional spaces may consider everywhere all the spaces as real finite-dimensional vector spaces with the range spaces all ordered by the natural componentwise partial order. We nevertheless didn't fail to state in this setting, whenever it is necessary, each result or comment being able to be formulated in simpler terms.

*Received by the editors September 28, 2007; accepted for publication (in revised form) October 12, 2008; published electronically March 13, 2009.
http://www.siam.org/journals/siopt/19-4/70404.html
†Math and Computer Department, Faculty of Sciences, Chouaïb Doukkali University, BP. 20, El Jadida, Morocco (elmaghri@yahoo.com, laghdirm@yahoo.fr).

The paper is organized as follows. Section 2 describes the different notions of vector subdifferential underlining their preliminary connections with the efficient sets. Section 3 is divided into three subsections. Subsection 3.1 presents the scalarization of these sets straightforwardly from the well-known one of the efficient sets. Scalarization principle is revealed in our investigations to be fundamental in the obtention of the most essential properties of these vector subdifferentials. In subsection 3.2, the Pareto subdifferentiability and the regular subdifferentiability are characterized under the well-known Attouch–Brézis and Moreau–Rockafellar conditions, respectively; the regular subdifferentiability concept due to Raffin [23] being revealed to be a key hypothesis for the obtention of our main results. As immediate result, we obtain the Pareto subdifferential in terms of the strong one. This relationship expresses, in fact, the gap between the two concepts allowing their mutual computation. Subsection 3.3 concerns the formulations of these sets by means of (directional or Gâteaux) derivatives when they exist. Section 4 represents the central part in this work for both theory and application. It is devoted to calculus rules of the Pareto subdifferentials for the sum and/or composition of two convex vector mappings. The obtained formulas are original and hold under the weak conditions of the Moreau–Rockafellar or Attouch–Brézis types. Let us point out the strange but indispensable appearance of the strong subdifferential, otherwise the formulas would not hold in general. This presence turns out to be rather favorable in applications: In the last section dealing with the constrained vector optimization, the extension of the penalty indicator function to the vector case is realized, in the sense that both necessary and sufficient efficiency conditions of the Kuhn–Tucker type are easily derived under only the weak qualification conditions of the Attouch–Brézis type; also, a certain gap between the efficient and the optimal sets is deduced.

The first studies of the vector subdifferential seem to appear in strong form due, respectively, to Raffin [23], Valadier [32], and Zowe [35], while the Pareto subdifferential was first considered by Sawaragi and Tanino [25] and next appeared in [26]. Concerning the chain rules, the first results concerned the sum operation due to Théra [29] for the strong subdifferential, Lin [18] and Taa [27] for the weak one. Their main results concerning us are described in detail as soon as it is required in the text.

2. Elements of vector convex analysis. Throughout the paper, we shall work with the following spaces and sets:

- X, Y, Z (real) topological vector spaces, with Y and Z separated;
- X^*, Y^*, Z^* their respective topological duals paired in duality by $\langle ., . \rangle$;
- Y_+ a nonempty proper[1] convex cone (resp. Z_+) of Y (resp. of Z), $\operatorname{int} Y_+$ its topological interior (resp. $\operatorname{int} Z_+$) sometimes will be required to be nonempty;
- $l(Y_+) = Y_+ \sqcap -Y_+$ (resp. $l(Z_+)$) the lineality of Y_+ (resp. Z_+), when it is null, the cone is said to be pointed;
- Y_+^* (resp. Z_+^*) the polar cone of Y_+ (resp. Z_+), which is the set of nonnegative forms $\lambda \in Y^*$, i.e., $\lambda(Y_+) \subseteq \mathbb{R}_+$;
- $(Y_+^*)^\circ$ (resp. $(Z_+^*)^\circ$) the strict polar cone of Y_+ (resp. Z_+), which is the set of positive forms $\lambda \in Y^*$, i.e., $\lambda(Y_+ \setminus l(Y_+)) \subseteq \mathbb{R}_+ \setminus \{0\}$, obviously, $(Y_+^*)^\circ \subseteq Y_+^* \setminus \{0\}$ since $Y_+ + Y_+ \setminus l(Y_+) \subseteq Y_+ \setminus l(Y_+)$;
- $L(X, Y)$ the space of linear continuous operators from X to Y;
- $L_+(Y, Z)$ the set of nonnegative operators $A \in L(Y, Z)$, i.e., $A(Y_+) \subseteq Z_+$;

[1] That is, not a linear subspace so that it cannot coincide with its lineality.

- $L^{-1}(X,Y)$ (resp. $L_+^{-1}(Y,Z)$) the set of invertible operators $A \in L(X,Y)$ (resp. $\in L_+(Y,Z)$).

The convex cone Y_+ (resp. Z_+) induces in Y (resp. in Z) preorder relations:

$$\begin{array}{lllll} y \leq_{Y_+} y' & \Longleftrightarrow & y' \geq_{Y_+} y & \Longleftrightarrow & y' - y \in Y_+, \\ y <_{Y_+} y' & \Longleftrightarrow & y' >_{Y_+} y & \Longleftrightarrow & y' - y \in \operatorname{int} Y_+, \\ y \lneq_{Y_+} y' & \Longleftrightarrow & y' \gneq_{Y_+} y & \Longleftrightarrow & y' - y \in Y_+ \setminus l(Y_+). \end{array}$$

We adjoin to Y (resp. Z) an abstract maximal element denoted $+\infty$ so that $y \leq_{Y_+} +\infty$ for all $y \in Y \sqcup \{+\infty\}$ on which we consider the operations $y+(+\infty) = +\infty$ and $\alpha \cdot (+\infty) = +\infty$ for all $\alpha \in \mathbb{R}_+ \setminus \{0\}$ by adopting the convention $0 \cdot (+\infty) = 0$.

Let us notice the following well-known properties on some fundamental relations between a cone and its polar.

PROPOSITION 2.1.

(1) *If Y is locally convex and Y_+ is closed, then*
$\exists y \in Y,\ \forall \lambda \in Y_+^* \setminus \{0\},\ \langle \lambda, y\rangle \geq 0 \Longrightarrow y \in Y_+$.

(2) *If, in addition, Y_+ is pointed, then*
$\exists y \in Y,\ \forall \lambda \in Y_+^* \setminus \{0\},\ \langle \lambda, y\rangle = 0 \Longrightarrow y = 0$.

(3) *If $\operatorname{int} Y_+ \neq \emptyset$, then*
$\forall \lambda \in Y_+^* \setminus \{0\},\ \forall y \in \operatorname{int} Y_+,\ \langle \lambda, y\rangle > 0$.

(4) *In particular,*
$\forall \lambda \in Y_+^* \setminus \{0\},\ \exists y_\lambda \in \operatorname{int} Y_+ : \langle \lambda, y_\lambda\rangle = 1$.

We recall that property (1) is obtained via contradiction by strict separation theorem (see [2, pp. 122] or [5, pp. 11–12], e.g.) knowing that Y_+ is a nonempty closed convex set.[2] Assertion (2) is a direct consequence of assertion (1) using that Y_+ is pointed. Property (3) is an application of the well-known property that a linear form is null if only if it is null on an open set.[3] Assertion (4) is obviously deduced from (3).

Also, we shall work with the following vector mappings:

- $F : X \to Y \sqcup \{+\infty\}$ is said to be
  - Y_+-convex, if

$$\forall x, x' \in X,\ \ \forall \alpha \in [0,1],\ \ F(\alpha x + (1-\alpha)x') \leq_{Y_+} \alpha F(x) + (1-\alpha)F(x');$$

  - sequentially Y_+-l.s.c at $\bar{x} \in X$, if

$$\forall y \leq_{Y_+} F(\bar{x}),\ \ \forall (x^n) \to \bar{x},\ \ \exists (y^n) \to y :\ \ y^n \leq_{Y_+} F(x^n),\ \ \forall n \in \mathbb{N};$$

  - proper, if its effective domain

$$\operatorname{dom} F = \{x \in X :\ F(x) \in Y\} \neq \emptyset.$$

- $G : Y \to Z \sqcup \{+\infty\}$ is said to be (Y_+, Z_+)-nondecreasing, if

$$y \leq_{Y_+} y' \quad \Longrightarrow \quad G(y) \leq_{Z_+} G(y').$$

[2] If $y \notin Y_+$, we can strictly separate it from Y_+, i.e., $\exists \lambda \in Y^* \setminus \{0\}$, $\exists \alpha > 0 : \langle \lambda, y\rangle < \alpha < \langle \lambda, y'\rangle$ $(\forall y' \in Y_+)$. Taking successively $y' = 0$ and $y' = ny''$, with $n \in \mathbb{N}^*$ and $y'' \in Y_+$, we obtain $\langle \lambda, y\rangle < 0$ and $\langle \lambda, y''\rangle > \frac{\alpha}{n}$. Letting $n \nearrow +\infty$, we get $\lambda \in Y_+^* \setminus \{0\}$ in contradiction with $\langle \lambda, y\rangle < 0$.

[3] Let $\lambda \in Y^*$ such that $\lambda(O) = \{0\}$ for some open set $O \subset Y$. Let $y_0 \in O$, then there exists V a neighborhood of 0 such that $y_0 + V \subset O$. For each $y \in Y$, the continuity at 0 of the mapping $\alpha \mapsto \alpha y$ implies the existence of $\alpha \in \mathbb{R}$, with $\alpha y \in V$. But $\langle \lambda, y\rangle = \frac{1}{\alpha}\langle \lambda, y_0 + \alpha y\rangle - \frac{1}{\alpha}\langle \lambda, y_0\rangle = 0$.

▶ $G \circ F : X \to Z \sqcup \{+\infty\}$ the composite vector mapping is defined by

$$G \circ F(x) = \begin{cases} G(F(x)) & \text{if } x \in \operatorname{dom} F, \\ +\infty & \text{else.} \end{cases}$$

It is immediate that

- $\operatorname{dom}(G \circ F) = F^{-1}(\operatorname{dom} G) \sqcap \operatorname{dom} F$.
- F is Y_+-convex $\Longrightarrow \operatorname{dom} F$ is a convex set.
- G is (Y_+, Z_+)-nondecreasing Z_+-convex and F is Y_+-convex $\Longrightarrow G \circ F$ is Z_+-convex.

We agree to denote the sets of such mappings in analogy with the scalar case:

▶ $\Gamma(X, Y)$ the set of proper Y_+-convex mappings from X to $Y \sqcup \{+\infty\}$,

▶ $\Gamma_0(X, Y)$ the set of sequentially Y_+-l.s.c mappings $F \in \Gamma(X, Y)$,

so that $\Gamma(X, \mathbb{R})$ reduces to $\Gamma(X)$, the set of proper convex functionals. In the same way, a functional $(Y_+, \mathbb{R}_+)$-nondecreasing is usually said to be Y_+-nondecreasing. While $\Gamma_0(X, \mathbb{R})$ reduces to $\Gamma_0(X)$, the set of l.s.c functionals in $\Gamma(X)$ if X is metrizable (in particular, a Fréchet space, i.e., complete metrizable space). Indeed, the vector lower semicontinuity concept due to Combari, Laghdir, and Thibault [11] is not other in the setting of metrizable variable and value spaces than a sequential characterization of the concept introduced by Penot and Théra [21]. Hence, sequential $\mathbb{R}_+$-l.s.c is no more than the classical l.s.c if X is metrizable. Sequential $\mathbb{R}^r_+$-l.s.c of $F = (f_1, \ldots, f_r)$, then, by definition, also becomes equivalent to l.s.c of each component f_i in such a space X. In general, the sequential cone l.s.c concept easily implies that each y-sublevel set ($y \in Y$) and the epigraph of F, respectively, defined by

$$Y_+\text{-Lev}(F; y) = \{x \in X : \ F(x) \leq_{Y_+} y\},$$
$$Y_+\text{-Epi } F = \{(x, y) \in X \times Y : \ F(x) \leq_{Y_+} y\}$$

are closed if Y_+ is closed; the converse being false, in general, even for finite-dimensional image spaces equipped with closed order cones (see example in [11] or [21]). It has been noted in [11] that sequential continuity (at a point in the effective domain) is equivalent to sequential Y_+-l.s.c and $-Y_+$-l.s.c (at this point) if the preorder Y_+ is normal, i.e., if there exists a base of neighborhoods V of the origin such that the order intervals

$$[a, b] = \{y \in Y : \ a \leq_{Y_+} y \leq_{Y_+} b\} \subset V \qquad \text{if} \quad a, b \in V.$$

Consider now the VOP associated with the vector map $F : X \supseteq S \to Y \sqcup \{+\infty\}$:

$$\text{VOP:} \qquad \operatorname*{Min}_{x \in S} F(x).$$

There are four kinds of solutions for VOP: a point $\bar{x} \in S \sqcap \operatorname{dom} F$ is said to be

- ▶ optimal or strongly efficient, if $\forall x \in S, \ F(\bar{x}) \leq_{Y_+} F(x)$;
- ▶ Pareto or efficient, if $\nexists x \in S, \ F(x) \lesssim_{Y_+} F(\bar{x})$;
- ▶ weak Pareto or weakly efficient, if $\nexists x \in S, \ F(x) <_{Y_+} F(\bar{x})$;
- ▶ proper Pareto or (Henig) properly efficient, if

$\exists \hat{Y}_+ \subsetneq Y$ a convex cone, with $Y_+ \setminus l(Y_+) \subseteq \operatorname{int} \hat{Y}_+, \ \nexists x \in S, \ F(x) \lesssim_{\hat{Y}_+} F(\bar{x})$.

The weak concept obviously supposes in the image space a nonempty interior preorder cone, while the proper concept supposes generally in counterpart a pointed

preorder cone. The efficient set and the strongly, properly, and weakly efficient sets for VOP will be denoted, respectively, by $E_e(F,S)$, $E_s(F,S)$, $E_p(F,S)$, and $E_w(F,S)$. To unify the presentation, we will denote by $E_\sigma(F,S)$ the set of σ-efficient points depending on the choice of $\sigma \in \{s,p,e,w\}$. By using the well-known property that $\operatorname{int} Y_+ \subseteq Y_+ \setminus l(Y_+)$, it is immediate that

$$E_p(F,S) \subseteq E_e(F,S) \subseteq E_w(F,S). \tag{2.1}$$

When $\dim Y = 1$, all these sets coincide with the ordinary optimal set $E_s(F,S)$. However, in vector optimization, i.e., $\dim Y \neq 1$, it does not often happen that

$$E_s(F,S) \neq \emptyset.$$

But if this situation occurs, one can easily show that $E_e(F,S) \subseteq E_s(F,S)$. The reverse inclusion $E_s(F,S) \subseteq E_e(F,S)$ being always satisfied, one therefore has

$$E_s(F,S) \neq \emptyset \Rightarrow E_s(F,S) = E_e(F,S).$$

The subdifferential of a vector mapping, when it is associated with a VOP, leads in connection with the different efficient sets to consider four concepts, namely, the strong, weak, efficient, and proper, defined, respectively, for $F : X \to Y \sqcup \{+\infty\}$ and $\bar{x} \in \operatorname{dom} F$ as follows:

- $\partial^s F(\bar{x}) = \{A \in L(X,Y) : \quad \forall x \in X,\ F(x) - F(\bar{x}) \geq_{Y_+} A(x-\bar{x})\}.$
- $\partial^w F(\bar{x}) = \{A \in L(X,Y) : \quad \nexists x \in X,\ F(x) - F(\bar{x}) <_{Y_+} A(x-\bar{x})\}.$
- $\partial^e F(\bar{x}) = \{A \in L(X,Y) : \quad \nexists x \in X,\ F(x) - F(\bar{x}) \lneq_{Y_+} A(x-\bar{x})\}.$
- $\partial^p F(\bar{x}) = \{A \in L(X,Y) : \quad \exists \hat{Y}_+ \subsetneq Y$ a convex cone that satisfies $Y_+ \setminus l(Y_+) \subseteq \operatorname{int} \hat{Y}_+,\ \nexists x \in X,\ F(x) - F(\bar{x}) \lneq_{\hat{Y}_+} A(x-\bar{x})\}.$

The notation $\partial^\sigma F(\bar{x})$ will also be used to stand for the σ-subdifferential according to the choice of $\sigma \in \{s,p,e,w\}$. By convention, we take $\partial^\sigma F(\bar{x}) = \emptyset$ if $\bar{x} \notin \operatorname{dom} F$. The σ-subdifferential, with $\sigma = s$ (resp. $\sigma \in \{e,w\}$) seems to first appear in [23] (resp. in [25]). All these definitions are justified by the importance of the following immediate property:

$$\bar{x} \in E_\sigma(F,X) \quad \Longleftrightarrow \quad 0 \in \partial^\sigma F(\bar{x}). \tag{2.2}$$

Similarly to inclusions (2.1), it is easily shown that

$$\partial^p F(\bar{x}) \subseteq \partial^e F(\bar{x}) \subseteq \partial^w F(\bar{x}), \tag{2.3}$$

while inclusion $\partial^s F(\bar{x}) \subseteq \partial^e F(\bar{x})$ is straightforward. If $(Y_+^*)^\circ \neq \emptyset$, we also have $\partial^s F(\bar{x}) \subseteq \partial^p F(\bar{x})$; see (3.3)–(3.4). But $\partial^s F(\bar{x}) \neq \emptyset \not\Rightarrow \partial^s F(\bar{x}) = \partial^e F(\bar{x})$ unlike the efficient sets. Indeed, let $X = \mathbb{R}$, $Y = \mathbb{R}^2$, $Y_+ = \mathbb{R}^2_+$, and $F(x) = (0,-x)$ if $x \geq 0$, $= (-x,-x)$ if $x < 0$. We have $(-1,0) \in \partial^e F(0)$, but $(-1,0) \notin \partial^s F(0) = [-1,0] \times \{-1\}$.

In scalar case ($Y = \mathbb{R}$, $Y_+ = \mathbb{R}_+$), all these sets coincide with the classical subdifferential of convex analysis ($\partial^s F$), usually denoted by ∂F.

3. Scalarization, σ-subdifferentiability, and relations with derivatives.

3.1. Characterization via scalarization. For $\lambda \in Y_+^* \setminus \{0\}$, the scalar function $\lambda \circ F : X \to \mathbb{R} \sqcup \{+\infty\}$ is defined by

$$\lambda \circ F(x) = \begin{cases} \langle \lambda, F(x) \rangle & \text{if } x \in \operatorname{dom} F, \\ +\infty & \text{else.} \end{cases}$$

For $\lambda = 0$, we take by convention $\lambda \circ F = 0$. Note that, $\forall\ \lambda \in Y_+^* \setminus \{0\}$,

- $\operatorname{dom}(\lambda \circ F) = \operatorname{dom} F$.
- F is Y_+-convex $\Longrightarrow \lambda \circ F$ is convex.
- F is sequentially Y_+-l.s.c $\Longrightarrow \lambda \circ F$ is l.s.c (see [11, Proposition 3.7]).

We unify the notation of the polar cones too by putting, for instance, for Y_+:

$$Y_+^\sigma = \begin{cases} Y_+^* \setminus \{0\} & \text{if } \sigma = w, \\ (Y_+^*)^\circ & \text{if } \sigma = p. \end{cases}$$

Then, recall the fundamental VOP scalarization principle whose proof in infinite-dimensional spaces can be found in Luc [19, pp. 91–92]. But we will give here an improved version by Bonnel [8]. Comparatively, the proof is direct and clearly simplified for the weak concept. For the proper one, the reflexivity assumption on Y, in fact not required, has been replaced by Y separated. It also seems that the essential assumption Y_+ is pointed, is only forgotten in [19].

THEOREM 3.1. *Let $F : X \supseteq S \to Y \sqcup \{+\infty\}$. Then,*

$$E_s(F,S) \subseteq \bigsqcap_{\lambda \in Y_+^* \setminus \{0\}} \operatorname*{argmin}_{x \in S} \langle \lambda, F(x) \rangle, \tag{3.1}$$

with equality if Y is locally convex and Y_+ is closed. For $\sigma \in \{p, w\}$,

$$E_\sigma(F,S) \supseteq \bigsqcup_{\lambda \in Y_+^\sigma} \operatorname*{argmin}_{x \in S} \langle \lambda, F(x) \rangle, \tag{3.2}$$

with equality if F is Y_+-convex and S is convex, with Y_+ pointed as $\sigma = p$.

Proof. The direct inclusion in (3.1) is straightforward, while the reverse one uses only Proposition 2.1(1).

Let us show the inclusion in (3.2).

Case $\sigma = p$. Let $\lambda \in Y_+^p$ and $\bar{x} \in \operatorname{argmin}_{x \in S} \langle \lambda, F(x) \rangle$. The pointed convex cone

$$\hat{Y}_+ = \{y \in Y :\ \langle \lambda, y \rangle > 0\} \sqcup \{0\}$$

obviously satisfies $\hat{Y}_+ \neq Y$ and $Y_+ \setminus l(Y_+) \subseteq \operatorname{int} \hat{Y}_+$. If $\bar{x} \notin E_p(F,S)$, there would exist $x \in S$ such that $F(x) \precneqq_{\hat{Y}_+} F(\bar{x})$, i.e., $F(\bar{x}) - F(x) \in \hat{Y}_+ \setminus \{0\}$, and hence, $\langle \lambda, F(\bar{x}) - F(x) \rangle > 0$ contradicting the choice of $\bar{x}$.

Case $\sigma = w$. Consider the convex cone $\widetilde{Y}_+ = \operatorname{int} Y_+ \sqcup \{0\}$. With the property that $\operatorname{int} Y_+ \subseteq Y_+ \setminus l(Y_+)$, we can see that $\widetilde{Y}_+$ is pointed. Moreover, by Proposition 2.1(3), we have that $Y_+^w \subseteq \widetilde{Y}_+^p$. Inclusion (3.2) showed for $\sigma = p$, when applied to $\widetilde{Y}_+$, gets $\bigsqcup_{\lambda \in Y_+^w} \operatorname{argmin}_{x \in S} \langle \lambda, F(x) \rangle \subseteq E_p(F,S,\widetilde{Y}_+)$ the p-efficient set with respect to $\widetilde{Y}_+$. But in view of (2.1), $E_p(F,S,\widetilde{Y}_+) \subseteq E_e(F,S,\widetilde{Y}_+) = E_w(F,S)$; the last equality being trivially true by the fact that $\widetilde{Y}_+$ is pointed.

Let us show now the reverse inclusion for (3.2).

Case $\sigma = w$. Let $\bar{x} \in E_w(F,S)$, then, by definition, $(F(\bar{x}) - \operatorname{int} Y_+) \sqcap F(S) = \emptyset$. Using the well-known property that $Y_+ + \operatorname{int} Y_+ \subseteq \operatorname{int} Y_+$, $(F(\bar{x}) - \operatorname{int} Y_+) \sqcap (F(S) + Y_+) = \emptyset$ holds too. Since F is Y_+-convex and S is convex, then the set $F(S) + Y_+$ is convex. So, by separation theorem, there exists $\lambda \in Y^* \setminus \{0\}$ such that

$$\forall x \in S,\ \forall y \in Y_+,\ \forall y' \in \operatorname{int} Y_+, \quad \langle \lambda, F(\bar{x}) - y' \rangle \leq \langle \lambda, F(x) + y \rangle.$$

Letting $y' \to 0$ and taking $y = 0$, we obtain $\bar{x} \in \operatorname{argmin}_{x \in S} \langle \lambda, F(x) \rangle$. Taking now $x = \bar{x}$ and $y = 0$, we see that $\langle \lambda, y' \rangle \geq 0\ \forall\ y' \in \operatorname{int} Y_+$. The well-known property $\operatorname{cl}(\operatorname{int} Y_+) = \operatorname{cl} Y_+$ ("cl" stands for the topological closure) shows that the last inequality extends to $y' \in Y_+$, i.e., $\lambda \in Y_+^w$.

Case $\sigma = p$. Let $\bar{x} \in E_p(F,S)$, then there exists $\hat{Y}_+ \subsetneq Y$ a convex cone such that $Y_+ \setminus \{0\} \subseteq \operatorname{int} \hat{Y}_+$ (Y_+ is now supposed pointed) and $\bar{x} \in E_e(F,S,\hat{Y}_+)$. In view of (2.1), $\bar{x} \in E_w(F,S,\hat{Y}_+)$ too. Moreover, F being Y_+-convex and $Y_+ \setminus \{0\} \subseteq \operatorname{int} \hat{Y}_+$, obviously F remains $\hat{Y}_+$-convex. So reverse inclusion (3.2) showed, for $\sigma = w$, may be applied with $\hat{Y}_+$, deducing thus $\lambda \in \hat{Y}_+^* \setminus \{0\}$ such that $\bar{x} \in \operatorname{argmin}_{x \in S} \langle \lambda, F(x) \rangle$. It remains to show that $\lambda \in Y_+^p$. If it was not the case, as $Y_+ \setminus \{0\} \subseteq \operatorname{int} \hat{Y}_+$, we therefore could find $y \in Y_+ \setminus \{0\}$ such that $\langle \lambda, y \rangle = 0$, contradicting Proposition 2.1(3). ∎

The next result characterizes scalarly the σ-subdifferential for $\sigma \in \{s,p,w\}$ and will be crucial for the basic results of this paper.

THEOREM 3.2. *Let $F : X \to Y \sqcup \{+\infty\}$. Then, $\forall \bar{x} \in X$,*

$$\partial^s F(\bar{x}) \subseteq \bigsqcap_{\lambda \in Y_+^* \setminus \{0\}} \{A \in L(X,Y) :\ \lambda \circ A \in \partial(\lambda \circ F)(\bar{x})\}, \tag{3.3}$$

with equality if Y is locally convex and Y_+ is closed. For $\sigma \in \{p,w\}$, $\forall \bar{x} \in X$,

$$\partial^\sigma F(\bar{x}) \supseteq \bigsqcup_{\lambda \in Y_+^\sigma} \{A \in L(X,Y) :\ \lambda \circ A \in \partial(\lambda \circ F)(\bar{x})\}, \tag{3.4}$$

with equality if F is Y_+-convex, with Y_+ pointed as $\sigma = p$.

Proof. It is clear that, for $\bar{x} \notin \operatorname{dom} F$, all the previous sets are empty. Hence it suffices to consider $\bar{x} \in \operatorname{dom} F$ and to see as in (2.2) that

$$A \in \partial^\sigma F(\bar{x}) \quad \Longleftrightarrow \quad A \in L(X,Y) :\ \bar{x} \in E_\sigma(F - A, X) \tag{3.5}$$

and then to apply scalarization formulas (3.1)–(3.2). ∎

In the finite-dimensional space $Y = \mathbb{R}^r$ equipped with its natural order $Y_+ = \mathbb{R}^r_+$, the strong subdifferential reduces by the very definition to

$$\partial^s (f_1, \ldots, f_r)(\bar{x}) = \partial f_1(\bar{x}) \times \cdots \times \partial f_r(\bar{x}). \tag{3.6}$$

The equality in (3.4) was already proved by Taa [27] for w-subdifferentials of set-valued maps in Banach spaces using a direct scalarizing proof. It was also proved for the e-subdifferential by Sawaragi, Nakayama, and Tanino [26] in finite-dimensional spaces with an additional hypothesis on a Fenchel-type conjugate map of F.

3.2. σ-subdifferentiability and regular subdifferentiability. The vector mapping $F : X \to Y \sqcup \{+\infty\}$ is said to be at $\bar{x} \in \operatorname{dom} F$:

▶ subdifferentiable (resp. strongly, properly, or weakly subdifferentiable) if

$$\partial^\sigma F(\bar{x}) \neq \emptyset$$

for $\sigma = e$ (resp., $\sigma = s, p, w$). In brief, σ-subdifferentiable when $\sigma \in \{e, s, p, w\}$;

▶ regular subdifferentiable, if

$$\partial(\lambda \circ F)(\bar{x}) = \lambda \circ \partial^s F(\bar{x}) \qquad \forall \lambda \in Y_+^*,$$

where it is understood that the set $\lambda \circ \partial^s F(\bar{x}) := \{\lambda \circ A : A \in \partial^s F(\bar{x})\}$.

This property, due to Raffin [23], essentially reposes upon the inclusion "$\subseteq$", the reverse one "$\supseteq$" being trivial. The absence of the word "strongly" in this definition will be justified next. Afterwards, it is normal that the subdifferential regularity supposes a priori a nonempty strong subdifferential, that is why, by definition, it includes the value $\lambda = 0$. But, in most cases, we will just need the following weaker concepts, depending on the choice of $\sigma \in \{p, w\}$ in which the strong subdifferentiability is anyway not necessary: $F : X \to Y \sqcup \{+\infty\}$ will be said to be at $\bar{x} \in \operatorname{dom} F$

▶ σ-regular subdifferentiable, with $\sigma \in \{p, w\}$, if

$$\partial(\lambda \circ F)(\bar{x}) = \lambda \circ \partial^s F(\bar{x}) \qquad \forall \lambda \in Y_+^\sigma.$$

From scalarization Theorem 3.2, we establish first the σ-subdifferentiability for $\sigma \in \{p, e, w\}$ under recent conditions similar to those of the scalar case.

PROPOSITION 3.1. *Let X be a Fréchet space and $F \in \Gamma(X, Y)$ be such that at $\bar{x} \in \operatorname{dom} F$, the following Attouch–Brézis qualification condition is fulfilled:*

$$\mathbb{R}_+[\operatorname{dom} F - \bar{x}] \quad \textit{is a closed vector subspace of } X.$$

Then, for $\sigma \in \{p, w\}$,

$$\partial^\sigma F(\bar{x}) \neq \emptyset \qquad \textit{iff} \qquad \exists \lambda \in Y_+^\sigma : \lambda \circ F \textit{ is l.s.c at } \bar{x},$$

where the necessary condition demands Y_+ to be pointed [4] *as $\sigma = p$.*

Proof. We use the following result recently established by Laghdir [15].[5]

LEMMA 3.1. *Let X be a Fréchet space and $f \in \Gamma(X)$ be such that at $\bar{x} \in \operatorname{dom} f$, $\mathbb{R}_+[\operatorname{dom} f - \bar{x}]$ is a closed vector subspace of X. Then $\partial f(\bar{x}) \neq \emptyset$ iff f is l.s.c at $\bar{x}$.*

Assume that F satisfies the general hypotheses of Proposition 3.1. Then it is clear that, for any fixed $\lambda \in Y_+^\sigma$, the functional $\lambda \circ F$ also satisfies the general hypotheses of the lemma. If $\lambda \circ F$ is l.s.c at $\bar{x}$, we therefore can apply Lemma 3.1 and deduce the existence of $x_\lambda^* \in \partial(\lambda \circ F)(\bar{x})$. But $\lambda \in Y_+^\sigma$, with $\sigma = p$ (resp. $\sigma = w$) implies, by definition (resp. by Proposition 2.1(4)), the existence of $y_\lambda \in Y_+ \setminus l(Y_+)$ (resp. $y_\lambda \in \operatorname{int} Y_+$) such that $\langle \lambda, y_\lambda \rangle = 1$. Let $A_\lambda : X \to Y$ be defined by

$$A_\lambda(x) = \langle x_\lambda^*, x \rangle \, y_\lambda. \tag{3.7}$$

[4] With Y_+ pointed, following the equality in (3.4) of Theorem 3.2, if $Y_+^p = \emptyset$, then $\partial^p F(\bar{x}) = \emptyset$.

[5] This result improves the well-known Attouch–Brézis theorem [3] about the subdifferentiability of a scalar function, in a sense that l.s.c is not required more than at the point in question, in which case, obviously, the condition becomes necessary and sufficient. In fact, necessity even holds in a topological vector space and without a qualification condition in Lemma 3.1 as well as Proposition 3.1. Note also that if $\bar{x} \in \operatorname{int} \operatorname{dom} f$, then $\mathbb{R}_+[\operatorname{dom} f - \bar{x}] = X$, in particular, if f is continuous at $\bar{x}$.

Then $A_\lambda \in L(X,Y)$ and $\lambda \circ A_\lambda = x^*_\lambda \in \partial(\lambda \circ F)(\bar{x})$. It follows by Theorem 3.2(3.4) that $A_\lambda \in \partial^\sigma F(\bar{x})$, and the sufficient condition is proved. The necessary condition is obtained easily using again Theorem 3.2 and Lemma 3.1. □

While until now, strong subdifferentiability requires conditions slightly stronger. Recently, Zălinescu [34] extended the strong subdifferentiability of $F \in \Gamma(X,Y)$ at $\bar{x} \in \text{core}(\text{dom}\, F) := \{x \in \text{dom}\, F : \forall x' \in X, \exists \alpha > 0,\ x + [-\alpha,\alpha]x' \subseteq \text{dom}\, F\}$ (the algebraic interior) established by Valadier [32, Theorem 6 and Remark 6] in real linear spaces X and Y, with Y an order complete vector lattice and Y_+ pointed, to $\bar{x} \in \text{ri}(\text{dom}\, F) := \{x \in \text{dom}\, F : \forall x' \in \text{dom}\, F, \exists \alpha > 0,\ x + \alpha(x - x') \in \text{dom}\, F\}$ (the relative interior). Well before, Zowe [35] extended some results of Valadier [32] from order complete vector lattices to separated locally convex topological vector spaces ordered by closed convex cones and established the strong subdifferentiability of $F \in \Gamma(X,Y)$ under the three following conditions:

(H_1) X is a reflexive Banach space,

(H_2) F is continuous at $\bar{x} \in \text{int}\, \text{dom}\, F$,

(H_3) $\text{int}\, Y^*_+ \neq \emptyset$.

He also proved that if, furthermore, the following condition is realized:

(H_4) all order intervals $[a,b]$ are relatively weakly compact in Y,

then F is even regular subdifferentiable at $\bar{x}$. Moreover, he showed next that hypotheses (H_3) and (H_4) can be replaced by the following condition:

(H_5) Y is semireflexive and Y_+ has a weakly compact base lying in a closed hyperplane not running through 0.

Recall that a nonempty convex subset $B \subseteq Y_+$ not containing 0 is called a base of Y_+, if each $y \in Y_+ \setminus \{0\}$ has a unique representation $y = \alpha b$, where $b \in B$ and $\alpha > 0$.

The (σ-)regular subdifferentiability concept will reveal to be crucial in what follows. So, we suggest making on it some commentaries surely in some special cases of helpful interest. In the infinite-dimensional case, we have, for instance, the following processes.

Remark 1. Hypothesis (H_2) holds, for example, if $F \in \Gamma_0(X,Y)$, where Y is a Fréchet space equipped with a closed pointed and normal preorder Y_+ (Théra [29]). Recall that a point of continuity is trivially necessarily in the interior of the effective domain. On the other hand, it has been observed in [35] that it is easy to construct closed pointed convex cones Y_+ satisfying (H_5): $Y_+ = \bigsqcup_{\alpha \geq 0} \alpha B$, with B a convex weakly compact subset of a closed hyperplane in Y not containing 0.

In finite dimension, (H_5) is always fulfilled, and (H_2) may be weakened. Let us see this in the two following remarks.

Remark 2. In finite dimension, a cone has a compact base iff it is closed and pointed (see [19], e.g.). But following Peressini–Borwein (see again [19]), a base in a vector space is always of the form

$$B = \{y \in Y_+ :\ \langle \lambda, y \rangle = 1\}$$

for some λ in the algebraic strict polar cone of Y_+. This means that every mapping of $\Gamma(X, \mathbb{R}^r)$ in the setting of a reflexive Banach space X and a preorder Y_+ closed pointed is, at every point of continuity, regular subdifferentiable.

Remark 3. When $Y_+ = \mathbb{R}^r_+$, we have $Y^*_+ = \mathbb{R}^r_+$ and $(Y^*_+)^\circ = \text{int}\, \mathbb{R}^r_+$. So, by (3.6), subdifferential regularity (resp. σ-regularity for $\sigma = w, p$) of $F = (f_1, \ldots, f_r)$ becomes exactly a well-known chain rule of convex analysis: $\forall (\lambda_1, \ldots, \lambda_r) \geq 0$ (resp. $\gneq 0, > 0$),

$$\partial \left(\sum_{i=1}^r \lambda_i f_i \right) (\bar{x}) = \sum_{i=1}^r \lambda_i \partial f_i(\bar{x}).$$

The formula holds with the respective λ_is under the well-known Moreau–Rockafellar qualification condition [20, 24]:

$$(MR)\quad \begin{cases} f_i \in \Gamma(X) \quad (i = 1 \cdots r),\ X \text{ locally convex}, \\ (r-1) \text{ functions } f_i \text{ are finite and continuous at a point of} \\ \text{the effective domain of the other one}, \end{cases}$$

provided that all the f_is are subdifferentiable at $\bar{x}$ (resp. every one or none of them is it at $\bar{x}$ as $\sigma = w$ and for all $\bar{x} \in X$ as $\sigma = p$). For the latter, the formula even holds under the following weaker qualification condition of the Attouch–Brézis kind [11, pp. 139–140]:

$$(AB)\quad \begin{cases} f_i \in \Gamma_0(X) \quad (i = 1 \cdots r),\ X \text{ a Fréchet space}, \\ \mathbb{R}_+[\Delta X^r - \prod_{i=1}^r \operatorname{dom} f_i] \quad \text{is a closed vector subspace of } X^r, \end{cases}$$

where $X^r = X \times \cdots \times X$ (r times) and $\Delta X^r = \{(x, \ldots, x) \in X^r\}$. The (AB)-type conditions have the advantage to be in some sense weaker, with those of (MR) easy to check.

The σ-regular subdifferentiability appears first to be of great interest for finding the gap between vector and scalar concepts, a well-known question in vector analysis. Indeed, for the vector subdifferentials, we obtain the following relationships.

THEOREM 3.3. *Let $F \in \Gamma(X, Y)$ be σ-regular subdifferentiable at $\bar{x}$, $\sigma \in \{p, w\}$ and Y_+ be pointed as $\sigma = p$. Then,*

$$\partial^\sigma F(\bar{x}) = \partial^s F(\bar{x}) + Z_\sigma(X, Y), \tag{3.8}$$

where $Z_\sigma(X, Y) = \{A \in L(X, Y) :\ \exists\ \lambda \in Y_+^\sigma,\ \lambda \circ A = 0\}$ is the set of σ-zerolike linear continuous operators ($A \sim_\sigma 0$), which represents here the gap between $\partial^\sigma F$ and $\partial^s F$.

Proof. By equality in (3.4) and σ-regular subdifferentiability, we have

$$\begin{aligned} A \in \partial^\sigma F(\bar{x}) &\Leftrightarrow A \in L(X, Y),\ \exists \lambda \in Y_+^\sigma :\ \lambda \circ A \in \partial(\lambda \circ F)(\bar{x}) = \lambda \circ \partial^s F(\bar{x}) \\ &\Leftrightarrow A \in L(X, Y),\ \exists B \in \partial^s F(\bar{x}),\ \exists \lambda \in Y_+^\sigma :\ \lambda \circ (A - B) = 0 \\ &\Leftrightarrow \exists B \in \partial^s F(\bar{x}) :\ A - B \in Z_\sigma(X, Y) \\ &\Leftrightarrow A \in \partial^s F(\bar{x}) + Z_\sigma(X, Y). \qquad \square \end{aligned}$$

As $Z_\sigma(X, Y) = -Z_\sigma(X, Y)$, this set may be replaced in (3.8) by its opposite.

In the finite-dimensional case $Y = \mathbb{R}^r$ and $Y_+ = \mathbb{R}^r_+$, we have that $Y_+^w = \mathbb{R}^r_+ \setminus \{0\}$ and $Y_+^p = \operatorname{int} \mathbb{R}^r_+$. So, a matrix $A = (A_1, \ldots, A_r) \in \mathbb{R}^{r \times n}$ is in $Z_w(\mathbb{R}^n, \mathbb{R}^r)$ (resp. in $Z_p(\mathbb{R}^n, \mathbb{R}^r)$) if there exists a vector λ of r nonnegative components λ_i not all equal to zero (resp. all positive) such that $A^T\lambda = \sum_{i=1}^r \lambda_i A_i = 0$. The set $Z_p(\mathbb{R}^n, \mathbb{R}^r)$ of p-zerolike matrices has already been evoked in [26, pp. 204], where the authors claim, for $F \in \Gamma(\mathbb{R}^n, \mathbb{R}^r)$, that[6]

$$\bar{x} \in E_p(F, \mathbb{R}^n) \quad \Leftrightarrow \quad \exists A \sim_p 0 :\ A \in \partial^s F(\bar{x}) \quad \Leftrightarrow \quad \exists A \sim_p 0 :\ \bar{x} \in E_s(F - A, \mathbb{R}^n).$$

The second relation being trivial from (3.5). While in comparison with (3.8), the first relation can be viewed by (2.2) as

$$0 \in \partial^p F(\bar{x}) \quad \Longleftrightarrow \quad 0 \in \partial^s F(\bar{x}) + Z_p(\mathbb{R}^n, \mathbb{R}^r).$$

[6] We just wonder if this result does not require a certain regularity on F, since it is announced without proof.

Remark 4. As predicted, let us underline that, in general spaces, for instance, Y is locally convex equipped with a closed pointed preorder Y_+, we effectively cannot talk about σ-subdifferential regularity for $\sigma \in \{p, e, w\}$. In fact, for all $\lambda \in Y_+^*$, $\partial(\lambda \circ F)(\bar{x}) = \lambda \circ \partial^\sigma F(\bar{x})$ for $\sigma \in \{p, e, w\}$ would imply in view of the equality in (3.3) that $\partial^\sigma F(\bar{x}) \subseteq \partial^s F(\bar{x})$. With the respective reverse inclusions, the subdifferentials of Pareto type would coincide with the strong one, and the function would be regular subdifferentiable too. This may be in contradiction with Theorem 3.3, and this justifies the absence of the word "strongly" in the regular subdifferentiability appellation.

3.3. Characterizations via derivatives. Let us recall the most popular notions of derivatives for vector mappings. $F : X \to Y \sqcup \{+\infty\}$ is said to be, at $\bar{x} \in \operatorname{dom} F$,

- directionally differentiable, in brief, D-differentiable, if $\forall d \in X$,
$$\lim_{t \searrow 0^+} \frac{F(\bar{x} + td) - F(\bar{x})}{t} = l(d) \in Y,$$
we denote $F'(\bar{x}; d) = l(d)$ the directional derivative of F at $\bar{x}$ in the direction d;
- Gâteaux differentiable, in brief, G-differentiable, if $\exists A \in L(X,Y)$, $\forall d \in X$,
$$F'(\bar{x}; d) = \lim_{t \to 0} \frac{F(\bar{x} + td) - F(\bar{x})}{t} = A(d),$$
we denote $F'_G(\bar{x}) = A$ the Gâteaux derivative of F at $\bar{x}$;
- Fréchet differentiable, in brief, F-differentiable, if $\exists A \in L(X, Y)$,
$$\lim_{h \to 0} \frac{F(\bar{x} + h) - F(\bar{x}) - A(h)}{\|h\|} = 0,$$
we denote $F'(\bar{x}) = A$ the Fréchet derivative of F at $\bar{x}$; here, X being, of course, normed.

Obviously,

- F-differentiable at $\bar{x} \Longrightarrow$ G-differentiable at $\bar{x} \Longrightarrow$ D-differentiable at $\bar{x}$.
- In this case, $F'(\bar{x}) = F'_G(\bar{x}) = F'(\bar{x}; \cdot)$.

For a finite-dimensional image space, the D-differentiability of $F = (f_1, \dots, f_r)$ is, of course, equivalent to D-differentiability of functionals f_is. But, it is well known that every proper convex functional always admits directional derivatives in all direction at any point of subdifferentiability. So the D-differentiability of F holds under the conditions of Lemma 3.1 and, in particular, at every point of $\operatorname{int} \operatorname{dom} F = \prod_{i=1}^r \operatorname{int} \operatorname{dom} f_i$, where F is sequentially $\mathbb{R}^r_+$-l.s.c or, strongly again, at every point of the continuity of F. For the case of infinite-dimensional image spaces, the D-differentiability of $F \in \Gamma(X, Y)$ was established in [35] under hypotheses (H_2)–(H_3) and the preorder normality assumption.

We express now the σ-subdifferentials in terms of the derivatives when they exist.

PROPOSITION 3.2. *Let Y be locally convex, Y_+ be closed, pointed as $\sigma = p$, and $F \in \Gamma(X, Y)$ be D-differentiable at $\bar{x}$. Then,*

$$\begin{aligned}
\partial^s F(\bar{x}) &= \{A \in L(X,Y) : \quad \forall d \in X,\ F'(\bar{x}; d) \geq_{Y_+} A(d)\},\\
\partial^w F(\bar{x}) &= \{A \in L(X,Y) : \quad \nexists d \in X,\ F'(\bar{x}; d) <_{Y_+} A(d)\},\\
\partial^p F(\bar{x}) &= \{A \in L(X,Y) : \quad \exists \hat{Y}_+ \subsetneq Y \text{ a convex cone that satisfies}\\
&\qquad Y_+ \setminus \{0\} \subseteq \operatorname{int} \hat{Y}_+,\ \nexists d \in X,\ F'(\bar{x}; d) \lneq_{\hat{Y}_+} A(d)\}.
\end{aligned}$$

Proof. We shall prove that, for $\sigma \in \{s, p, w\}$,

$$A \in \partial^\sigma F(\bar{x}) \iff A \in L(X,Y) :\ 0 \in E_\sigma(F'(\bar{x};\cdot) - A, X), \tag{3.9}$$

which, by the very definitions of σ-efficient sets, show the proposition. To this, observe first that, for all $\lambda \in Y_+^*$, by continuity of λ, for all $d \in X$,

$$\lambda \circ F'(\bar{x};d) = \lim_{t \searrow 0^+} \frac{\lambda \circ F(\bar{x}+td) - \lambda \circ F(\bar{x})}{t} = (\lambda \circ F)'(\bar{x};d). \tag{3.10}$$

On the other hand, by Theorem 3.2, to say $A \in \partial^\sigma F(\bar{x})$ for $\sigma = s$ (resp. $\sigma \in \{p, w\}$) is equivalent to saying that $\lambda \circ A \in \partial(\lambda \circ F)(\bar{x})\ \forall\ \lambda \in Y_+^* \setminus \{0\}$ (resp. for some $\lambda \in Y_+^\sigma$), with $A \in L(X,Y)$. As $\lambda \circ F \in \Gamma(X)$ is D-differentiable at $\bar{x}$, then we have

$$\partial(\lambda \circ F)(\bar{x}) = \{x^* \in X^* :\ \forall d \in X,\ (\lambda \circ F)'(\bar{x};d) \geq \langle x^*, d\rangle\}.$$

So $A \in \partial^\sigma F(\bar{x})$ is also equivalent to $\lambda \circ F'(\bar{x};d) \geq \lambda \circ A(d)\ \forall\ d \in X$ or equivalently, $0 \in \operatorname{argmin}_{d \in X} \langle \lambda, F'(\bar{x};d) - A(d)\rangle$, with the respective σs and λs. Property (3.9) thus follows from scalarization Theorem 3.1, since for $\sigma \in \{p, w\}$, $F'(\bar{x};\cdot)$ is well Y_+-convex. More precisely, as shown in what follows.

LEMMA 3.2. *With the same hypotheses as Proposition* 3.2*, vector function* $d \mapsto F'(\bar{x};d)$ *is positively homogeneous* Y_+*-subadditive, i.e.,* $\forall \alpha \in \mathbb{R}_+,\ \forall d_1,\ d_2 \in X$*,*

$$\begin{aligned} &F'(\bar{x};\alpha d_1) = \alpha F'(\bar{x};d_1),\\ &F'(\bar{x};d_1+d_2) \leq_{Y_+} F'(\bar{x};d_1) + F'(\bar{x};d_2). \end{aligned}$$

Proof of Lemma 3.2. Homogeneous positivity is immediate from the definition of $F'(\bar{x};d)$. While subadditivity, it follows from the one of $(\lambda \circ F)'(\bar{x};.)$, relation (3.10), and Proposition 2.1(1). ∎

PROPOSITION 3.3. *Let* Y *be locally convex,* Y_+ *be closed pointed, and* $F \in \Gamma(X,Y)$ *be both D-differentiable, p-regular subdifferentiable, and s-subdifferentiable at* $\bar{x}$*. Then,* $\forall d \in X$*,*

$$\begin{aligned} F'(\bar{x};d) &= \operatorname*{MAX_s}_{A \in \partial^s F(\bar{x})} A(d),\\ F'(\bar{x};d) &\in \operatorname*{MAX}_{\sigma\ A \in \partial^\sigma F(\bar{x})} A(d) \qquad \text{for } \sigma \in \{p, w\}, \end{aligned}$$

where $\operatorname{MAX}_{\sigma\, x \in D} G(x) = G(E_\sigma(-G, D))$ *denotes the set*[7] *of* σ*-maximal values of a vector mapping* G *over a subset* D*.*

Proof. For $\sigma = s$, taking into account the first result of Proposition 3.2, we have to prove that, for any fixed $d \in X$, there exists $A \in \partial^s F(\bar{x})$ such that $A(d) = F'(\bar{x};d)$. If we proceed by contradiction, it would mean that $F'(\bar{x};d) - A(d) \in Y_+ \setminus \{0\}$ for every $A \in \partial^s F(\bar{x})$. Picking $\lambda \in (Y_+^*)^\circ$, we would have

$$\langle \lambda, F'(\bar{x};d) - A(d)\rangle > 0 \qquad \forall A \in \partial^s F(\bar{x}). \tag{3.11}$$

Since $\lambda \circ F \in \Gamma(X)$ is subdifferentiable at $\bar{x}$ (see (3.3)), it holds that

$$(\lambda \circ F)'(\bar{x};d) = \max_{x^* \in \partial(\lambda \circ F)(\bar{x})} \langle x^*, d\rangle.$$

[7] It is easy to see that the set of strong maximal values $\operatorname{MAX}_{s\, x \in D} G(x)$ is a singleton if the preorder is pointed. So, we consider it a vector instead of a set.

Let $x^* \in \partial(\lambda \circ F)(\bar{x})$ be such that $(\lambda \circ F)'(\bar{x}; d) = \langle x^*, d\rangle$ realizing this maximum. Let $A^* \in \partial^s F(\bar{x})$ be given by p-regular subdifferentiability such that $x^* = \lambda \circ A^*$. Then, with (3.10), $\lambda \circ F'(\bar{x}; d) = (\lambda \circ F)'(\bar{x}; d) = \lambda \circ A^*(d)$ so in contradiction with (3.11). Case $\sigma \in \{p, w\}$ follows easily by contradiction too, using both the previous result and the two last formulas of Proposition 3.2. □

PROPOSITION 3.4. *Let Y be locally convex, Y_+ be closed, and $F \in \Gamma(X, Y)$.*

(1) *If F is G-differentiable at $\bar{x}$, with Y_+ pointed as $\sigma = p$, then*

$$\partial^s F(\bar{x}) = \{F'_G(\bar{x})\},$$
$$\partial^\sigma F(\bar{x}) = \{F'_G(\bar{x})\} + Z_\sigma(X, Y) \qquad \text{for } \sigma \in \{p, w\}.$$

(2) *Conversely, if F is both D-differentiable and p-regular subdifferentiable at $\bar{x}$, with Y_+ pointed, and $\partial^s F(\bar{x}) = \{A\}$, then F is G-differentiable at $\bar{x}$ and $F'_G(\bar{x}) = A$.*

Proof. (1) Case $\sigma = s$ is directly coming from G-differentiability and Proposition 3.2. For $\sigma \in \{p, w\}$, let us prove first the following lemma.

LEMMA 3.3. *In Y locally convex with Y_+ closed, vector mapping $F \in \Gamma(X, Y)$, which is G-differentiable at $\bar{x}$, is regular subdifferentiable at this point.*

Proof of Lemma 3.3. Similarly to (3.10), we have that $(\lambda \circ F)'_G(\bar{x}) = \lambda \circ F'_G(\bar{x})$ for each $\lambda \in Y_+^*$, so functional $\lambda \circ F \in \Gamma(X)$ is G-differentiable at $\bar{x}$. Hence $\partial(\lambda \circ F)(\bar{x}) = \{(\lambda \circ F)'_G(\bar{x})\} = \lambda \circ \{F'_G(\bar{x})\} = \lambda \circ \partial^s F(\bar{x})$. □

If F is regular subdifferentiable, then it is obviously a w-regular one. If $Y_+^p = \emptyset$, then $\partial^p F(\bar{x})$ and $Z_p(X, Y)$ are empty; otherwise, F is also p-regular subdifferentiable. The result of assertion (1) for $\sigma \in \{p, w\}$ then follows from Theorem 3.3.

(2) This assertion is an immediate consequence of Proposition 3.3. □

Propositions 3.2 and 3.3 were proved for $\sigma = s$ by Valadier [32] in the setting of normal order complete vector lattices with continuity hypothesis (H_2), while Proposition 3.4, it was again shown for $\sigma = s$ by Zowe [35] in the more general setting of ordered separated locally convex topological vector spaces using the order normality assumption and hypotheses (H_1) to (H_4).

We end the section by the following corollaries.

COROLLARY 3.1. *Let Y be locally convex, Y_+ be closed, and pointed as $\sigma = p$, $F \in \Gamma(X, Y)$, and $\sigma \in \{p, w\}$.*

(1) *If F is σ-regular subdifferentiable at $\bar{x}$, then*

$$\partial^\sigma F(\bar{x}) = \bigsqcup_{A \in \partial^s F(\bar{x})} \partial^\sigma A(x) \qquad \forall x \in X.$$

(2) *If F is, in particular, G-differentiable at $\bar{x}$, then*

$$\partial^\sigma F(\bar{x}) = \partial^\sigma (F'_G(\bar{x}))(x) \qquad \forall x \in X.$$

(3) *In particular, one has*

$$Z_\sigma(X, Y) = \partial^\sigma 0(x) \qquad \forall x \in X.$$

Indeed, since a continuous linear operator A is G-differentiable and $A'_G(x) = A \ \forall\ x \in X$, then according to Theorem 3.3 and Proposition 3.4(1), we have that $\partial^\sigma F(\bar{x}) = \bigsqcup_{A \in \partial^s F(\bar{x})} \{A\} + Z_\sigma(X, Y) = \bigsqcup_{A \in \partial^s F(\bar{x})} \partial^\sigma A(x) \quad (\forall x \in X)$.

4. σ-subdifferential calculus rules. We are concerned in this section with the subdifferential calculus of the sum and/or the composition of convex vector mappings.

4.1. Addition. In literature, one may find some calculus rules about the sum of two convex vector mappings. The following formula was established by Théra [29] in the setting of a normal order complete vector topological space Y, where one of the functions is supposed to be continuous over a separated locally convex space X:

$$\partial^s(F+G)(\bar{x}) = \partial^s F(\bar{x}) + \partial^s G(\bar{x}).$$

In the more general setting of Banach spaces partially ordered, a similar result was proved in the case $\sigma = w$ by Lin [18] and Taa [27] for convex set-valued maps under a Slater-type qualification condition:

$$\partial^w(F+G)(\bar{x}) \subset \partial^w F(\bar{x}) + \partial^w G(\bar{x}).$$

Let us emphasize that unlike the extended scalar concept, i.e., $\sigma = s$, for which the reverse inclusion is trivial, in fact, with Pareto concepts, this converse cannot hold. Look at this with this simple example of very regular vector mappings: Let $F,\ G \in L(\mathbb{R}^2, \mathbb{R}^2)$ be defined by $F(x_1, x_2) = (x_1, 0)$ and $G(x_1, x_2) = (0, x_2)$. At $\bar{x} = (0,0)$, $F'(0,0) = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$, $G'(0,0) = \left(\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, and $(F+G)'(0,0) = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$. But $A = \left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$ is a matrix which verifies $\langle \lambda, A - F'(0,0)\rangle = 0$, with $\lambda = (1,0)$ or equivalently, by taking natural order $\mathbb{R}^2_+$, $A \in F'(0,0) + Z_w(\mathbb{R}^2, \mathbb{R}^2) = \partial^w F(0,0)$ according to Proposition 3.4(1). In a similar way, by taking $\lambda = (0,1)$, we verify that the matrix $-A \in \partial^w G(0,0)$. Hence, the matrix $0 = A - A \in \partial^w F(0,0) + \partial^w G(0,0)$. However, $0 \notin \partial^w(F+G)(0,0)$, since otherwise, by the Proposition 3.4(1) again, we obtain, for some $\lambda \in \mathbb{R}^2_+ \setminus \{0\}$, contradiction $0 = \langle \lambda, (F+G)'(0,0)\rangle = \lambda$. Thus

$$\partial^w F(0,0) + \partial^w G(0,0) \not\subset \partial^w(F+G)(0,0).$$

In fact, the presence of the strong subdifferential establishes the desired equality.

THEOREM 4.1. *Let $F,\ G : X \to Y \sqcup \{+\infty\}$, and $\sigma \in \{p, w\}$. Then, $\forall \bar{x} \in X$,*

$$\partial^\sigma(F+G)(\bar{x}) \supseteq \partial^\sigma F(\bar{x}) + \partial^s G(\bar{x}).$$

Assume now that G is σ-regular subdifferentiable at $\bar{x}$, Y_+ is pointed as $\sigma = p$, and one of the two following qualification conditions is satisfied:

$$(MR)_1 \quad \begin{cases} F,\ G \in \Gamma(X,Y),\ X \text{ locally convex}, \\ \text{one of the two functions is finite and continuous} \\ \text{at some point of the effective domain of the other one.} \end{cases}$$

$$(AB)_1 \quad \begin{cases} F,\ G \in \Gamma_0(X,Y),\ X \text{ Fréchet space}, \\ \mathbb{R}_+[\operatorname{dom} F - \operatorname{dom} G] \quad \text{is a closed vector subspace of } X. \end{cases}$$

Then,

$$\partial^\sigma(F+G)(\bar{x}) = \partial^\sigma F(\bar{x}) + \partial^s G(\bar{x}).$$

Proof. Let us prove the first inclusion for $\sigma = w$. Let $A \in \partial^w F(\bar{x})$, $B \in \partial^s G(\bar{x})$. We proceed by contradiction: If $A + B \notin \partial^w(F+G)(\bar{x})$, we would have

$$F(x_0) + G(x_0) - F(\bar{x}) - G(\bar{x}) - A(x_0 - \bar{x}) - B(x_0 - \bar{x}) \in -\operatorname{int} Y_+$$

for some $x_0 \in \operatorname{dom}(F+G) = \operatorname{dom} F \sqcap \operatorname{dom} G$. Then since $B \in \partial^s G(\bar{x})$, we would have

$$-G(x_0) + G(\bar{x}) + B(x_0 - \bar{x}) \in -Y_+.$$

Adding term by term and taking into account the fact that $-Y_+ - \operatorname{int} Y_+ \subseteq -\operatorname{int} Y_+$, we would obtain

$$F(x_0) - F(\bar{x}) - A(x_0 - \bar{x}) \in -\operatorname{int} Y_+,$$

which would contradict $A \in \partial^w F(\bar{x})$. Case $\sigma = p$ is obtained similarly by using now the fact that $-\operatorname{int} \hat{Y}_+ \subseteq -\hat{Y}_+ \setminus l(\hat{Y}_+)$ and $-\hat{Y}_+ \setminus l(\hat{Y}_+) - \hat{Y}_+ \setminus l(\hat{Y}_+) \subseteq -\hat{Y}_+ \setminus l(\hat{Y}_+)$; $\hat{Y}_+$ being given by the p-subdifferential definition.

We show the reverse inclusion by scalarization for $\sigma \in \{p, w\}$ simultaneously. So let $A \in \partial^\sigma(F+G)(\bar{x})$, then, by Theorem 3.2, there exists $\lambda \in Y_+^\sigma$ such that

$$\lambda \circ A \in \partial[\lambda \circ (F+G)](\bar{x}) = \partial(\lambda \circ F + \lambda \circ G)(\bar{x}).$$

Following condition $(MR)_1$ or $(AB)_1$, scalar functions $\lambda \circ F$ and $\lambda \circ G$, then satisfy exactly the qualification hypothesis of Moreau–Rockafellar or Attouch–Brézis. Hence, the subdifferential addition formula for functionals $\lambda \circ F$ and $\lambda \circ G$ applies, and, with the σ-regular subdifferentiability assumption on G, we therefore obtain

$$\partial(\lambda \circ F + \lambda \circ G)(\bar{x}) = \partial(\lambda \circ F)(\bar{x}) + \partial(\lambda \circ G)(\bar{x}) = \partial(\lambda \circ F)(\bar{x}) + \lambda \circ \partial^s G(\bar{x}).$$

Thus we deduce $B \in \partial^s G(\bar{x})$ such that $\lambda \circ (A - B) \in \partial(\lambda \circ F)(\bar{x})$ which shows, again by Theorem 3.2, that $A - B \in \partial^\sigma F(\bar{x})$ or equivalently $A \in \partial^\sigma F(\bar{x}) + \partial^s G(\bar{x})$. □

Using Lemma 3.3 and Proposition 3.4(1), we easily obtain the following corollary.

COROLLARY 4.1. *Let X and Y be locally convex, Y_+ be closed, pointed as $\sigma = p$, $F \in \Gamma(X, Y)$, $A \in L(X, Y)$, and $\sigma \in \{p, w\}$. Then, $\forall \bar{x} \in X$,*

$$\partial^\sigma(F + A)(\bar{x}) = \partial^\sigma F(\bar{x}) + A.$$

In finite-dimensional image space, according to Remark 3, Theorem 4.1 becomes the following.

COROLLARY 4.2. *Let X be locally convex, f_i, $g_i \in \Gamma(X)$ $(i = 1 \cdot\cdot\, r)$, and consider the following hypotheses:*

(MR) *$(r-1)$ g_is are finite and continuous at a point of domain of the other one.*

(AB) *The g_is are l.s.c, $\mathbb{R}_+[\Delta X^r - \prod_{i=1}^r \operatorname{dom} g_i]$ is a closed vector subspace of X^r.*

$(MR)_1$ *All the g_is are finite and continuous at some point of $\prod_{i=1}^r \operatorname{dom} f_i$.*

$(MR)'_1$ *All the f_is are finite and continuous at some point of $\prod_{i=1}^r \operatorname{dom} g_i$.*

$(AB)_1$ *The f_is, g_is are l.s.c, $\mathbb{R}_+[\prod_{i=1}^r \operatorname{dom} f_i - \prod_{i=1}^r \operatorname{dom} g_i]$ is closed vector subspace.*

If one of conditions $(MR)_1$, $\big[(MR)'_1$ and $(MR)\big]$, $\big[(MR)'_1$ and (AB), with X Fréchet$\big]$, $\big[(AB)_1$ and (MR), with X Fréchet$\big]$, or $\big[(AB)_1$ and (AB), with X Fréchet$\big]$ is satisfied, then, $\forall \bar{x} \in X$,

$$\partial^p(f_1 + g_1, \ldots, f_r + g_r)(\bar{x}) = \partial^p(f_1, \ldots, f_r)(\bar{x}) + \prod_{i=1}^r \partial g_i(\bar{x}).$$

If, in addition to condition $(MR)_1$, $\big[(MR)'_1$ and $(MR)\big]$, or $\big[(AB)_1$ and (MR), with X Fréchet$\big]$, every one or none of the g_is is subdifferentiable at $\bar{x}$, then

$$\partial^w(f_1 + g_1, \ldots, f_r + g_r)(\bar{x}) = \partial^w(f_1, \ldots, f_r)(\bar{x}) + \prod_{i=1}^r \partial g_i(\bar{x}).$$

4.2. Composition. We present in this subsection several results concerning the σ-subdifferential calculus of the composite of two convex vector mappings. The main one of them is the following.

THEOREM 4.2. *Let $F : X \to Y \sqcup \{+\infty\}$, $G : Y \to Z \sqcup \{+\infty\}$, and $\sigma \in \{p, w\}$. Then, $\forall \bar{x} \in X$,*

$$\partial^\sigma (G \circ F)(\bar{x}) \supseteq \bigsqcup_{A \in \partial^s G(F(\bar{x}))} \partial^\sigma (A \circ F)(\bar{x}).$$

Assume now that G is (Y_+, Z_+)-nondecreasing and σ-regular subdifferentiable at $F(\bar{x})$, Z_+ is pointed as $\sigma = p$, and one of the two following qualification conditions is satisfied:

$$(MR)_2 \quad \begin{cases} F \in \Gamma(X, Y),\ G \in \Gamma(Y, Z),\ X \text{ and } Y \text{ locally convex}, \\ G \text{ is finite and continuous at some point of } \operatorname{Im} F := F(\operatorname{dom} F). \end{cases}$$

$$(AB)_2 \quad \begin{cases} F \in \Gamma_0(X, Y),\ G \in \Gamma_0(Y, Z),\ X \text{ and } Y \text{ Fréchet spaces}, \\ \mathbb{R}_+[\operatorname{dom} G - \operatorname{Im} F] \text{ is a closed vector subspace of } Y. \end{cases}$$

Then,

$$\partial^\sigma (G \circ F)(\bar{x}) = \bigsqcup_{A \in \partial^s G(F(\bar{x}))} \partial^\sigma (A \circ F)(\bar{x}).$$

Proof. Let us prove the first inclusion for $\sigma = w$. Let $B \in \partial^w (A \circ F)(\bar{x})$ for some $A \in \partial^s G(F(\bar{x}))$. We proceed by contradiction: If $B \notin \partial^w (G \circ F)(\bar{x})$, we would have

$$G \circ F(x_0) - G \circ F(\bar{x}) - B(x_0 - \bar{x}) \in -\operatorname{int} Z_+$$

for some $x_0 \in \operatorname{dom}(G \circ F) = F^{-1}(\operatorname{dom} G) \sqcap \operatorname{dom} F$. But $A \in \partial^s G(F(\bar{x}))$ would imply

$$-G \circ F(x_0) + G \circ F(\bar{x}) + A \circ F(x_0) - A \circ F(\bar{x}) \in -Z_+.$$

Adding term by term, by using that $-Z_+ - \operatorname{int} Z_+ \subseteq -\operatorname{int} Z_+$, we would obtain

$$A \circ F(x_0) - A \circ F(\bar{x}) - B(x_0 - \bar{x}) \in -\operatorname{int} Z_+,$$

which would contradict $B \in \partial^w (A \circ F)(\bar{x})$. The case of $\sigma = p$ is similarly obtained by the same arguments as for Theorem 4.1.

To show the reverse inclusion, we proceed by scalarization. To this end, we need the following subdifferential formula of a scalar composite recently proved by Combari, Laghdir, and Thibault [11], obtained well before by Lemaire [16] in the setting of Banach spaces under the Moreau–Rockafellar-type hypothesis.

LEMMA 4.1. *Let $F : X \to Y \sqcup \{+\infty\}$ and $g : Y \to \mathbb{R} \sqcup \{+\infty\}$ be a functional Y_+-nondecreasing such that one of the two following qualification conditions holds:*

$$(MR)_2 \quad \begin{cases} F \in \Gamma(X, Y),\ g \in \Gamma(Y),\ X \text{ and } Y \text{ locally convex}, \\ g \text{ is finite and continuous at some point of } \operatorname{Im} F. \end{cases}$$

$$(AB)_2 \quad \begin{cases} F \in \Gamma_0(X, Y),\ g \in \Gamma_0(Y),\ X \text{ and } Y \text{ Fréchet spaces}, \\ \mathbb{R}_+[\operatorname{dom} g - \operatorname{Im} F] \text{ is a closed vector subspace of } Y. \end{cases}$$

Then, $\forall \bar{x} \in X$,

$$\partial(g \circ F)(\bar{x}) = \bigsqcup_{\lambda \in \partial g(F(\bar{x}))} \partial(\lambda \circ F)(\bar{x}).$$

We come back to Theorem 4.2. Let $B \in \partial^\sigma(G \circ F)(\bar{x})$, then, by Theorem 3.2,

$$\mu \circ B \in \partial(\mu \circ G \circ F)(\bar{x}) \quad \text{for some} \quad \mu \in Z_+^\sigma.$$

It is easy to see that F and $\mu \circ G$ satisfies the conditions of Lemma 4.1. Hence,

$$\partial(\mu \circ G \circ F)(\bar{x}) = \bigsqcup_{\lambda \in \partial(\mu \circ G)(F(\bar{x}))} \partial(\lambda \circ F)(\bar{x}).$$

So, by σ-regular subdifferentiability of G, we deduce the existence of $A \in \partial^s G(F(\bar{x}))$ such that $\mu \circ B \in \partial(\mu \circ A \circ F)(\bar{x})$. Let us observe that

$$A \in \partial^s G(F(\bar{x})), \quad G \text{ is } (Y_+, Z_+)\text{-nondecreasing} \quad \Longrightarrow \quad A \in L_+(Y, Z),$$

since, in fact, $A \in L(Y, Z)$ and $0 \leq_{Z_+} G(F(\bar{x})) - G(F(\bar{x}) - y) \leq_{Z_+} A(y)\ \forall\ y \in Y_+$. Consequently, $A \circ F \in \Gamma(X, Z)$ and then, by Theorem 3.2, $B \in \partial^\sigma(A \circ F)(\bar{x})$. □

We continue with some specializations of Theorem 4.2.

COROLLARY 4.3. *Let Z be locally convex, $G \in \Gamma(Y, Z)$ be (Y_+, Z_+)-nondecreasing σ-regular subdifferentiable at $A(\bar{x})$, $\sigma \in \{p, w\}$, Z_+ be closed, pointed as $\sigma = p$. Then,*

$$\begin{aligned} \partial^\sigma(G \circ A)(\bar{x}) &= \partial^s G(A(\bar{x})) \circ A + Z_\sigma(X, Z) && \forall A \in L(X, Y), \\ &= \partial^\sigma G(A(\bar{x})) \circ A && \forall A \in L^{-1}(X, Y), \end{aligned}$$

respectively, if

(i) *X, Y are locally convex and G is finite and continuous at some point of* $\operatorname{Im} A$*, or X, Y are Fréchet spaces, G is sequentially Z_+-l.s.c and $\mathbb{R}_+[\operatorname{dom} G - \operatorname{Im} A]$ is a closed vector subspace of Y,*

respectively,

(ii) *X, Y are locally convex and G is finite and continuous at some point of Y, or X, Y are Fréchet spaces and G is sequentially Z_+-l.s.c.*

Proof. The first formula is obtained via Theorem 4.2 and Proposition 3.4. Indeed,

$$\begin{aligned} \partial^\sigma(G \circ A)(\bar{x}) = \bigsqcup_{B \in \partial^s G(A(\bar{x}))} \partial^\sigma(B \circ A)(\bar{x}) &= \bigsqcup_{B \in \partial^s G(A(\bar{x}))} B \circ A + Z_\sigma(X, Z) \\ &= \partial^s G(A(\bar{x})) \circ A + Z_\sigma(X, Z). \end{aligned}$$

The second formula requires the following relation, which may be easily obtained by definition of the σ-zerolike sets using the fact that A is invertible:

$$Z_\sigma(Y, Z) \circ A = Z_\sigma(X, Z).$$

Indeed, because $\operatorname{Im} A = Y$, the first formula still holds under condition (ii), so with Theorem 3.3, $\partial^\sigma(G \circ A)(\bar{x}) = [\partial^s G(A(\bar{x})) + Z_\sigma(Y, Z)] \circ A = \partial^\sigma G(A(\bar{x})) \circ A$. □

The following result may be viewed as a vector extension of the subdifferential regularity in the sense of Pareto.

PROPOSITION 4.1. *Let $F \in \Gamma(X, Y)$ be regular subdifferentiable at $\bar{x}$, $\sigma \in \{p, w\}$ and Z_+ be pointed as $\sigma = p$. Then,*

$$\begin{aligned} \partial^\sigma(A \circ F)(\bar{x}) &= A \circ \partial^s F(\bar{x}) + Z_\sigma(X, Z) && \forall A \in L_+(Y, Z), \\ &= A \circ \partial^\sigma F(\bar{x}) && \forall A \in L_+^{-1}(Y, Z), \end{aligned}$$

where the last equality demands Y_+ to be pointed as $\sigma = p$.

Proof. We already pointed out that $F \in \Gamma(X,Y)$ and $A \in L_+(Y,Z)$ imply that $A \circ F \in \Gamma(X,Z)$. On the other hand, let $\mu \in Z_+^\sigma$, then $\lambda = \mu \circ A \in Y^*$ and $\lambda(Y_+) = \mu(A(Y_+)) \subseteq \mu(Z_+) \subseteq \mathbb{R}_+$, i.e., $\lambda \in Y_+^*$. Hence, by Theorem 3.2 and the regular subdifferentiability assumption, we have that

$$\begin{aligned} B \in \partial^\sigma(A \circ F)(\bar{x}) &\Leftrightarrow B \in L(X,Z), \exists \mu \in Z_+^\sigma : \mu \circ B \in \partial(\mu \circ A \circ F)(\bar{x}) = \lambda \circ \partial^s F(\bar{x}) \\ &\Leftrightarrow B \in L(X,Z), \exists C \in \partial^s F(\bar{x}), \exists \mu \in Z_+^\sigma : \mu \circ (B - A \circ C) = 0 \\ &\Leftrightarrow \exists C \in \partial^s F(\bar{x}) : B - A \circ C \in Z_\sigma(X,Z) \\ &\Leftrightarrow B \in A \circ \partial^s F(\bar{x}) + Z_\sigma(X,Z). \end{aligned}$$

If, furthermore, A is invertible, then we easily have like for Corollary 4.3 that

$$A \circ Z_\sigma(X,Y) = Z_\sigma(X,Z).$$

The proof ends by applying Theorem 3.3. □

Remark 5. We already justified in Remark 4 that σ-subdifferential regularity for $\sigma \in \{p,e,w\}$ cannot generally hold, and this does not contradict Proposition 4.1 since any $A \in L(Y,\mathbb{R})$ cannot be invertible except for the trivial case where $Y = \mathbb{R}$ too.

In finite-dimensional image space, according to Remark 3, Theorem 4.2 becomes the following.

COROLLARY 4.4. *Let X, Y be locally convex, $F \in \Gamma(X,Y)$, $g_i \in \Gamma(Y)$ $(i = 1 \cdots r)$ be Y_+-nondecreasing and consider the following hypotheses:*

- (MR) *$(r-1)$ g_is are finite and continuous at a point of domain of the other one.*
- (AB) *The g_is l.s.c, $\mathbb{R}_+[\Delta Y^r - \prod_{i=1}^r \operatorname{dom} g_i]$ closed vector subspace of Y^r.*
- $(MR)_2$ *All the g_is are finite and continuous at some point of $\operatorname{Im} F$.*
- $(AB)_2$ *The g_is l.s.c, F sequentially Y_+-l.s.c, $\mathbb{R}_+[\prod_{i=1}^r \operatorname{dom} g_i - \operatorname{Im} F]$ closed vector subspace.*

If one of conditions $(MR)_2$, $\big[(AB)_2$ and (MR), with X, Y Fréchet$\big]$ or $\big[(AB)_2$ and (AB), with X, Y Fréchet$\big]$ is satisfied, then, $\forall \bar{x} \in X$,

$$\partial^p(g_1 \circ F, \ldots, g_r \circ F)(\bar{x}) = \bigsqcup_{\substack{\lambda_i \in \partial g_i(F(\bar{x})) \\ i=1\cdots r}} \partial^p(\lambda_1 \circ F, \ldots, \lambda_r \circ F)(\bar{x}).$$

If, in addition to condition $(MR)_2$ or $\big[(AB)_2$ and (MR), with X, Y Fréchet$\big]$, every one or none of the g_is is subdifferentiable at $F(\bar{x})$, then

$$\partial^w(g_1 \circ F, \ldots, g_r \circ F)(\bar{x}) = \bigsqcup_{\substack{\lambda_i \in \partial g_i(F(\bar{x})) \\ i=1\cdots r}} \partial^w(\lambda_1 \circ F, \ldots, \lambda_r \circ F)(\bar{x}).$$

4.3. Addition/composition. In what follows, we establish a formula concerning the σ-subdifferential of the sum of a vector mapping with a vector composite.

THEOREM 4.3. *Let $F : X \to Y \sqcup \{+\infty\}$, $G : X \to Z \sqcup \{+\infty\}$, $H : Z \to Y \sqcup \{+\infty\}$, and $\sigma \in \{p,w\}$. Then, $\forall \bar{x} \in X$,*

$$\partial^\sigma(F + H \circ G)(\bar{x}) \supseteq \bigsqcup_{A \in \partial^s H(G(\bar{x}))} \partial^\sigma(F + A \circ G)(\bar{x}).$$

Assume now that H is (Z_+, Y_+)-nondecreasing and σ-regular subdifferentiable at $G(\bar{x})$, Y_+ is pointed as $\sigma = p$, and one of the two following qualification conditions is satisfied:

$$(MR)_3 \quad \begin{cases} F \in \Gamma(X,Y),\ G \in \Gamma(X,Z),\ H \in \Gamma(Z,Y),\ X \text{ and } Z \text{ locally convex}, \\ H \text{ is finite and continuous at some point of } G(\operatorname{dom} F \sqcap \operatorname{dom} G). \end{cases}$$

$$(AB)_3 \quad \begin{cases} F \in \Gamma_0(X,Y),\ G \in \Gamma_0(X,Z),\ H \in \Gamma_0(Z,Y),\ X \text{ and } Z \text{ Fréchet spaces}, \\ \mathbb{R}_+[\operatorname{dom} H - G(\operatorname{dom} F \sqcap \operatorname{dom} G)] \text{ is a closed vector subspace of } Z. \end{cases}$$

Then,

$$\partial^\sigma(F + H \circ G)(\bar{x}) = \bigsqcup_{A \in \partial^s H(G(\bar{x}))} \partial^\sigma(F + A \circ G)(\bar{x}).$$

The proof is obtained identically to Theorem 4.2 using, instead of Lemma 4.1, the following result due to Combari, Laghdir, and Thibault [11].

LEMMA 4.2. *Let $f : X \to \mathbb{R} \sqcup \{+\infty\}$, $G : X \to Z \sqcup \{+\infty\}$, and $h : Z \to \mathbb{R} \sqcup \{+\infty\}$ be a functional Z_+-nondecreasing such that one of the two following conditions holds:*

$$(MR)_3 \quad \begin{cases} f \in \Gamma(X),\ G \in \Gamma(X,Z),\ h \in \Gamma(Z),\ X \text{ and } Z \text{ locally convex}, \\ h \text{ is finite and continuous at some point of } G(\operatorname{dom} f \sqcap \operatorname{dom} G). \end{cases}$$

$$(AB)_3 \quad \begin{cases} f \in \Gamma_0(X),\ G \in \Gamma_0(X,Z),\ h \in \Gamma_0(Z),\ X \text{ and } Z \text{ Fréchet spaces}, \\ \mathbb{R}_+[\operatorname{dom} h - G(\operatorname{dom} f \sqcap \operatorname{dom} G)] \text{ is a closed vector subspace of } Z. \end{cases}$$

Then, $\forall \bar{x} \in X$,

$$\partial(f + h \circ G)(\bar{x}) = \bigsqcup_{\mu \in \partial h(G(\bar{x}))} \partial(f + \mu \circ G)(\bar{x}).$$

Though popular, the (MR)- and (AB)-type conditions are by their simplicity and feasibility, there exist other weaker interior point conditions ensuring the previous formula, as the following Azé- [4] type condition recently established in [13]:

$$\begin{cases} f \in \Gamma(X),\ G \in \Gamma(X,Z),\ h \in \Gamma(Z),\ X \text{ and } Z \text{ locally convex}, \\ \text{for each neighborhood } V \text{ of } 0 \text{ in } X,\ \text{there exists } \alpha \in \mathbb{R} \text{ such that} \\ 0 \in \operatorname{int}_E[\operatorname{Lev}(h;\alpha) - G(\operatorname{Lev}(f;\alpha) \sqcap \operatorname{dom} G \sqcap \alpha V)], \end{cases}$$

where int_E stands for the topological interior in $E := \operatorname{Vect}[\operatorname{dom} h - G(\operatorname{dom} f \sqcap \operatorname{dom} G)]$, $\operatorname{Vect}(S)$ means vector space generated by a set S, and the symbol Lev refers to $\mathbb{R}_+$-Lev. Weaker again in a sense, another class of qualification conditions, called of closedness-type, is introduced recently by Boţ, Grad, and Wanka [9, 10]. Main among them is

$$\begin{cases} f \in \Gamma_0(X),\ h \in \Gamma_0(Z),\ X \text{ and } Z \text{ separated locally convex}, \\ G \in \Gamma(X,Z) \text{ star } Z_+\text{-l.s.c},\ Z_+ \text{ closed},\ [G(\operatorname{dom} f) + Z_+] \sqcap \operatorname{dom} h \neq \emptyset, \\ \bigsqcup_{\mu \in \operatorname{dom} h^*}(0, h^*(\mu)) + \operatorname{Epi}(f + \mu \circ G)^* \ \text{ is closed}, \end{cases}$$

where h^* stands for the Fenchel conjugate of h (idem for $(f + \mu \circ G)^*$), Epi refers to $\mathbb{R}_+$-Epi, and star Z_+-l.s.c means l.s.c of $\mu \circ G$ for each $\mu \in Z_+^*$. Star and sequential Z_+-l.s.c was relaxed in [10] and [12], respectively, to Z_+-Epi-closedness (Z_+-Epi G closed),

in another variant of the above condition and in the (AB) condition of Lemma 4.2, respectively. The interior point conditions turn out to be globally easily extensible to vector maps, while the closedness ones seem well to require a separate study.

We end this section by the following corollary.

COROLLARY 4.5. *Assume, in addition to the assumptions of Theorem* 4.3*, that F or G is finite and continuous at a point of the effective domain of the other one.*

(1) *If F is σ-regular subdifferentiable at $\bar{x}$, then*

$$\partial^\sigma(F + H \circ G)(\bar{x}) = \partial^s F(\bar{x}) + \bigsqcup_{A \in \partial^s H(G(\bar{x}))} \partial^\sigma(A \circ G)(\bar{x}).$$

(2) *If G is regular subdifferentiable at $\bar{x}$, then*

$$\partial^\sigma(F + H \circ G)(\bar{x}) = \partial^\sigma F(\bar{x}) + \bigsqcup_{A \in \partial^s H(G(\bar{x}))} \partial^s(A \circ G)(\bar{x}).$$

Proof. According to Theorem 4.3, it remains to apply the sum calculus rule for $\partial^\sigma(F + A \circ G)(\bar{x})$ knowing that $A \in L_+(Z, Y)$ and $A \circ G \in \Gamma(X, Y)$ (see the proof of Theorem 4.2). Indeed, Theorem 4.1 directly applies for $F + A \circ G$, taking into account for assertion (2) the following property.

LEMMA 4.3. *With G regular subdifferentiable at $\bar{x}$ and $A \in L_+(Z, Y)$, $A \circ G$ is also regular subdifferentiable at $\bar{x}$.*

Proof of Lemma 4.3. By the regular subdifferentiability of G at $\bar{x}$, we have that $\partial(\lambda \circ A \circ G)(\bar{x}) = \lambda \circ A \circ \partial^s G(\bar{x})$ for any $\lambda \in Y_+^*$ and $A \in L_+(Z, Y)$ since $\lambda \circ A \in Z_+^*$. Also, we easily prove that $A \circ \partial^s G(\bar{x}) \subseteq \partial^s(A \circ G)(\bar{x})$. So $\partial(\lambda \circ A \circ G)(\bar{x}) \subseteq \lambda \circ \partial^s(A \circ G)(\bar{x})$. The reverse inclusion being trivial, the lemma is therefore proved. □

5. Applications to constrained VOPs.

5.1. Vector σ-efficiency conditions for convex VOPs. One main objective in this subsection is to realize the extension of the indicator function technique to the vector case in a way similar to the scalar case, so that necessary and sufficient efficiency conditions under vector forms for a constrained convex VOP may easily be derived from the subdifferential formulas previously obtained. First, the constrained problem

$$\text{VOP:} \qquad \operatorname*{Min}_{x \in S} F(x)$$

may effectively be penalized by the vector indicator function defined for the nonempty set $S \subseteq X$ by

$$\begin{aligned} \delta_S^v : X &\to Y \sqcup \{+\infty\} \\ x &\mapsto \delta_S^v(x) = \begin{cases} 0 & \text{if } x \in S, \\ +\infty & \text{else,} \end{cases} \end{aligned}$$

so that it becomes equivalent to the unconstrained vector problem

$$\operatorname*{Min}_{x \in X} F(x) + \delta_S^v(x)$$

in the following senses.

LEMMA 5.1. *Let $F : X \to Y \sqcup \{+\infty\}$. Then, for $\sigma \in \{s, p, e, w\}$,*
(1) $E_\sigma(F, S) = E_\sigma(F + \delta^v_S, X)$.
(2) $\mathrm{MIN}_\sigma F(S) = \mathrm{MIN}_\sigma(F + \delta^v_S)(X)$.

Proof. Case $\sigma = s$ is easily proved like the scalar case and cases $\sigma \in \{e, p\}$ like $\sigma = w$. Thus, it is enough to show the lemma for $\sigma = w$.

(1) Let $\bar{x} \in E_w(F, S)$, then by definition, $\bar{x} \in S \sqcap \mathrm{dom}\, F = \mathrm{dom}(F + \delta^v_S)$. If $\bar{x} \notin E_w(F + \delta^v_S, X)$, then it would exist $x \in X$ such that $F(x) + \delta^v_S(x) <_{Y_+} F(\bar{x})$, which would imply $x \in S \sqcap \mathrm{dom}\, F$ contradicting $\bar{x} \in E_w(F, S)$. Conversely, let $\bar{x} \in E_w(F + \delta^v_S, X)$, then $\bar{x} \in S \sqcap \mathrm{dom}\, F$. If we suppose that $\bar{x} \notin E_w(F, S)$, then it would exist $x \in S$ such that $F(x) <_{Y_+} F(\bar{x})$, that is, $F(x) + \delta^v_S(x) <_{Y_+} F(\bar{x}) + \delta^v_S(\bar{x})$ contradicting $\bar{x} \in E_w(F + \delta^v_S, X)$. Assertion (1) is therefore proved.

(2) By definition, $\mathrm{MIN}_\sigma F(S) = F(E_\sigma(F, S))$. As $E_w(F, S) \subseteq S$, with (1), it follows that $F(E_w(F, S)) = (F + \delta^v_S)(E_w(F, S)) = (F + \delta^v_S)(E_w(F + \delta^v_S, X))$. ∎

The vector indicator function appears to possess properties like the scalar one. Define for that the normal cone at $\bar{x} \in S$ in a vector sense:

$$N^v_S(\bar{x}) = \{A \in L(X, Y) :\ \forall x \in S,\ A(x - \bar{x}) \leq_{Y_+} 0\}.$$

LEMMA 5.2. *$\delta^v_S \in \Gamma_0(X, Y)$ if S is convex closed. Furthermore, $\forall \bar{x} \in S$,*

$$\partial^s \delta^v_S(\bar{x}) = N^v_S(\bar{x}).$$

In particular, $\delta^v_{-Z_+} \in \Gamma_0(Z, Y)$ if Z_+ is closed and is always (Z_+, Y_+)-nondecreasing. Moreover, if Y_+ is pointed, then, $\forall \bar{z} \in -Z_+$,

$$\partial^s \delta^v_{-Z_+}(\bar{z}) = \{A \in L_+(Z, Y) :\ A(\bar{z}) = 0\}.$$

Proof. The s-subdifferential expression is immediate from the very definitions.

The properness of δ^v_S is immediate, since $\mathrm{dom}\, \delta^v_S = S \neq \emptyset$. Its Y_+-convexity follows easily from the convexity of S and Y_+, more precisely, from its epigraph given by

$$Y_+\text{-Epi}\ \delta^v_S = \{(x, y) \in X \times Y :\ \delta^v_S(x) \leq_{Y_+} y\} = S \times Y_+.$$

While comes the sequential Y_+-l.s.c from the closedness of S. Indeed, let $\bar{x} \in X$, $y \leq_{Y_+} \delta^v_S(\bar{x})$, and $(x^n) \to \bar{x}$ be given. Does $(y^n) \to y$ exist such that $y^n \leq_{Y_+} \delta^v_S(x^n)$ $(\forall n)$? Two cases are distinguished: $\bar{x} \in S$ and $\bar{x} \notin S$. If $\bar{x} \in S$, then it suffices to take $y^n = y$ for all n. If $\bar{x} \in X \setminus S$ which is an open set, then $(x^n) \subset X \setminus S$ for all n large enough. We therefore can take $y^n = y$ for all the n large enough and $= 0$ for the other n. This proves well the sequential Y_+-l.s.c of δ^v_S.

Let us show now that $\delta^v_{-Z_+}$ is (Z_+, Y_+)-nondecreasing. Being given $z_1 \leq_{Z_+} z_2$, do we have $\delta^v_{-Z_+}(z_1) \leq_{Y_+} \delta^v_{-Z_+}(z_2)$? This amounts to showing that if $z_2 \in -Z_+$, then $z_1 \in -Z_+$. But, $z_1 = z_1 - z_2 + z_2 \in -Z_+ - Z_+ \subseteq -Z_+$.

Let us express now its s-subdifferential. One already has $\partial^s \delta^v_{-Z_+}(\bar{z}) = N^v_{-Z_+}(\bar{z})$. By definition, $A \in N^v_{-Z_+}(\bar{z})$ is equivalent to $A \in L(Z, Y)$ and $A(z - \bar{z}) \leq_{Y_+} 0\ \forall\ z \in -Z_+$. Because $\bar{z} \in -Z_+$, by taking successively $z = 0$ and $z = 2\bar{z}$, we get $A(\bar{z}) = 0$. Consequently, $A(z) \geq_{Y_+} 0\ \forall\ z \in Z_+$, i.e., $A \in L_+(Z, Y)$. The reverse inclusion is immediate, since $A(z - \bar{z}) = A(z) \leq_{Y_+} 0\ \forall\ z \in -Z_+$. ∎

The vector indicator function reveals mostly to satisfy the following important property.

LEMMA 5.3. *δ_S^v is regular (resp. p-regular) subdifferentiable on S if* $\operatorname{int} Y_+ \neq \emptyset$ *(resp. if $Y_+^p \neq \emptyset$). In fact, for $\sigma = w$ (resp. $\sigma = p$), $\forall \bar{x} \in S$,*

$$N_S(\bar{x}) = \lambda \circ N_S^v(\bar{x}) \qquad \forall \lambda \in Y_+^\sigma,$$

where $N_S(\bar{x}) = \{x^* \in X : \forall x \in S, \langle x^*, x - \bar{x}\rangle \leq 0\}$ is the usual normal cone to S at $\bar{x}$.

Proof. Let us show first the relation between the vector normal cone and the ordinary one. Let $A \in N_S^v(\bar{x})$ and $\lambda \in Y_+^\sigma$, then $\langle \lambda \circ A, x - \bar{x}\rangle = \langle \lambda, A(x - \bar{x})\rangle \leq 0$ $(\forall x \in S)$. Thus, $\lambda \circ N_S^v(\bar{x}) \subseteq N_S(\bar{x})$. Conversely, let $x^* \in N_S(\bar{x})$ and $\lambda \in Y_+^\sigma$, then we define, exactly as in (3.7), $A_\lambda : X \to Y$ by $A_\lambda(x) = \langle x^*, x\rangle\ y_\lambda$ for some $y_\lambda \in \operatorname{int} Y_+$ if $\sigma = w$ and $y_\lambda \in Y_+ \setminus l(Y_+)$ if $\sigma = p$ so that $A_\lambda \in L(X, Y)$ and $\lambda \circ A_\lambda = x^*$. On the other hand, $A_\lambda(x - \bar{x}) = \langle x^*, x - \bar{x}\rangle y_\lambda \in -Y_+$ $(\forall x \in S)$, i.e., $A_\lambda \in N_S^v(\bar{x})$. Consequently, $N_S(\bar{x}) \subseteq \lambda \circ N_S^v(\bar{x})$.

Let us prove now the regular (resp. p-regular) subdifferentiability of δ_S^v over S. Indeed, by considering the scalar indicator function $\delta_S : X \to \mathbb{R} \sqcup \{+\infty\}$ defined by $\delta_S(x) = 0$ if $x \in S$ and $+\infty$ otherwise, we observe that $\lambda \circ \delta_S^v = \delta_S\ \forall\ \lambda \in Y_+^* \setminus \{0\}$. Hence, for all $\bar{x} \in S$ and $\lambda \in Y_+^\sigma$,

$$\partial(\lambda \circ \delta_S^v)(\bar{x}) = \partial\delta_S(\bar{x}) = N_S(\bar{x}) = \lambda \circ N_S^v(\bar{x}) = \lambda \circ \partial^s \delta_S^v(\bar{x}).$$

This proves the σ-regular subdifferentiability following the hypotheses depending on $\sigma \in \{p, w\}$. But δ_S^v being s-subdifferentiable on S, the above equality remains valid for $\lambda = 0$. This means (taking $\sigma = w$) that δ_S^v is regular subdifferentiable on S. ∎

We are now ready to state the vector σ-efficiency conditions in terms of the Lagrange–Kuhn–Tucker (operator) multiplier and the (vector) normal cone for the following general convex vector mathematical programming problem:

$$\text{VOP:} \qquad \text{Min } F(x) \quad \begin{cases} G(x) \in -Z_+, \\ x \in C, \end{cases}$$

where feasible set $S = \{x \in X : G(x) \in -Z_+\} \sqcap C$, with C a closed convex set.

The Lagrangian map associated to this VOP may be defined as vector function $\mathcal{L} : X \times L_+(Z, Y) \to Y \sqcup \{+\infty\}$ given by

$$\mathcal{L}(x, \Lambda) = F(x) + \Lambda \circ G(x).$$

THEOREM 5.1. *Let $F : X \to Y \sqcup \{+\infty\}$ and $G : X \to Z \sqcup \{+\infty\}$ be such that one of the two following qualification conditions is satisfied:*

$$(MR)_4 \quad \begin{cases} F \in \Gamma(X, Y),\ G \in \Gamma(X, Z),\ X \text{ and } Z \text{ locally convex}, \\ \delta^v_{-Z_+} \text{ is finite and continuous at a point of } G(C \sqcap \operatorname{dom} F \sqcap \operatorname{dom} G). \end{cases}$$

$$(AB)_4 \quad \begin{cases} F \in \Gamma_0(X, Y),\ G \in \Gamma_0(X, Z),\ Z_+ \text{ closed},\ X \text{ and } Z \text{ Fréchet spaces}, \\ \mathbb{R}_+[Z_+ + G(C \sqcap \operatorname{dom} F \sqcap \operatorname{dom} G)] \text{ is a closed vector subspace of } Z. \end{cases}$$

Then, with Y_+ pointed, for $\sigma \in \{p, w\}$,

$$\bar{x} \in E_\sigma(F, S) \iff \begin{cases} G(\bar{x}) \in -Z_+, \\ \exists \bar{\Lambda} \in L_+(Z, Y) :\ \bar{\Lambda}(G(\bar{x})) = 0, \\ 0 \in \partial^\sigma(F + \bar{\Lambda} \circ G + \delta_C^v)(\bar{x}),\ \textit{i.e.,}\ \bar{x} \in E_\sigma(\mathcal{L}(., \bar{\Lambda}), C). \end{cases}$$

Proof. Taking into account that $\delta^v_S = \delta^v_{-Z_+} \circ G + \delta^v_C$, by Lemma 5.1 and relation (2.2), we can write that

$$\bar{x} \text{ is } \sigma\text{-efficient for VOP} \iff 0 \in \partial^\sigma(F + \delta^v_{-Z_+} \circ G + \delta^v_C)(\bar{x}).$$

Next we easily verify that all the hypotheses of Theorem 4.3 are satisfied with $F + \delta^v_C$, G, and $\delta^v_{-Z_+}$ underlining that $F + \delta^v_C$ is well sequentially Y_+-l.s.c as a sum of such maps (see [11]). With Lemmas 5.2 and 5.3 and that $\bar{x} \in S \sqcap \operatorname{dom} F$, we thus obtain that

$$\partial^\sigma(F + \delta^v_C + \delta^v_{-Z_+} \circ G)(\bar{x}) = \bigsqcup_{\substack{\bar{\Lambda} \in L_+(Z,Y) \\ \bar{\Lambda}(G(\bar{x}))=0}} \partial^\sigma(F + \delta^v_C + \bar{\Lambda} \circ G)(\bar{x}).$$

Significance $\bar{x} \in E_\sigma(\mathcal{L}(.,\bar{\Lambda}), C)$ is a consequence of (2.2) and Lemma 5.1. ∎

Remark 6. The condition in $(MR)_4$ is not other than the Slater-type condition "$\exists a \in C \sqcap \operatorname{dom} F \sqcap \operatorname{dom} G : G(a) \in -\operatorname{int} Z_+$," which, in turn, is a particular case of the condition in $(AB)_4$. Indeed, it easily implies that $\mathbb{R}_+[Z_+ + G(C \sqcap \operatorname{dom} F \sqcap \operatorname{dom} G)] = Z$ (see [11]). However, it is, in most cases, easier to check when $\operatorname{int} Z_+ \neq \emptyset$. Theorem 5.1 has been obtained under the Slater condition by several authors using different methods; see, e.g., [17] for the (Benson) proper concept and [1, 27] for the weak one.

COROLLARY 5.1. *If in addition to the assumptions of Theorem* 5.1*, F and G are finite and continuous at some point of C, F (and resp. G) is σ-regular (resp. regular) subdifferentiable at $\bar{x}$ and $\sigma \in \{p, w\}$, then*

$$\bar{x} \in E_\sigma(F,S) \iff \begin{cases} G(\bar{x}) \in -Z_+, \\ \exists \bar{\Lambda} \in L_+(Z,Y) : \ \bar{\Lambda}(G(\bar{x})) = 0, \\ 0 \in \partial^\sigma F(\bar{x}) + \bar{\Lambda} \circ \partial^s G(\bar{x}) + N^v_C(\bar{x}). \end{cases}$$

Proof. It suffices to apply to Theorem 5.1, Theorem 4.1 two times successively followed by Proposition 4.1, and finally Theorem 3.3 and Lemma 5.2 to obtain

$$\begin{aligned} 0 \in \partial^\sigma(F + \bar{\Lambda} \circ G + \delta^v_C)(\bar{x}) &= \partial^\sigma(\bar{\Lambda} \circ G)(\bar{x}) + \partial^s F(\bar{x}) + \partial^s \delta^v_C(\bar{x}) \\ &= \bar{\Lambda} \circ \partial^s G(\bar{x}) + Z_\sigma(X,Y) + \partial^s F(\bar{x}) + \partial^s \delta^v_C(\bar{x}) \\ &= \partial^\sigma F(\bar{x}) + \bar{\Lambda} \circ \partial^s G(\bar{x}) + N^v_C(\bar{x}). \end{aligned}$$ ∎

Remark 7. If space Y (resp. Z) is finite-dimensional endowed with the natural order, then the σ-regular (resp. regular) subdifferentiability assumption on F (resp. on G) in Corollary 5.1 may be omitted, since condition (MR) of Remark 3 is satisfied, and, as $\partial^\sigma(F + \bar{\Lambda} \circ G + \delta^v_C)(\bar{x}) \neq \emptyset$, mappings F and G are sensed to be s-subdifferentiable at $\bar{x}$, i.e., their components are subdifferentiable at this point.

5.2. Gap between weakly or properly and strongly efficient sets. We end up the paper with another application to VOPs. We already presented the gap between the efficient and optimal sets associated to unconstrained VOPs. In what follows, similar results are also derived for constrained VOPs via combinations of several basic results.

THEOREM 5.2. *Let S be convex closed in X locally convex, $F \in \Gamma(X,Y)$ be σ-regular subdifferentiable at $\bar{x}$, $\sigma \in \{p, w\}$, Y_+ be pointed as $\sigma = p$, and one of the three following conditions be satisfied:*

$(MR)_5$ $\left\{ \begin{array}{l} F \text{ is finite and continuous at a point of } S, \text{ or } \operatorname{dom} F \sqcap \operatorname{int} S \neq \emptyset, \text{ or,} \end{array} \right.$

$(AB)_5$ $\left\{ \begin{array}{l} F \text{ sequentially } Y_+\text{-l.s.c}, \ X \text{ Fréchet space}, \\ \mathbb{R}_+[S - \operatorname{dom} F] \quad \text{is a closed vector subspace of } X. \end{array} \right.$

Then,

$$\bar{x} \in E_\sigma(F,S) \quad \Longleftrightarrow \quad \exists A \sim_\sigma 0 \ : \ \bar{x} \in E_s(F-A,S).$$

Proof. Let us show the nontrivial direct implication. According to Lemma 5.1, Theorem 4.1, Theorem 3.3, and finally, relation (3.5), we have that

$$\begin{aligned}\bar{x} \in E_\sigma(F,S) = E_\sigma(F+\delta_S^v,X) &\Leftrightarrow 0 \in \partial^\sigma(F+\delta_S^v)(\bar{x}) = \partial^s F(\bar{x}) + Z_\sigma(X,Y) + \partial^s \delta_S^v(\bar{x}) \\ &\Leftrightarrow \exists A \sim_\sigma 0 : \ A \in \partial^s F(\bar{x}) + \partial^s \delta_S^v(\bar{x}) \subseteq \partial^s(F+\delta_S^v)(\bar{x}) \\ &\Rightarrow \exists A \sim_\sigma 0 : \ \bar{x} \in E_s(F-A+\delta_S^v,X) = E_s(F-A,S).\end{aligned}$$

The reverse implication follows straightforwardly by the definition of the s-efficient set and the σ-zerolike operator one, and, by scalarization, Theorem 3.1. ∎

Remark 8. For the same reasons as in Remark 7, if in Theorem 5.2, the range space is finite-dimensional endowed with the natural order and the first condition in $(MR)_5$ is fulfilled, then the σ-regular subdifferentiability assumption on F may be omitted. Since, for $\sigma = p$, F becomes p-regular subdifferentiable everywhere in its domain (see Remark 3), the result of this theorem then may rather be written in this case as

$$E_p(F,S) = \bigsqcup_{A \sim_p 0} E_s(F-A,S).$$

Conclusion. All the well-known results, in scalar case, coincide with our results. The unique additional assumption, namely σ-regular subdifferentiability, obviously holds in this case: $\partial(\alpha f)(x) = \alpha \partial f(x) \ \forall \ \alpha \in \mathbb{R}_+^\sigma = \mathbb{R}_+ \setminus \{0\} = \operatorname{int} \mathbb{R}_+$.

Further research avenues may interest some other subdifferential operations such as maximum of functions, marginal function, infimal convolution, etc., and the applications that may engender. Also, a particular attention to the Pareto concept ($\sigma = e$) and/or the strong ordinary notion ($\sigma = s$), which certainly will contribute more in multicriteria optimization, may be the subject of future investigations.

Acknowledgments. The authors are grateful to the reviewers for their valuable reports and their judicious remarks and suggestions.

REFERENCES

[1] M. Adán and V. Novo, *Optimality conditions for vector optimization problems with generalized convexity in real linear spaces*, Optimization, 51 (2002), pp. 73–91.
[2] V. Alexéev, V. Tikhomirov, and S. Fomine, *Commande Optimale*, MIR, Moscou, 1982.
[3] H. Attouch and H. Brézis, *Duality for the sum of convex functions in general Banach spaces*, in Aspects of Mathematics and its Applications, J. Barroso, ed., Elsevier, Amsterdam, 1986, pp. 125–133.
[4] D. Azé, *Duality for the sum of convex functions in general normed spaces*, Arch. Math., 62 (1994), pp. 554–561.
[5] D. Azé, *Eléments d'Analyse Convexe et Variationnelle*, Ellipses, Paris, 1997.
[6] S. Bolintinéanu and M. El Maghri, *Pénalisation dans l'optimisation sur l'ensemble faiblement efficient*, RAIRO Oper. Res., 31 (1997), pp. 295–310.
[7] S. Bolintinéanu (H. Bonnel), *Vector variational principles; ε-efficiency and scalar stationarity*, J. Convex Anal., 8 (2001), pp. 71–85.
[8] H. Bonnel, A. Iusem, and B. Svaiter, *Proximal methods in vector optimization*, SIAM J. Optim., 15 (2005), pp. 953–970, (improved version of S. Bolintinéanu, *Cours de Maîtrise*, Université de Perpignan, Perpignan, France, 1995).
[9] R. I. Boţ, S.-M. Grad, and G. Wanka, *Generalized Moreau–Rockafellar Results for Composed Convex Functions*, Preprint 16, Chemnitz University of Technology, Faculty of Mathematics, Chemnitz, Germany, 2007.

[10] R. I. Boţ, S.-M. Grad, and G. Wanka, *A new constraint qualification for the formula of the subdifferential of composed convex functions in infinite dimensional spaces*, Math. Nachr., 281 (2008), pp. 1065–1068.
[11] C. Combari, M. Laghdir, and L. Thibault, *Sous-différentiels de fonctions convexes composées*, Ann. Sci. Math. Québec, 18 (1994), pp. 119–148.
[12] C. Combari, M. Laghdir, and L. Thibault, *A note on subdifferentials of convex composite functionals*, Arch. Math., 67 (1996), pp. 239–252.
[13] C. Combari, M. Laghdir, and L. Thibault, *On subdifferential calculus for convex functions defined on locally convex spaces*, Ann. Sci. Math. Québec, 23 (1999), pp. 23–26.
[14] M. El Maghri and B. Bernoussi, *Pareto optimizing and Kuhn–Tucker stationary sequences*, Numer. Funct. Anal. Optim., 28 (2007), pp. 287–305.
[15] M. Laghdir, *Some remarks on subdifferentiability of convex functions*, Appl. Math. E-Notes, 5 (2005), pp. 150–156.
[16] B. Lemaire, *Application of a Subdifferential of a Convex Composite Functional to Optimal Control in Variational Inequalities*, Lecture Notes Econom. Math. Systems 255, Springer-Verlag, New York, 1985, pp. 103–117.
[17] Z. F. Li, *Benson proper efficiency in the vector optimization of set-valued maps*, J. Optim. Theory Appl., 98 (1998), pp. 623–649.
[18] L. J. Lin, *Optimization of set-valued functions*, J. Math. Anal. Appl., 186 (1994), pp. 30–51.
[19] D. T. Luc, *Theory of Vector Optimization*, Lecture Notes Econom. Math. Systems 319, Springer-Verlag, Berlin, 1989.
[20] J. J. Moreau, *Fonctionnelles convexes*, Séminaire "Equations aux Dérivées Partielles", Collège de France, Paris, 1966.
[21] J. P. Penot and M. Théra, *Semicontinuous mappings in general topology*, Arch. Math., 38 (1982), pp. 158–166.
[22] A. L. Peressini, *Ordered Topological Vector Spaces*, Harper, New York, 1967.
[23] C. Raffin, *Contribution à l'étude des programmes convexes définis dans des espaces vectoriels topologiques*, Thèse, University of Paris, Paris, 1969.
[24] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[25] Y. Sawaragi and T. Tanino, *Conjugate maps and duality in multiobjective optimization*, J. Optim. Theory Appl., 31 (1980), pp. 473–499.
[26] Y. Sawaragi, H. Nakayama, and T. Tanino, *Theory of Multiobjective Optimization*, Academic Press, New York, 1985.
[27] A. Taa, *Subdifferentials of multifunctions and Lagrange multipliers for multiobjective optimization*, J. Math. Anal. Appl., 283 (2003), pp. 398–415.
[28] Ch. Tammer and K. Tammer, *Duality results for convex vector optimization problems with linear restrictions*, in System Modelling and Optimization, P. Kall, ed., Springer-Verlag, Berlin, 1992, pp. 55–64.
[29] M. Théra, *Subdifferential calculus for convex operators*, J. Math. Anal. Appl., 80 (1981), pp. 78–91.
[30] J. Thierfelder, *Nonvertical affine manifolds and separation theorems in product spaces*, Math. Nachr., 151 (1991), pp. 329–344.
[31] J. Thierfelder, *Separation theorems for sets in product spaces*, Kybernetika, 27 (1991), pp. 522–534.
[32] M. Valadier, *Sous-différentiabilité de fonctions convexes à valeurs dans un espace vectoriel ordonné*, Math. Scand., 30 (1972), pp. 65–74.
[33] C. Zălinescu, *Convex Analysis in General Vector Spaces*, World Scientific, Singapore, 2002.
[34] C. Zălinescu, *Hahn–Banach extension theorems for multifunctions revisited*, Math. Methods Oper. Res., 68 (2008), pp. 493–508.
[35] J. Zowe, *Subdifferentiability of convex functions with values in an ordered vector space*, Math. Scand., 34 (1974), pp. 69–83.

SIAM J. OPTIM.
Vol. 19, No. 4, pp. 1995–2014

CONVEXITY IN SEMIALGEBRAIC GEOMETRY AND POLYNOMIAL OPTIMIZATION*

JEAN B. LASSERRE†

Abstract. We review several (and provide new) results on the theory of moments, sums of squares, and basic semialgebraic sets when convexity is present. In particular, we show that, under convexity, the hierarchy of semidefinite relaxations for polynomial optimization simplifies and has finite convergence, a highly desirable feature as convex problems are in principle easier to solve. In addition, if a basic semialgebraic set $\mathbf{K}$ is convex but its defining polynomials are not, we provide two algebraic *certificates* of convexity which can be checked numerically. The second is simpler and holds if a sufficient (and almost necessary) condition is satisfied; it also provides a new condition for $\mathbf{K}$ to have semidefinite representation. For this we use (and extend) some of the recent results from the author and Helton and Nie [*Math. Program.*, to appear]. Finally, we show that, when restricting to a certain class of convex polynomials, the celebrated Jensen's inequality in convex analysis can be extended to linear functionals that are not necessarily probability measures.

Key words. convex polynomials, sums of squares, basic semialgebraic sets, convex sets, Jensen inequality, semidefinite programming

AMS subject classifications. Primary, 14P10, 90C22; Secondary, 11E25, 12D15, 90C25

DOI. 10.1137/080728214

1. Introduction.

Motivation. This paper is a contribution to the new emerging field of convex semialgebraic geometry, and its purpose is threefold: First, we show that the moment approach for global polynomial optimization proposed in [13], and based on semidefinite programming (SDP), is consistent as it simplifies and/or has better convergence properties when solving convex problems. In other words, the SDP moment approach somehow "recognizes" convexity, a highly desirable feature for a general purpose method because, in principle, convex problems should be easier to solve.

We next review some recent results (and provide a new one) on the representation of convex basic semialgebraic sets by linear matrix inequalities which show how convexity permits to derive relatively simple and *explicit* semidefinite representations. In doing so, we also provide a *certificate* of convexity for $\mathbf{K}$ when its defining polynomials are not convex.

Finally, we consider the important Jensen's inequality in convex analysis. When restricting its application to a class of convex polynomials, we provide an extension to a class of linear functionals that are not necessarily probability measures.

To do so, we use (and sometimes extend) some recent results of the author [16, 17] and Helton and Nie [6]. We hope to convince the reader that convex semialgebraic geometry is, indeed, a very specific subarea of real algebraic geometry which should deserve more attention from both the optimization and real algebraic geometry research communities.

*Received by the editors June 23, 2008; accepted for publication (in revised form) December 3, 2008; published electronically March 13, 2009. This research was supported by the (French) ANR under grant NT05-3-41612.

http://www.siam.org/journals/siopt/19-4/72821.html

†LAAS-CNRS and Institute of Mathematics, University of Toulouse, 7 Avenue du Colonel Roche, 31077 Toulouse cédex 4, France (lasserre@laas.fr).

Background. I. Relatively recent results in the theory of moments and its dual theory of positive polynomials have been proved useful in polynomial optimization as they provide the basis of a specific convergent numerical approximation scheme. Namely, one can define a hierarchy of semidefinite relaxations (in short SDP relaxations) of the original optimization problem whose associated monotone sequence of optimal values converges to the global optimum. For a more detailed account of this approach, the interested reader is referred to, e.g., Lasserre [13, 14], Parrilo [21], Schweighofer [29], and the many references therein.

Remarkably, practice seems to reveal that convergence is often fast and even finite. However, the size of the SDP relaxations grows rapidly with the rank in the hierarchy; typically the rth SDP relaxation in the hierarchy has $O(n^{2r})$ variables and semidefinite matrices of $O(n^r)$ sizes (where n is the number of variables in the original problem). On the other hand, it is well-known that a large class of convex optimization problems can be solved efficiently; see, e.g., Ben-Tal and Nemirovski [1]. Therefore, as the SDP-based moment approach is dedicated to solving difficult nonconvex (most of the time NP-hard) problems, it should have the highly desirable feature to somehow *recognize* "easy" problems like convex ones. That is, when applied to such easy problems, it should show some significant improvement or a particular nice behavior not necessarily valid in the general case. Notice that this is *not* the case of the moment approach based on linear programming (LP) described in [14, 15] for which only asymptotic (and *not* finite) convergence occurs in general (and especially for convex problems), a rather annoying feature. However, for SDP relaxations, some results of [17] already show that, indeed, convexity helps as one provides specialized representation results for convex polynomials that are nonnegative on a basic semialgebraic set.

II. Next, in view of the potential of semidefinite programming techniques, an important issue is the characterization of convex sets that are semidefinite representable (in short called SDr). An SDr set $\mathbf{K} \subset \mathbb{R}^n$ is the projection of a set defined by linear matrix inequalities. That is,

$$\mathbf{K} := \left\{ x \in \mathbb{R}^n \,:\, \exists\, y \in \mathbb{R}^s \text{ such that (s.t.)} \quad A_0 + \sum_{i=1}^{n} x_i\, A_i + \sum_{j=1}^{s} y_j\, B_j \succeq 0 \right\}$$

for some real symmetric matrices (A_i, B_j) (and where $A \succeq 0$ stands for A is positive semidefinite). For more details, the interested reader is referred to Ben Tal and Nemirovski [1], Lewis, Parrilo, and Ramana [19], and Parrilo [22] and more recently, Chua and Tuncel [2], Helton and Nie [6, 7], Henrion [8], and Lasserre [16]. For compact basic semialgebraic sets

$$\mathbf{K} := \{x \in \mathbb{R}^n \,:\, g_j(x) \geq 0, \quad j = 1, \ldots, m\}, \tag{1.1}$$

recent results of Helton and Nie [6, 7] and the author [16] provide sufficient conditions on the defining polynomials $(g_j) \subset \mathbb{R}[X]$ for the convex hull $\mathrm{co}\,(\mathbf{K})$ ($\equiv \mathbf{K}$ if $\mathbf{K}$ is convex) to be SDr. Again, an interesting issue is to analyze whether convexity of $\mathbf{K}$ (with or without concavity of the defining polynomials (g_j)) provides some additional insights and/or simplifications. Another interesting issue is how to detect whether a basic semialgebraic set $\mathbf{K}$ is convex or, equivalently, how to obtain an algebraic *certificate* of convexity of $\mathbf{K}$ from its defining polynomials (g_j). By certificate we mean a mathematical statement that obviously implies convexity of $\mathbf{K}$, can be checked numerically, and does not require infinitely many tests. So far, and to the best of our knowledge, such a certificate does not exist.

III. The celebrated Jensen's inequality is an important result in convex analysis which states that $E_\mu(f(x)) \geq f(E_\mu(x))$ for a convex function $f : \mathbb{R}^n \to \mathbb{R}$ and a probability measure μ with $E_\mu(x) < \infty$. A third goal of this paper is to analyze whether, when restricted to a certain class of convex polynomials, Jensen's inequality can be extended to a class of linear functionals larger than the class of probability measures.

Contribution. Concerning issue I, we first recall two previous results proved in [17]: (a) the cone of convex sums of squares (SOS) is dense (for the l_1-norm of coefficients) in the cone of nonnegative convex polynomials, and (b) a convex positivstellensatz holds for convex polynomials nonnegative on $\mathbf{K}$ (a specialization of Putinar's positivstellensatz). We then analyze the role of convexity for the polynomial optimization problem

$$\mathbf{P}: \quad f^* = \min_x \{ f(x) : x \in \mathbf{K} \}, \tag{1.2}$$

with $\mathbf{K}$ as in (1.1), and show that, indeed, convexity helps and makes the SDP relaxations more efficient. In particular, when $\mathbf{K}$ is convex and Slater's condition[1] holds, by using some recent results of Helton and Nie [6], we show the following.

(i) If the polynomials f and $(-g_j)$ are all convex and $\nabla^2 f$ is positive definite (and so f is strictly convex) on $\mathbf{K}$, then the hierarchy of SDP relaxations has *finite* convergence.

(ii) If f and $(-g_j)$ are all SOS-convex (i.e., their Hessian is an SOS matrix polynomial), then $\mathbf{P}$ reduces to solving a *single* SDP whose index in the hierarchy is readily available.

Concerning II, under certain sufficient conditions on the (g_j) (typically some second-order positive curvature conditions) Helton and Nie [6, 7] have proved that $\mathrm{co}\,(\mathbf{K})$ (or $\mathbf{K}$ if convex) has a semidefinite representation that uses the Schmüdgen or Putinar SOS representation of polynomials positive on $\mathbf{K}$; see [6, 17]. Yet, in general its dimension depends on an unknown degree parameter in the Schmüdgen (or Putinar) SOS representation. Our contribution is to provide a new sufficient condition for the existence of a semidefinite representation when $\mathbf{K}$ is compact with nonempty interior and its boundary satisfies some nondegeneracy assumption. It translates the geometric property of convexity of $\mathbf{K}$ into an SOS Putinar representation of some appropriate polynomial obtained from each g_j. When satisfied, this representation provides an algebraic certificate of convexity for $\mathbf{K}$ and it is almost necessary in the sense that it always holds true when relaxed by an arbitrary $\epsilon > 0$. It also contains as special cases Helton and Nie [6] sufficient conditions of SOS convexity or strict convexity on $\partial\mathbf{K}$ of the $-g_j$'s and leads to an explicit semidefinite representation of $\mathbf{K}$. We also provide a more general algebraic certificate based on Stengle's positivstellensatz, but more complex and heavy to implement and so not very practical. In practice, both certificates are obtained by solving a semidefinite program. Therefore, because of unavoidable numerical inaccuracies, the certificate is valid only up to machine precision.

Concerning III, we prove that, when restricting its application to the subclass of SOS-convex polynomials, Jensen's inequality can be extended to all linear functionals $L_{\mathbf{y}}$ (with $L_{\mathbf{y}}(1) = 1$) in the dual cone of SOS polynomials and, hence, *not* necessarily probability measures.

Some of the results already obtained in [6, 16] and in the present paper strongly suggest that the class of SOS-convex polynomials introduced in Helton and Nie [6] is particularly nice and should deserve more attention.

[1] Slater's condition holds for $\mathbf{K}$ in (1.1) if for some $x_0 \in \mathbf{K}$, $g_j(x_0) > 0$, $j = 1, \ldots, m$.

2. Notation, definitions, and preliminary results. Let $\mathbb{R}[X]$ be the ring of real polynomials in the variables $X = (X_1, \ldots, X_n)$, and let $\Sigma^2[X] \subset \mathbb{R}[X]$ be the subset of SOS polynomials. Let $\mathbb{R}[X]_d \subset \mathbb{R}[X]$ be the set of polynomials of degree at most d, which forms a vector space of dimension $s(d) = \binom{n+d}{d}$. If $f \in \mathbb{R}[X]_d$, write $f(X) = \sum_{\alpha \in \mathbb{N}^n} f_\alpha X^\alpha$ in the usual canonical basis (X^α), and denote by $\mathbf{f} = (f_\alpha) \in \mathbb{R}^{s(d)}$ its vector of coefficients. Also, write $\|f\|_1 \,(= \|\mathbf{f}\|_1 := \sum_\alpha |f_\alpha|)$ the l_1-norm of f. Finally, denote by $\Sigma^2[X]_d \subset \Sigma^2[X]$ the subset of SOS polynomials of degree at most $2d$.

We use the notation X for the variable of a polynomial $X \mapsto f(X)$ and x when x is a point of $\mathbb{R}^n$ as, for instance, in $\{x \in \mathbb{R}^n \,:\, f(x) \geq 0\}$.

Moment matrix. With $\mathbf{y} = (y_\alpha)$ being a sequence indexed in the canonical basis (X^α) of $\mathbb{R}[X]$, let $L_{\mathbf{y}} : \mathbb{R}[X] \to \mathbb{R}$ be the linear functional

$$f \quad \left(= \sum_\alpha f_\alpha \, X^\alpha \right) \quad \mapsto \quad L_{\mathbf{y}}(f) \,=\, \sum_\alpha f_\alpha \, y_\alpha,$$

and let $M_d(\mathbf{y})$ be the symmetric matrix with rows and columns indexed in the canonical basis (X^α) and defined by

$$M_d(\mathbf{y})(\alpha, \beta) \,:=\, L_{\mathbf{y}}\left(X^{\alpha+\beta}\right) \,=\, y_{\alpha+\beta}, \quad \alpha, \beta \in \mathbb{N}^n_d,$$

with $\mathbb{N}^n_d := \{\alpha \in \mathbb{N}^n \,:\, |\alpha| \,(= \sum_i \alpha_i) \leq d\}$.

Localizing matrix. Similarly, with $\mathbf{y} = (y_\alpha)$ and $g \in \mathbb{R}[X]$ written

$$X \mapsto g(X) \,=\, \sum_{\gamma \in \mathbb{N}^n} g_\gamma \, X^\gamma,$$

let $M_d(g\,\mathbf{y})$ be the symmetric matrix with rows and columns indexed in the canonical basis (X^α) and defined by

$$M_d(g\,\mathbf{y})(\alpha, \beta) \,:=\, L_{\mathbf{y}}\left(g(X)\, X^{\alpha+\beta}\right) \,=\, \sum_\gamma g_\gamma \, y_{\alpha+\beta+\gamma}$$

for every $\alpha, \beta \in \mathbb{N}^n_d$.

Putinar positivstellensatz. Let $Q(g) \subset \mathbb{R}[X]$ be the quadratic module generated by the polynomials $(g_j) \subset \mathbb{R}[X]$, that is,

$$Q(g) \,:=\, \left\{ \sigma_0 + \sum_{j=1}^m \sigma_j \, g_j \,:\quad (\sigma_j) \subset \Sigma^2[X] \right\}. \tag{2.1}$$

ASSUMPTION 2.1. *$\mathbf{K} \subset \mathbb{R}^n$ is a compact basic semialgebraic set defined as in* (1.1)*, and the quadratic polynomial $X \mapsto M - \|X\|^2$ belongs to $Q(g)$.*

Assumption 2.1 is not very restrictive. For instance, it holds if every g_j is affine (i.e., $\mathbf{K}$ is a convex polytope) or if the level set $\{x \,:\, g_j(x) \geq 0\}$ is compact for some $j \in \{1, \ldots, m\}$. In addition, if $M - \|x\| \geq 0$ for all $x \in \mathbf{K}$, then it suffices to add the redundant quadratic constraint $M^2 - \|x\|^2 \geq 0$ to the definition (1.1) of $\mathbf{K}$ and Assumption 2.1 will hold true.

THEOREM 2.2 (Putinar's positivstellensatz [24]). *Let Assumption* 2.1 *hold. If* $f \in \mathbb{R}[X]$ *is (strictly) positive on* $\mathbf{K}$*, then* $f \in Q(g)$*. That is,*

$$f = \sigma_0 + \sum_{j=1}^{m} \sigma_j\, g_j \tag{2.2}$$

for some SOS polynomials $(\sigma_j) \subset \Sigma^2[X]$.

2.1. A hierarchy of semidefinite relaxations (SDP relaxations). Let $\mathbf{P}$ be the optimization problem (1.2) with $\mathbf{K}$ as in (1.1), and let $r_j = \lceil (\deg g_j)/2 \rceil$, $j = 1, \ldots, m$. With $f \in \mathbb{R}[X]$ and $2r \geq \max[\deg f,\ \max_j 2r_j]$, consider the hierarchy of semidefinite relaxations $(\mathbf{Q}_r)$ defined by

$$\mathbf{Q}_r : \quad \begin{cases} \inf\limits_{\mathbf{y}} & L_{\mathbf{y}}(f) \\ \text{s.t.} & M_r(\mathbf{y}) \succeq 0, \\ & M_{r-r_j}(g_j\, \mathbf{y}) \succeq 0, \qquad j = 1, \ldots, m, \\ & y_0 = 1, \end{cases} \tag{2.3}$$

with optimal value denoted by $\inf \mathbf{Q}_r$. One says that $\mathbf{Q}_r$ is solvable if it has an optimal solution (in which case one writes $\inf \mathbf{Q}_r = \min \mathbf{Q}_r$). The dual of $\mathbf{Q}_r$ reads

$$\mathbf{Q}_r^* : \quad \begin{cases} \sup & \lambda \\ \text{s.t.} & f - \lambda = \sigma_0 + \sum\limits_{j=1}^{m} \sigma_j\, g_j, \\ & \sigma_j \in \Sigma^2[X], \quad j = 0, 1, \ldots, m, \\ & \deg \sigma_0,\ \deg \sigma_j + \deg g_j \leq 2r, \quad j = 1, \ldots, m, \end{cases} \tag{2.4}$$

with optimal value denoted by $\sup \mathbf{Q}_r^*$ (or $\max \mathbf{Q}_r^*$ if the sup is attained).

By weak duality $\sup \mathbf{Q}_r^* \leq \inf \mathbf{Q}_r$ for every $r \in \mathbb{N}$ and under Assumption 2.1, $\inf \mathbf{Q}_r \uparrow f^*$ as $r \to \infty$. For a more detailed account see, e.g., [13].

2.2. Convexity and SOS convexity. We first briefly recall basic facts on a multivariate convex function. If $C \subseteq \mathbb{R}^n$ is a nonempty convex set, a function $f : C \to \mathbb{R}$ is convex on C if and only if

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \qquad \forall\, \lambda \in (0,1),\ x, y \in C.$$

Similarly, f is strictly convex on C if and only if the above inequality is strict for every $x, y \in C$, $x \neq y$, and all $\lambda \in (0,1)$.

If $C \subseteq \mathbb{R}^n$ is an open convex set and f is twice differentiable on C, then f is convex on C if and only if its Hessian $\nabla^2 f$ is positive semidefinite on C (denoted $\nabla^2 f \succeq 0$ on C). Finally, if $\nabla^2 f$ is positive definite on C (denoted $\nabla^2 f \succ 0$ on C), then f is strictly convex on C.

SOS convexity. Helton and Nie [6] have introduced the following interesting subclass of convex polynomials, called SOS-convex polynomials.

DEFINITION 2.3 (Helton and Nie [6]). *A polynomial* $f \in \mathbb{R}[X]_{2d}$ *is said to be SOS-convex if* $\nabla^2 f$ *is SOS; that is,* $\nabla^2 f = LL^T$ *for some real matrix polynomial* $L \in \mathbb{R}[X]^{n \times s}$ *(for some* $s \in \mathbb{N}$*).*

As noted in [6], an important feature of SOS convexity is that it can be checked numerically by solving an SDP. They have also proved the following important property.

LEMMA 2.4 (Helton and Nie [6, Lemma 7]). *If a symmetric matrix polynomial $P \in \mathbb{R}[X]^{r\times r}$ is SOS, then, for any $u \in \mathbb{R}^n$, the double integral*

$$X \mapsto \quad F(X,u) := \int_0^1 \int_0^t P(u + s(X-u))\, ds\, dt$$

is also a symmetric SOS matrix polynomial in $\mathbb{R}[X]^{r\times r}$.

See also the following.

LEMMA 2.5 (Helton and Nie [6, Lemma 8]). *For a polynomial $f \in \mathbb{R}[X]$ and every $x, u \in \mathbb{R}^n$,*

$$f(x) = f(u) + \nabla f(u)^T(x-u) + (x-u)^T \underbrace{\int_0^1 \int_0^t \nabla^2 f(u + s(x-u))dsdt}_{F(x,u)} (x-u).$$

So if f is SOS-convex and $f(u) = 0, \nabla f(u) = 0$, then f is an SOS polynomial.

2.3. An extension of Jensen's inequality. Recall that if μ is a probability measure on $\mathbb{R}^n$ with $E_\mu(x) < \infty$, Jensen's inequality states that if $f \in L_1(\mu)$ and f is convex, then

$$E_\mu(f(x)) \;\geq\; f(E_\mu(x)),$$

a very useful property in many applications.

We now provide an extension of Jensen's inequality when one restricts its application to the class of SOS-convex polynomials. Namely, we may consider the linear functionals $L_{\mathbf{y}} : \mathbb{R}[X]_{2d} \to \mathbb{R}$ in the dual cone of $\Sigma^2[X]_d$, that is, vectors $\mathbf{y} = (y_\alpha)$ such that $M_d(\mathbf{y}) \succeq 0$ and $y_0 = L_{\mathbf{y}}(1) = 1$; hence, $\mathbf{y}$ is *not* necessarily the (truncated) moment sequence of some probability measure μ. Crucial in the proof is Lemma 2.4 of Helton and Nie [6].

THEOREM 2.6. *Let $f \in \mathbb{R}[X]_{2d}$ be SOS-convex, and let $\mathbf{y} = (y_\alpha)_{\alpha\in\mathbb{N}^n_{2d}}$ satisfy $y_0 = 1$ and $M_d(\mathbf{y}) \succeq 0$. Then*

$$L_{\mathbf{y}}(f(X)) \;\geq\; f(L_{\mathbf{y}}(X)), \tag{2.5}$$

where $L_{\mathbf{y}}(X) = (L_{\mathbf{y}}(X_1), \ldots, L_{\mathbf{y}}(X_n))$.

Proof. Let $z \in \mathbb{R}^n$ be fixed and arbitrary, and consider the polynomial $X \mapsto f(X) - f(z)$. Then

$$f(X) - f(z) \;=\; \langle \nabla f(z), X - z\rangle + \langle (X-z), F(X)(X-z)\rangle, \tag{2.6}$$

with $F : \mathbb{R}^n \to \mathbb{R}[X]^{n\times n}$ being the matrix polynomial

$$X \mapsto \quad F(X) := \int_0^1 \int_0^t \nabla^2 f(z + s(X-z))\, ds\, dt.$$

As f is SOS-convex, by Lemma 2.4, F is an SOS matrix polynomial and so the polynomial $X \mapsto \Delta(X) := \langle (X-z), F(X)(X-z)\rangle$ is SOS, i.e., $\Delta \in \Sigma^2[X]$. Then applying $L_{\mathbf{y}}$ to the polynomial $X \mapsto f(X) - f(z)$ and using (2.6) yields (recall that $y_0 = 1$)

$$\begin{aligned} L_{\mathbf{y}}(f(X)) - f(z) &= \langle \nabla f(z), L_{\mathbf{y}}(X) - z\rangle + L_{\mathbf{y}}(\Delta(X)) \\ &\geq \langle \nabla f(z), L_{\mathbf{y}}(X) - z\rangle \quad [\text{because } L_{\mathbf{y}}(\Delta(X)) \geq 0]. \end{aligned}$$

As $z \in \mathbb{R}^n$ was arbitrary, taking $z := L_{\mathbf{y}}(X)\,(= (L_{\mathbf{y}}(X_1), \ldots, L_{\mathbf{y}}(X_n)))$ yields the desired result. □

As a consequence we also get the following corollary.

COROLLARY 2.7. *Let f be a convex univariate polynomial, $g \in \mathbb{R}[X]$ (and so $f \circ g \in \mathbb{R}[X]$). Let $d := \lceil (\deg f \circ g)/2 \rceil$, and let $\mathbf{y} = (y_\alpha)_{\alpha \in \mathbb{N}^n_{2d}}$ be such that $y_0 = 1$ and $M_d(\mathbf{y}) \succeq 0$. Then*

$$L_{\mathbf{y}}[\, f(g(X))\,] \;\geq\; f(L_{\mathbf{y}}[\, g(X)\,]). \tag{2.7}$$

Proof. Again, let $z \in \mathbb{R}$ be fixed and arbitrary, and consider the univariate polynomial $Y \mapsto f(Y) - f(z)$ so that (2.6) holds. That is,

$$f(Y) - f(z) \;=\; f'(z)\,(Y - z) + F(Y)(Y - z)^2,$$

with $F : \mathbb{R} \to \mathbb{R}[Y]$ being the univariate polynomial

$$Y \mapsto \quad F(Y) \;:=\; \int_0^1 \int_0^t f''(z + s(Y - z))\, ds\, dt.$$

As f is convex, $f'' \geq 0$, and so the univariate polynomial $Y \mapsto F(Y)(Y - z)^2$ is nonnegative and, being univariate, is SOS. Therefore, with $Y := g(X)$,

$$f(g(X)) - f(z) \;=\; f'(z)\,(g(X) - z) + F(g(X))(g(X) - z)^2,$$

and so

$$\begin{aligned} L_{\mathbf{y}}[\, f(g(X))] - f(z) &= f'(z)\,(L_{\mathbf{y}}[\, g(X)\,] - z) + L_{\mathbf{y}}\left[\, F(g(X))\,(g(X) - z)^2\,\right] \\ &\geq f'(z)(L_{\mathbf{y}}[\, g(X)\,] - z), \end{aligned}$$

and taking $z := L_{\mathbf{y}}[g(X)]$ yields the desired result. □

Hence, the class of SOS-convex polynomials has the very interesting property of extending Jensen's inequality to some linear functionals that are not necessarily coming from a probability measure.

3. Semidefinite relaxations in the convex case.

3.1. A convex positivstellensatz. Let $\mathbf{K}$ be as in (1.1), and define $Q_c(g) \subset \mathbb{R}[X]$ to be the set

$$Q_c(g) \;:=\; \left\{ \sigma_0 + \sum_{j=1}^m \lambda_j\, g_j \;:\quad \lambda \in \mathbb{R}^m_+ ;\; \sigma_0 \in \Sigma^2[X],\; \sigma_0 \text{ convex} \right\} \subset Q(g). \tag{3.1}$$

The set $Q_c(g)$ is a specialization of $Q(g)$ in (2.1) to the convex case, in that the weights asociated with the g_j's are nonnegative scalars, i.e., SOS polynomials of degree 0, and the SOS polynomial σ_0 is convex. In particular, every $f \in Q_c(g)$ is nonnegative on $\mathbf{K}$. Let $\mathcal{F}_{\mathbf{K}} \subset \mathbb{R}[X]$ be the convex cone of convex polynomials nonnegative on $\mathbf{K}$.

THEOREM 3.1 (Lasserre [17]). *Let $\mathbf{K}$ be as in* (1.1), *Slater's condition hold, and g_j be concave for every $j = 1, \ldots, m$.*

Then, with $Q_c(g)$ as in (3.1), *the set $Q_c(g) \cap \mathcal{F}_{\mathbf{K}}$ is dense in $\mathcal{F}_{\mathbf{K}}$ for the l_1-norm $\|\cdot\|_1$. In particular, if $\mathbf{K} = \mathbb{R}^n$ (so that $\mathcal{F}_{\mathbb{R}^n} =: \mathcal{F}$ is now the set of nonnegative convex polynomials), then $\Sigma^2[X] \cap \mathcal{F}$ is dense in $\mathcal{F}$.*

Theorem 3.1 states that if f is convex and nonnegative on $\mathbf{K}$ (including the case $\mathbf{K} \equiv \mathbb{R}^n$), then one may approximate f by a sequence $\{f_{\epsilon r}\} \subset Q_c(g) \cap \mathcal{F}_{\mathbf{K}}$ with $\|f - f_{\epsilon r}\|_1 \to 0$ as $\epsilon \to 0$ (and $r \to \infty$). For instance, with $r_0 := \lfloor (\deg f)/2 \rfloor + 1$,

$$
\begin{aligned}
&X \mapsto f_{\epsilon r}(X) := f + \epsilon(\theta_{r_0}(X) + \theta_r(X)), \quad \text{with} \\
&X \mapsto \theta_r(X) := 1 + \sum_{k=1}^{r} \sum_{i=1}^{n} \frac{X_i^{2k}}{k!}, \qquad r \geq r_\epsilon,
\end{aligned} \tag{3.2}
$$

for some r_ϵ; see Lasserre [17] for details. Observe that Theorem 3.1 provides f with a *certificate* of nonnegativity on $\mathbf{K}$. Indeed, let $x \in \mathbf{K}$ be fixed arbitrary. Then as $f_{\epsilon r} \in Q_c(g)$ one has $f_{\epsilon r}(x) \geq 0$. Letting $\epsilon \downarrow 0$ yields $0 \leq \lim_{\epsilon \to 0} f_{\epsilon r}(x) = f(x)$. As $x \in \mathbf{K}$ was arbitrary, $f \geq 0$ on $\mathbf{K}$.

Theorem 3.1 is a convex (weak) version of Theorem 2.2 (Putinar's positivstellensatz) where one replaces the quadratic module $Q(g)$ with its subset $Q_c(g)$. We call it a *weak* version of Theorem 2.2 because it invokes a density result (i.e., $f_{\epsilon r} \in Q_c(g)$, whereas f might not be an element of $Q_c(g)$). Notice that f is allowed to be nonnegative (instead of strictly positive) on $\mathbf{K}$ and $\mathbf{K}$ need *not* be compact; recall that extending Theorem 2.2 to noncompact basic semialgebraic sets $\mathbf{K}$ and to polynomials f nonnegative on $\mathbf{K}$ is hopeless in general; see Scheiderer [26].

COROLLARY 3.2. *Let* $\mathbf{K}$ *be as in* (1.1), $f \in \mathbb{R}[X]$ *with* $f^* := \min_x\{f(x) : x \in \mathbf{K}\}$, *and let* $d := \max[\lceil (\deg f)/2 \rceil, \max_j \lceil (\deg g_j)/2 \rceil]$. *Consider the simplified SDP relaxation*

$$
\widehat{\mathbf{Q}}: \quad \begin{cases} \inf\limits_{\mathbf{y}} & L_{\mathbf{y}}(f) \\ s.t. & M_d(\mathbf{y}) \succeq 0, \\ & L_{\mathbf{y}}(g_j) \geq 0, \qquad j = 1, \ldots, m, \\ & y_0 = 1 \end{cases} \tag{3.3}
$$

and its dual

$$
\widehat{\mathbf{Q}}^*: \quad \begin{cases} \sup\limits_{\gamma, \sigma_0, \lambda} & \gamma \\ s.t. & f - \gamma = \sigma_0 + \sum\limits_{j=1}^{m} \lambda_j g_j, \\ & \sigma_0 \in \Sigma^2[X]_d;\ \lambda_j \geq 0, \quad j = 1, \ldots, m. \end{cases} \tag{3.4}
$$

(a) *If* $f - f^* \in Q_c(g)$, *then the SDP relaxation* $\widehat{\mathbf{Q}}$ *and its dual* $\widehat{\mathbf{Q}}^*$ *are exact.*

(b) *If* $f, -g_j \in \mathbb{R}[X]$ *are convex,* $j = 1, \ldots, m$, *and if* $\mathbf{y}$ *is an optimal solution of* $\widehat{\mathbf{Q}}$ *which satisfies*

$$
\operatorname{rank} M_d(\mathbf{y}) = \operatorname{rank} M_{d-1}(\mathbf{y}), \tag{3.5}
$$

then $\widehat{\mathbf{Q}}$ *is exact and* $x^* := (L_{\mathbf{y}}(X_i)) \in \mathbf{K}$ *is a (global) minimizer of* f *on* $\mathbf{K}$.

Proof. (a) If $f - f^* \in Q_c(g)$, i.e., if $f - f^* = \sigma_0 + \sum_{j=1}^m \lambda_j g_j$, with $\sigma_0 \in \Sigma^2[X]_d$ and $\lambda \in \mathbb{R}^m_+$, the triplet (f^*, σ_0, λ) is a feasible solution of $\widehat{\mathbf{Q}}^*$ with value f^*. Therefore, as $\sup \widehat{\mathbf{Q}}^* \leq \inf \widehat{\mathbf{Q}} \leq f^*$, the SDP relaxation $\widehat{\mathbf{Q}}$ and its dual $\widehat{\mathbf{Q}}^*$ are exact. In fact, (f^*, σ_0, λ) is an optimal solution of $\widehat{\mathbf{Q}}^*$.

(b) If $\mathbf{y}$ satisfies the rank condition (3.5), then by the *flat extension* theorem of Curto and Fialkow [4], $\mathbf{y}$ is the (truncated) moment sequence of an atomic probability measure μ on $\mathbb{R}^n$, say, $\mu = \sum_{k=1}^s \lambda_k \delta_{x(k)}$ with $s = \operatorname{rank} M_d(\mathbf{y})$, $0 < \lambda_k \leq 1$,

$\sum_k \lambda_k = 1$, and $\delta_{x(k)}$ being the Dirac measure at $x(k) \in \mathbb{R}^n$, $k = 1, \ldots, s$. Let $x^* := \sum_k \lambda_k x(k) = (L_{\mathbf{y}}(X_i)) \in \mathbb{R}^n$. Then $f^* \geq L_{\mathbf{y}}(f)$ and by convexity of f, $L_{\mathbf{y}}(f) = \sum_k \lambda_k f(x(k)) \geq f(\sum_k \lambda_k x(k)) = f(x^*)$. Similarly, by convexity of $-g_j$, $0 \leq L_{\mathbf{y}}(g_j) = \sum_k \lambda_k g_j(x(k)) \leq g_j(\sum_k \lambda_k x(k)) = g_j(x^*)$, $j = 1, \ldots, m$. Therefore, $x^* \in \mathbf{K}$ and as $f(x^*) \leq f^*$, x^* is a global minimizer of f on $\mathbf{K}$. □

Notice that $\mathbf{K}$ in Corollary 3.2 need not be compact. Also, Corollary 3.2(b) has practical value because in general one does not know whether $f - f^* \in Q_c(g)$ (despite that, in the convex case, $f - f^* \in \mathcal{F}_{\mathbf{K}}$ and $Q_c(g) \cap \mathcal{F}_{\mathbf{K}}$ is dense in $\mathcal{F}_{\mathbf{K}}$). However, one may still solve $\widehat{\mathbf{Q}}$ and check whether the rank condition (3.5) is satisfied. If, in solving $\widehat{\mathbf{Q}}_r$, the rank condition (3.5) is not satisfied, then other sufficient conditions can be exploited as we next see.

3.2. The SOS-convex case. Part (a) of the following result is already contained in Lasserre [17, Corollary 2.5].

THEOREM 3.3. *Let $\mathbf{K}$ be as in* (1.1) *and Slater's condition hold. Let $f \in \mathbb{R}[X]$ be such that $f^* := \inf_x\{f(x) : x \in \mathbf{K}\} = f(x^*)$ for some $x^* \in \mathbf{K}$. If f is SOS-convex and $-g_j$ is SOS-convex for every $j = 1, \ldots, m$, then*

(a) $f - f^* \in Q_c(g)$;

(b) *the simplified SDP relaxation $\widehat{\mathbf{Q}}$ in* (3.3) *and its dual* (3.4) *are exact and solvable. If $\mathbf{y}$ is an optimal solution of $\widehat{\mathbf{Q}}$, then $x^* := (L_{\mathbf{y}}(X_i)) \in \mathbf{K}$ is a global minimizer of f on $\mathbf{K}$.*

Proof. (a) is proved in [17, Corollary 2.5]. (b) That $\widehat{\mathbf{Q}}$ is exact follows from (a) and Corollary 3.2(a). Hence, it is solvable (e.g., take $\mathbf{y}$ to be the moment sequence associated with the Dirac measure at a global minimizer $x^* \in \mathbf{K}$). So let $\mathbf{y}$ be an optimal solution of $\widehat{\mathbf{Q}}$, hence, with $f^* = L_{\mathbf{y}}(f)$. As $-g_j$ is SOS-convex for every j, then by Theorem 2.6, $0 \leq L_{\mathbf{y}}(g_j) \leq g_j(x^*)$ with $x^* := (L_{\mathbf{y}}(X_i))$, and so $x^* \in \mathbf{K}$. Similarly, as f is SOS-convex, we also have $f^* = L_{\mathbf{y}}(f) \geq f(x^*)$ which proves that $f(x^*) = f^*$ and x^* is a global minimizer of f on $\mathbf{K}$. Finally, as by (a) $f - f^* \in Q_c(g)$, then $\widehat{\mathbf{Q}}^*$ is exact and solvable. □

(Again notice that $\mathbf{K}$ in Theorem 3.3 need not be compact.) So the class of SOS-convex polynomials is particularly interesting. Not only Jensen's inequality can be extended to some linear functionals that are not coming from a probability measure but also one may also solve SOS-convex optimization problems $\mathbf{P}$ in (1.2) (i.e., with f and $\mathbf{K}$ defined with SOS-convex polynomials) by solving the single semidefinite program (3.3).

Notice that a self-concordant[2] logarithmic barrier function exists for (3.3), whereas the logarithmic barrier function with barrier parameter μ,

$$x \mapsto \phi_\mu(x) := \mu\, f(x) - \sum_{j=1}^{m} \ln\,(g_j(x)), \tag{3.6}$$

associated with $\mathbf{P}$, is not self-concordant in general. Therefore, despite the fact that (3.3) involves additional variables (a lifting), solving (3.3) via an interior point method might be more efficient than solving $\mathbf{P}$ by using the logarithmic barrier function (3.6) with no lifting. In addition, all SOS-convex polynomials nonnegative on $\mathbf{K}$, and which attain their minimum on $\mathbf{K}$ belong to $Q_c(g)$, a very specific version of Putinar positivstellensatz (as f is only nonnegative and $\mathbf{K}$ need not be compact).

[2]The self-concordance property introduced in [20] is fundamental in the design and efficiency of interior point methods for convex programming.

3.3. The strictly convex case. If f or some of the $-g_j$'s are not SOS-convex but $\nabla^2 f \succ 0$ (so that f is strictly convex) and $-g_j$ is convex for every $j = 1, \ldots, m$, then, inspired by a nice argument from Helton and Nie [6] for the existence of a semidefinite representation of convex sets, one obtains the following result.

THEOREM 3.4. *Let* $\mathbf{K}$ *be as in* (1.1)*, and let Assumption* 2.1 *and Slater's condition hold. Assume that* $f, -g_j \in \mathbb{R}[X]$ *are convex,* $j = 1, \ldots, m$*, with* $\nabla^2 f \succ 0$ *on* $\mathbf{K}$.

Then the hierarchy of SDP relaxations defined in (2.3) *has finite convergence. That is,* $f^* = \sup \mathbf{Q}_r^* = \inf \mathbf{Q}_r$ *for some index* r*. In addition,* $\mathbf{Q}_r$ *and* $\mathbf{Q}_r^*$ *are solvable so that* $f^* = \max \mathbf{Q}^* = \min \mathbf{Q}_r$.

Proof. Let $x^* \in \mathbf{K}$ be a global minimizer (i.e., $f^* = f(x^*)$). As Slater's condition holds, there exists a vector of Karush–Kuhn–Tucker (KKT) multipliers $\lambda \in \mathbb{R}^m_+$ such that the (convex) Lagrangian $L_f \in \mathbb{R}[X]$ defined by

$$X \mapsto L_f(X) \,:=\, f(X) - f^* - \sum_{j=1}^m \lambda_j \, g_j(X) \tag{3.7}$$

has a global minimum at $x^* \in \mathbf{K}$, i.e., $\nabla L_f(x^*) = 0$. In addition, $\lambda_j g_j(x^*) = 0$ for every $j = 1, \ldots, m$ and $L_f(x^*) = 0$. Then, by Lemma 2.5,

$$L_f(X) \,=\, \langle (X - x^*), F(X, x^*)(X - x^*) \rangle$$

with

$$F(X, x^*) \,:=\, \left(\int_0^1 \int_0^t \nabla^2 L_f(x^* + s(X - x^*)) \, ds \, dt \right).$$

Next, let I_n be the $n \times n$ identity matrix. As $\nabla^2 f \succ 0$ on $\mathbf{K}$, continuity of the (strictly positive) smallest eigenvalue of $\nabla^2 f$ and compactness of $\mathbf{K}$ yield that $\nabla^2 f \succeq \delta I_n$ on $\mathbf{K}$ for some $\delta > 0$. Next, as $-g_j$ is convex for every j and in view of the definition (3.7) of L_f, $\nabla^2 L_f \succeq \nabla^2 f \succeq \delta I_n$ on $\mathbf{K}$. Hence, for every $\xi \in \mathbb{R}^n$, $\xi^T F(x, x^*) \xi \geq \delta \int_0^1 \int_0^t \xi^T \xi \, ds dt = \frac{\delta}{2} \xi^T \xi$, and so $F(x, x^*) \succeq \frac{\delta}{2} I_n$ for every $x \in \mathbf{K}$. Therefore, by the matrix polynomial version of Putinar positivstellensatz,

$$F(X, x^*) \,=\, F_0(X) + \sum_{j=1}^m F_j(X) \, g_j(X)$$

for some real SOS matrix polynomials $X \mapsto F_j(X) = L_j(X) L_j(X)^T$ (for some appppropriate $L_j \in \mathbb{R}[X]^{n \times p_j}$), $j = 0, \ldots, m$. See Helton and Nie [6], Kojima and Maramatsu [10], and Hol and Scherer [11]. But then

$$X \mapsto \langle (X - x^*), F_j(X, x^*)(X - x^*) \rangle \,=\, \sigma_j(X) \in \Sigma^2[X], \qquad j = 0, \ldots, m,$$

and so

$$\begin{aligned} f(X) - f^* &= L_f(X) + \sum_{j=1}^m \lambda_j g_j(X) \\ &= \sigma_0(X) + \sum_{j=1}^m (\lambda_j + \sigma_j(X)) \, g_j(X). \end{aligned}$$

Let $2s$ be the maximum degree of the SOS polynomials (σ_j). Then $(f^*, \{\sigma_j + \lambda_j\})$ is a feasible solution of the SDP relaxation $\mathbf{Q}_r^*$ in (2.4) with $r := s + \max_j r_j$. Therefore, as

$\sup \mathbf{Q}_r^* \leq \inf \mathbf{Q}_r \leq f^*$, the SDP relaxations $\mathbf{Q}_r$ and $\mathbf{Q}_r^*$ are exact, finite convergence occurs, and $\mathbf{Q}_r^*$ is solvable. But this also implies that $\mathbf{Q}_r$ is solvable (take $\mathbf{y}$ to be the moment sequence of the Dirac measure δ_{x^*} at any global minimizer $x^* \in \mathbf{K}$). □

When compared to Theorem 3.3 for the SOS-convex case, in the strictly convex case the simplified SDP relaxation $\widehat{\mathbf{Q}}$ in (3.3) is not guaranteed to be exact. However, finite convergence still occurs for the SDP relaxations $(\mathbf{Q}_r)$ in (2.3).

Remark 3.5. It is worth emphasizing that in general the hierarchy of LP relaxations (as opposed to SDP relaxations) defined in [15] and based on Krivine's representation [12, 30] and Handelman [**?**] for polynomials positive on $\mathbf{K}$ *cannot* have finite convergence, especially in the convex case! For more details, the interested reader is referred to [14, 15]. Therefore and despite the fact that LP software packages can solve LP problems of very large size, using LP relaxations does not seem a good idea even for solving a convex polynomial optimization problem.

4. Convexity and semidefinite representation of convex sets. We now consider the semidefinite representation of convex sets. First, recall the following result.

THEOREM 4.1 (Lasserre [16]). *Let* $\mathbf{K}$ *in* (1.1) *be compact with* g_j *concave,* $j = 1, \ldots, m$, *and assume that Slater's condition holds. If the Lagrangian polynomial* L_f *in* (3.7) *associated with every linear polynomial* $f \in \mathbb{R}[X]$ *is SOS, then, with* $d := \max_j \lceil (\deg g_j)/2 \rceil$, *the set*

$$\Omega := \left\{ (x, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^{s(2d)} \; : \; \begin{cases} M_d(\mathbf{y}) \succeq 0, \\ L_{\mathbf{y}}(g_j) \geq 0, \quad j = 1, \ldots, m, \\ L_{\mathbf{y}}(X_i) = x_i, \quad i = 1, \ldots, n, \\ y_0 = 1 \end{cases} \right\} \tag{4.1}$$

is a semidefinite representation of $\mathbf{K}$.

Next, Helton and Nie [6, 7] have provided several interesting second-order positive curvature (sufficient and necessary) conditions on the defining polynomials (g_j) for $\mathbf{K}$ (or its convex hull co $(\mathbf{K})$) to be an SDr set. In particular (recall that $r_j = \lceil (\deg g_j)/2 \rceil$ for every $j = 1, \ldots, m$), see the following theorem.

THEOREM 4.2 (Helton and Nie [6]). *Let* $\mathbf{K}$ *in* (1.1) *be convex and Asssumption* 2.1 *hold, and assume that Slater's condition holds and* g_j *is concave on* $\mathbf{K}$, $j = 1, \ldots, m$.

(a) *If* $-g_j$ *is SOS-convex for every* $j = 1, \ldots, m$, *then, for every linear* $f \in \mathbb{R}[X]$, *the associated Lagrangian* L_f (3.7) *is SOS and the set* Ω *in* (4.1) *is a semidefinite representation of* $\mathbf{K}$.

(b) *If every* $-g_j$ *is either SOS-convex or satisfies* $-\nabla^2 g_i \succ 0$ *on* $\mathbf{K} \cap \{x \; : \; g_j(x) = 0\}$, *then there exists* $r \in \mathbb{N}$ *such that the set*

$$\Omega := \left\{ (x, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^{s(2r)} \; : \; \begin{cases} M_r(\mathbf{y}) \succeq 0 \\ M_{r-r_j}(g_j \mathbf{y}) \succeq 0, \quad j = 1, \ldots, m \\ L_{\mathbf{y}}(X_i) = x_i, \quad i = 1, \ldots, n \\ y_0 = 1 \end{cases} \right\} \tag{4.2}$$

is a semidefinite representation of $\mathbf{K}$.

See [6, Theorems 6 and 9]. This follows from the fact that the Hessian $\nabla^2 L_f$ associated with a linear $f \in \mathbb{R}[X]$ has a Putinar representation in terms of SOS matrix polynomials and with degree of the weights bounded uniformly in f. In principle, the degree parameter d in Theorem 4.2(b) may be computed by solving a hierarchy of semidefinite programs. Some other (more technical) weaker second-order positive

curvature sufficient conditions (merely for the existence of a semidefinite representation) are also provided in [6, 7], but the semidefinite representation is not explicit anymore in terms of the defining polynomials (g_j). Notice that if $\mathbf{K}$ is compact but Assumption 2.1 does not hold, then one still obtains a semidefinite representation for $\mathbf{K}$ but more complicated as it is now based on Schmüdgen's representation [27] instead of Putinar's representation; see [6, Theorem 5].

We next provide a sufficient condition in the case where $\mathbf{K}$ is convex but its defining polynomials $(-g_j)$ are *not* necessarily convex. Among its distinguishing features, it is checkable numerically, contains Theorem 4.2 as a special case, and leads to the explicit semidefinite representation (4.2) of $\mathbf{K}$.

4.1. Algebraic certificate of convexity. We first present the following characterization of convexity when $\mathbf{K}$ is closed, satisfies a nondegeneracy assumption on its boundary, and Slater's condition holds.

LEMMA 4.3. *Let $\mathbf{K}$ be as in* (1.1) *(hence, closed), let Slater's condition hold, and assume that for every $j = 1, \ldots, m$, $\nabla g_j(y) \neq 0$ if $y \in \mathbf{K}$ and $g_j(y) = 0$. Then $\mathbf{K}$ is convex if and only if for every $j = 1, \ldots, m$,*

$$\langle \nabla g_j(y), x - y \rangle \geq 0 \qquad \forall x \in \mathbf{K} \text{ and } \forall\, y \in \mathbf{K} \text{ with } g_j(y) = 0. \tag{4.3}$$

Proof. The *only if* part is obvious. Indeed, if $\langle \nabla g_j(y), x - y \rangle < 0$ for some $x \in \mathbf{K}$ and $y \in \mathbf{K}$ with $g_j(y) = 0$, then there is some $\overline{t} > 0$ such that $g_j(y + t(x - y)) < 0$ for all $t \in (0, \overline{t})$, and so the point $x' := tx + (1 - t)y$ does not belong to $\mathbf{K}$, which in turn implies that $\mathbf{K}$ is not convex.

For the *if* part, (4.3) implies that at, every point of the boundary, there exists a supporting hyperplane for $\mathbf{K}$. As $\mathbf{K}$ is closed with nonempty interior, the result follows from [28, Theorem 1.3.3][3]. □

The nondegeneracy assumption is crucial as demonstrated in the following simple example kindly provided by an anonymous referee.

Example 4.1. Consider the nonconvex set $\mathbf{K} \subset \mathbb{R}^2$ defined by

$$\mathbf{K} := \left\{ x \in \mathbb{R}^2 \,:\, \left(1 - x_1^2 + x_2^2\right)^3 \geq 0,\ 10 - x_1^2 - x_2^2 \geq 0 \right\}.$$

Then it is straightforward to see that (4.3) is satisfied. This is because ∇g_1 vanishes on the piece of boundary determined by $g_1(x) = 0$.

Next, using the above characterization (4.3), we provide an algebraic certificate of convexity.

COROLLARY 4.4 (algebraic certificate of convexity). *Let $\mathbf{K}$ be as in* (1.1), *let Slater's condition hold, and assume that for every $j = 1, \ldots, m$, $\nabla g_j(y) \neq 0$ if $y \in \mathbf{K}$ and $g_j(y) = 0$. Then $\mathbf{K}$ is convex if and only if for every $j = 1, \ldots, m$,*

$$h_j(X, Y)\langle \nabla g_j(Y), X - Y \rangle \,=\, \langle \nabla g_j(Y), X - Y \rangle^{2l} + \theta_j(X, Y) + \varphi_j(X, Y) g_j(Y) \tag{4.4}$$

for some integer $l \in \mathbb{N}$, some polynomial $\varphi_j \in \mathbb{R}[X, Y]$, and some polynomials h_j, θ_j in the preordering[4] of $\mathbb{R}[X, Y]$ generated by the family of polynomials $(g_k(X), g_p(Y))$, $k, p \in \{1, \ldots, m\}$, $p \neq j$.

Proof. By Lemma 4.3, $\mathbf{K}$ is convex if and only if, for every $j = 1, \ldots, m$, the polynomial $(X, Y) \mapsto \langle \nabla g_j(Y), X - Y \rangle$ is nonnegative on the set Ω_j defined by

$$\Omega_j \,:=\, \{(x, y) \in \mathbf{K} \times \mathbf{K} \,:\, g_j(y) = 0\}. \tag{4.5}$$

[3] The author is grateful to L. Tuncel for providing us with [28].

[4] The preordering of $\mathbb{R}[X]$ generated by a family $(g_1, \ldots, g_m) \subset \mathbb{R}[X]$ is the set of polynomials $\{p \,:\, p = \sum_{J \subseteq \{1, \ldots, m\}} \sigma_J (\prod_{j \in J} g_j)$, with $\sigma_J \in \Sigma^2[X]\}$.

Equivalently, $\mathbf{K}$ is convex if and only if for every $j = 1, \ldots, m$

$$\emptyset = \{(x,y) \in \mathbb{R}^n : (x,y) \in \mathbf{K} \times \mathbf{K}; \quad g_j(y) = 0; \\ \langle \nabla g_j(y), x - y \rangle \leq 0; \langle \nabla g_j(y), x - y \rangle \neq 0\}.$$

Then (4.4) follows from Stengle's positivstellensatz [25, Theorem 4.4.2, p. 92]. □

Observe that Corollary 4.4 provides an algebraic certificate of convexity when $\mathbf{K}$ is closed with nonempty interior and a nondegeneracy assumption holds on its boundary. If one fixes an a priori bound s on $l \in \mathbb{N}$ and on the degree of h_j, θ_j, and φ_j, then checking whether (4.4) holds reduces to solving a semidefinite program. If $\mathbf{K}$ is convex, by increasing s, eventually one would obtain such a certificate if one could solve semidefinite programs exactly. In practice and because of unavoidable numerical inaccuracies, one obtains only a numerical approximation of the optimal value and so a certificate valid *up to machine precision* only.

However, implementing such a procedure is extremely costly because one has potentially 2×2^m unknown SOS polynomials to define h_j and θ_j in (4.4)! Therefore, it is highly desirable to provide a less costly certificate but with no guarantee to hold for every $\mathbf{K}$ as in Corollary 4.4.

In particular, one considers only compact sets $\mathbf{K}$. Indeed, if $\mathbf{K}$ is compact, one has the following result (recall that $g_0 \equiv 1$).

LEMMA 4.5. *Let $\mathbf{K}$ be convex, and let Assumption 2.1 and Slater's condition hold. Assume that for every $j = 1, \ldots, m$, $\nabla g_j(y) \neq 0$ if $y \in \mathbf{K}$ and $g_j(y) = 0$. Then for every $\epsilon > 0$ and every $j = 1, \ldots, m$*

$$\langle \nabla g_j(Y), X - Y \rangle + \epsilon = \sum_{k=0}^{m} \sigma_{jk}(X,Y)\, g_k(X) + \sum_{k=0, k\neq j}^{m} \psi_{jk}(X,Y)\, g_k(Y) \\ + \psi_j(X,Y)\, g_j(Y) \tag{4.6}$$

for some SOS polynomials (σ_{jk}) and $(\psi_{jk})_{k\neq j} \subset \Sigma^2[X,Y]$ and some polynomial $\psi_j \in \mathbb{R}[X,Y]$.

Proof. By Lemma 4.3, for every $j = 1, \ldots, m$ and every $x, y \in \mathbf{K}$ such that $g_j(y) = 0$, (4.3) holds, and, therefore, for every $j = 1, \ldots, m$,

$$\langle \nabla g_j(y), x - y \rangle + \epsilon > 0 \qquad \forall (x,y) \in \Omega_j, \tag{4.7}$$

where Ω_j has been defined in (4.5). As $\mathbf{K}$ satisfies Assumption 2.1, then so does Ω_j for every $j = 1, \ldots, m$. Hence, (4.6) follows from (4.7) and Theorem 2.2. □

Therefore, inspired by Lemma 4.5, we introduce the following condition.

ASSUMPTION 4.6 (certificate of convexity). *For every $j = 1, \ldots, m$, (4.6) holds with $\epsilon = 0$. Then let $d_j \in \mathbb{N}$ be such that $2d_j$ is larger than the maximum degree of the polynomials $\sigma_{jk} g_k, \psi_{jk} g_k, \psi_j g_j \in \mathbb{R}[X,Y]$ in (4.6), $j = 1, \ldots, m$.*

When $\mathbf{K}$ is closed (and not necessarily compact), Slater's condition holds and the nondegeneracy assumption on the boundary holds (i.e., $\nabla g_j(y) \neq 0$ if $y \in \mathbf{K}$ and $g_j(y) = 0$). Assumption 4.6 is, indeed, a certificate of convexity because then (4.3) holds for every $x, y \in \mathbf{K}$ with $g_j(y) = 0$, and, by Lemma 4.3, $\mathbf{K}$ is convex. It translates the geometric property of convexity of $\mathbf{K}$ into an algebraic SOS Putinar representation of the polynomial $(X,Y) \mapsto \langle \nabla g_j(Y), X - Y \rangle$ nonnegative on Ω_j, $j = 1, \ldots, m$. On the other hand, if $\mathbf{K}$ is convex and Assumption 2.1, Slater's condition, and the nondegeneracy assumption all hold, then Assumption 4.6 is almost necessary as, by Lemma 4.5, (4.6) holds with $\epsilon > 0$ arbitrary.

With d_j fixed a priori, checking whether (4.6) holds with $\epsilon = 0$ can be done numerically. (However, again it provides a certificate of convexity valid *up to machine precision* only.) For instance, for every $j = 1, \ldots, m$, it suffices to solve the semidefinite program (recall that $r_k = \lceil (\deg g_k)/2 \rceil$, $k = 1 \ldots, m$)

$$
\begin{cases}
\rho_j := \min\limits_{\mathbf{z}} \; L_{\mathbf{z}}(\langle \nabla g_j(Y), X - Y \rangle) \\
\text{s.t.} \quad M_{d_j}(\mathbf{z}) \succeq 0, \\
\qquad M_{d_j - r_k}(g_k(X)\,\mathbf{z}) \succeq 0, \quad k = 1, \ldots, m, \\
\qquad M_{d_j - r_k}(g_k(Y)\,\mathbf{z}) \succeq 0, \quad k = 1, \ldots, m;\; k \neq j, \\
\qquad M_{d_j - r_j}(g_j(Y)\,\mathbf{z}) \;=\; 0, \\
\qquad y_0 = 1.
\end{cases} \tag{4.8}
$$

If $\rho_j = 0$ for every $j = 1, \ldots, m$, then Assumption 4.6 holds. This is in contrast to the PP-BDR property defined in [17] that cannot be checked numerically as it involves infinitely many linear polynomials f.

Remark 4.7. Observe that the usual rank condition (3.5) used as a stopping criterion to detect whether (4.8) is exact (i.e., $\rho_1 = 0$) cannot be satisfied in solving (4.8) with primal dual interior point methods (as in the SDP solvers used by GloptiPoly) because one tries to find an optimal solution $\mathbf{z}^*$ in the *relative interior* of the feasible set of (4.8), and this gives maximum rank to the moment matrix $M_{d_j}(\mathbf{z}^*)$. Therefore, in the context of (4.8), if, indeed, $\rho_j = 0$, then $\mathbf{z}^*$ corresponds to the moment vector of some probability measure μ supported on the set of points $(x, x) \in \mathbf{K} \times \mathbf{K}$ that satisfy $g_j(x) = 0$ (as, indeed, $L_{\mathbf{z}^*}(\langle \nabla g_j(Y), X - Y \rangle) = 0 = \rho_j$). Therefore, $\rho_j = 0$ as d_j increases, but the rank of $M_{d_j}(\mathbf{z}^*)$ does not stabilize because μ is not finitely supported. In particular, a good candidate $\mathbf{z}^*$ for an optimal solution is the moment vector of the probability measure uniformly distributed on the set $\{(x, x) \in \mathbf{K} \times \mathbf{K} \,:\, g_j(x) = 0\}$.

Alternatively, if $\rho_j \approx 0$ and the dual of (4.8) has an optimal solution $(\sigma_{jk}, \psi_{jk}, \psi_j)$, then in some cases one may check if (4.6) holds exactly after appropriate rounding of coefficients of the solution. But in general obtaining an exact certificate (i.e., $\rho_j = 0$ in the primal or (4.6) with $\epsilon = 0$ in the dual) numerically is hopeless.

Example 4.2. Consider the following simple illustrative example in $\mathbb{R}^2$:

$$
\mathbf{K} := \left\{ x \in \mathbb{R}^2 \,:\, x_1 x_2 - 1/4 \geq 0;\; 0.5 - (x_1 - 0.5)^2 - (x_2 - 0.5)^2 \geq 0 \right\}. \tag{4.9}
$$

Obviously, $\mathbf{K}$ is convex but its defining polynomial $x \mapsto g_1(x) := x_1 x_2 - 1/4$ is not concave, whereas $x \mapsto g_2(x) := 0.5 - (x_1 - 0.5)^2 - (x_2 - 0.5)^2$ is.

With $d_1 = 3$, solving (4.8) using GloptiPoly 3[5] yields the optimal value $\rho_1 \approx -4.58.10^{-11}$ which, in view of the machine precision for the SDP solvers used in GloptiPoly, could be considered to be zero, but of course with no guarantee. However, and according to Remark 4.7, we could check that (again up to machine precision) for every $\alpha \in \mathbb{N}^n$ with $|\alpha| \leq 2d_j$, $z^*_{\alpha,\alpha} = z^*_{2\alpha,0}$ and $z^*_{\alpha,0} = z^*_{0,\alpha}$. In addition, because of symmetry, $z_{\alpha,\beta} = z_{\alpha',\beta'}$ whenever $\alpha'_1 = \alpha_2$ and $\alpha'_2 = \alpha_1$ (and similarly for β and β'). Indeed, for moments of order 1 we have $z^*_{\alpha,\beta} = (0.5707, 0.5707, 0.5707, 0.5707)$ and for moments of order 2,

$$
z^*_{\alpha,\beta} = (0.4090, 0.25, 0.4090, 0.25, 0.4090, 0.25, 0.4090, 0.4090, 0.25, 0.4090).
$$

[5] GloptiPoly 3 (a Matlab-based public software) is an extension of GloptiPoly [9] to solve the generalized problem of moments described in [18]. For more details see www.laas.fr/∼henrion/software/.

For $j = 2$ there is no test to perform because $-g_2$ being quadratic and convex yields

$$\langle \nabla g_2(Y), X - Y \rangle \;=\; g_2(X) - g_2(Y) + \underbrace{(X-Y)^T \left(-\nabla^2 g_2(Y)\right)(X-Y)}_{SOS} \tag{4.10}$$

which is in the form (4.6) with $d_2 = 1$.

We next show the role of Assumption 4.6 in obtaining a semidefinite representation of $\mathbf{K}$.

THEOREM 4.8. *Let Assumption* 2.1 *and Slater's condition hold. Moreover, assume that for every* $j = 1, \ldots, m$, $\nabla g_j(y) \neq 0$ *whenever* $y \in \mathbf{K}$ *and* $g_j(y) = 0$. *If Assumption* 4.6 *holds, then* $\mathbf{K}$ *is convex and* Ω *in* (4.2) *with* $d := \max_j d_j$ *is a semidefinite representation of* $\mathbf{K}$.

Proof. That $\mathbf{K}$ is convex follows from Lemma 4.3. We next prove that the PP-BDR property defined in Lasserre [16] holds for $\mathbf{K}$. Let $f \in \mathbb{R}[X]$ be a linear polynomial with coefficient vector $\mathbf{f} \in \mathbb{R}^n$ (i.e., $X \mapsto f(X) = \mathbf{f}^T X$), and consider the optimization problem $\mathbf{P}: \min\{\mathbf{f}^T x \;:\; x \in \mathbf{K}\}$. As $\mathbf{K}$ is compact, let $x^* \in \mathbf{K}$ be a global minimizer of f. The Fritz John optimality conditions state that there exists $0 \neq \lambda \in \mathbb{R}^{m+1}_+$ such that

$$\lambda_0 \, \mathbf{f} = \sum_{j=1}^{m} \lambda_j \, \nabla g_j(x^*); \quad \lambda_j \, g_j(x^*) = 0 \quad \forall j = 1, \ldots, m. \tag{4.11}$$

(See, e.g., [3].) We first prove by contradiction that if Slater's condition and the nondegeneracy assumption hold, then $\lambda_0 > 0$. Suppose that $\lambda_0 = 0$ and let $J := \{j \in \{1, \ldots, m\} \;:\; \lambda_j > 0\}$; hence, J is nonempty as $\lambda \neq 0$. With $x_0 \in \mathbf{K}$ such that $g_j(x_0) > 0$ (as Slater's condition holds, one such x_0 exists), let $B(x_0, \rho) := \{z \;:\; \|z - x_0\| \leq \rho\}$. For ρ sufficiently small, $B(x_0, \rho) \subset \mathbf{K}$ and $g_j(z) > 0$ for all $z \in B(x_0, \rho)$ and every $j = 1, \ldots, m$. Then by (4.11) and $\lambda_0 = 0$,

$$0 = \sum_{j=1}^{m} \lambda_j \, \langle \nabla g_j(x^*), z - x^* \rangle \qquad \forall z \in B(x_0, \rho),$$

which in turn implies (by nonnegativity of each term in the above sum)

$$\langle \nabla g_j(x^*), z - x^* \rangle = 0 \qquad \forall z \in B(x_0, \rho), \; j \in J.$$

But this clearly implies $\nabla g_j(x^*) = 0$ for every $j \in J$, in contradiction with the nondegeneracy assumption. Hence, $\lambda_0 > 0$ and, by homogeneity, we may and will take $\lambda_0 = 1$.

Therefore, letting $Y := x^*$ in (4.6), the polynomial $X \mapsto f(X) - f^*$ can be written

$$\begin{aligned}
\mathbf{f}^T X - f^* &= \sum_{j=1}^{m} \lambda_j \, [\, \langle \nabla g_j(x^*), X - x^* \rangle \,] \\
&= \sum_{j=1}^{m} \lambda_j \left[\sum_{k=0}^{m} \sigma_{jk}(X, x^*) \, g_k(X) + \sum_{k=0, k \neq j}^{m} \psi_{jk}(X, x^*) \, g_k(x^*) \right. \\
&\qquad \left. + \, \psi_j(X, x^*) \, g_j(x^*) \right],
\end{aligned}$$

where we have used (4.6) with $Y = x^*$ and $\epsilon = 0$. Next, observe that

$$\begin{aligned} X \mapsto \sigma_{jk}(X, x^*) \in \Sigma^2[X] \qquad & [\text{as } \sigma_{jk} \in \Sigma^2[X, Y]], \\ X \mapsto \psi_{jk}(X, x^*)\, g_k(x^*) \in \Sigma^2[X] \qquad & [\text{as } \psi_{jk} \in \Sigma^2[X, Y] \text{ and } g_j(x^*) \geq 0], \\ \lambda_j g_j(x^*) = 0, \qquad & j = 1, \ldots, m. \end{aligned}$$

So as $\lambda \in \mathbb{R}^m_+$,

$$X \mapsto \mathbf{f}^T X - f^* \;=\; \Delta_0(X) + \sum_{j=1}^m \Delta_j(X)\, g_j(X) \tag{4.12}$$

for SOS polynomials $(\Delta_j)_{j=0}^m \subset \Sigma^2[X]$ defined by

$$\begin{aligned} X \mapsto \Delta_0(X) &= \sum_{j=1}^m \lambda_j \left(\sum_{k=0, k\neq j}^m \psi_{jk}(X, x^*)\, g_k(x^*) \right), \\ X \mapsto \Delta_j(X) &= \sum_{l=1}^m \lambda_l\, \sigma_{lj}(X, x^*), \qquad j = 1, \ldots, m. \end{aligned}$$

Write every affine polynomial $f \in \mathbb{R}[X]$ as $\mathbf{f}^T X + f_0$ for some $\mathbf{f} \in \mathbb{R}^n$ and $f_0 = f(0)$. If f is nonnegative on $\mathbf{K}$, then from (4.12),

$$\begin{aligned} f(X) \;=\; \mathbf{f}^T X - f^* + f^* + f_0 &= f^* + f_0 + \Delta_0(X) + \sum_{j=1}^m \Delta_j(X)\, g_j(X) \\ &= \widehat{\Delta}_0(X) + \sum_{j=1}^m \Delta_j(X)\, g_j(X) \qquad \forall X, \end{aligned}$$

with $\widehat{\Delta}_0 \in \Sigma^2[X]$ (because $f^* + f_0 \geq 0$), and so the PP-BDR property holds for $\mathbf{K}$ with order d. By [16, Theorem 2], $\mathbf{K}$ is SDr with the semidefinite representation (4.2). ∎

We next show that the two sufficient conditions of strict convexity and SOS convexity of Helton and Nie [6] in Theorem 4.2 both imply that Assumption 4.6 holds, and so Theorem 4.8 contains Theorem 4.2 as a special case.

COROLLARY 4.9. *Let* $\mathbf{K}$ *in* (1.1) *be convex and both Assumption* 2.1 *and Slater's condition hold. Assume that either* $-g_j$ *is SOS-convex or* $-g_j$ *is convex on* $\mathbf{K}$ *and* $-\nabla^2 g_j \succ 0$ *on* $\mathbf{K} \cap \{x \,:\, g_j(x) = 0\}$ *for every* $j = 1, \ldots, m$. *Then Assumption* 4.6 *holds and so Theorem* 4.8 *applies.*

Proof. By Lemma 2.5, for every $j = 1, \ldots, m$, write

$$\begin{aligned} (X, Y) \quad &\mapsto \quad g_j(X) - g(Y) - \langle \nabla g_j(Y), X - Y \rangle \\ &= \left\langle (X - Y), \underbrace{\left(\int_0^1 \int_0^t \nabla^2 g_j(Y + s(X - Y))\, ds dt \right)}_{F_j(X,Y)} (X - Y) \right\rangle. \end{aligned}$$

If $-\nabla^2 g_j \succ 0$ on $y \in \mathbf{K}$ with $g_j(y) = 0$, then from the proof of [6, Lemma 19], $-F_j(x, y) \succ 0$ for all $x, y \in \mathbf{K}$ with $g_j(y) = 0$. In other words, $-F_j(x, y) \succeq \delta I_n$ on

Ω_j (defined in (4.5)) for some $\delta > 0$. Therefore, by the matrix polynomial version of Putinar positivstellensatz in [6, Theorem 29],

$$-F_j(X,Y) = \sum_{k=0}^{m} \widehat{\sigma}_{jk}(X,Y)g_k(X) + \sum_{k=0,k\neq j}^{m} \widehat{\psi}_{jk}(X,Y)g_k(Y) + \widehat{\psi}_j(X,Y)g_j(Y) \tag{4.13}$$

for some SOS matrix polynomials $(\widehat{\sigma}_{jk}(X,Y))$, $(\widehat{\psi}_{jk}(X,Y))$ and some matrix polynomial $\widehat{\psi}_j(X,Y)$.

On the other hand, if $-g_j$ is SOS-convex, then by Lemma 2.4, $-F_j(X,Y)$ is SOS and, therefore, (4.13) also holds (take $\widehat{\sigma}_{jk} \equiv 0$ for all $k \neq 0$, and $\widehat{\psi}_{jk} \equiv 0$ for all k and $\widehat{\psi}_j \equiv 0$). But then

$$\begin{aligned}
g_j(X) - g(Y) - \langle \nabla g_j(Y), X - Y\rangle &= \langle (X-Y), F_j(X,Y)(X-Y)\rangle \\
&= -\sum_{k=0}^{m} \langle (X-Y), \widehat{\sigma}_{jk}(X,Y)(X-Y)\rangle\, g_k(X) \\
&\quad - \sum_{k=0,k\neq j}^{m} \Big\langle (X-Y), \widehat{\psi}_{jk}(X,Y)(X-Y)\Big\rangle\, g_k(Y) \\
&\quad - \Big\langle (X-Y), \widehat{\psi}_j(X,Y)(X-Y)\Big\rangle\, g_j(Y) \\
&= -\sum_{k=0}^{m} \sigma_{jk}(X,Y)\, g_k(X) \\
&\quad - \sum_{k=0,k\neq j}^{m} \psi_{jk}(X,Y)\, g_k(Y) - \psi_j(X,Y)\, g_j(Y)
\end{aligned}$$

for all X, Y, for some SOS polynomials $\sigma_{jk}, \psi_{jk} \in \mathbb{R}[X,Y]$, and some polynomial $\psi_j \in \mathbb{R}[X,Y]$. Equivalently,

$$\begin{aligned}
\langle \nabla g_j(Y), X - Y\rangle &= g_j(X) - g_j(Y) + \sum_{k=0}^{m} \sigma_{jk}(X,Y)\, g_k(X) \\
&\quad + \sum_{k=0,k\neq j}^{m} \psi_{jk}(X,Y)\, g_k(Y) + \psi_j(X,Y)\, g_j(Y) \\
&= \sum_{k=0}^{m} \sigma'_{jk}(X,Y)\, g_k(X) + \sum_{k=0,k\neq j}^{m} \psi_{jk}(X,Y)\, g_k(Y) \\
&\quad + \psi'_j(X,Y)\, g_j(Y)
\end{aligned}$$

for some SOS polynomials $\sigma'_{jk}, \psi_{jk} \in \Sigma^2[X,Y]$ and some polynomial $\psi'_j \in \mathbb{R}[X,Y]$. In other words, Assumption 4.6 holds, which concludes the proof. □

Hence, if each $-g_j$ is SOS-convex or convex on $\mathbf{K}$ with $-\nabla^2 g_j \succ 0$ on $\mathbf{K} \cap \{x : g_j(x) = 0\}$, one obtains a numerical scheme to obtain the parameter d in Theorem 4.8 as well as the semidefinite representation (4.2) of $\mathbf{K}$. Solve the semidefinite programs (4.8) with degree parameter d_j. Eventually, $\rho_j = 0$ for every $j = 1, \ldots, m$.

Example 4.3. Consider the convex set $\mathbf{K}$ in (4.9) of Example 4.2 for which the defining polynomial g_1 of $\mathbf{K}$ is not concave. We have seen that Assumption 4.6 holds

(up to $\rho_1 \approx 10^{-11}$, close to machine precision) and $\max[d_1, d_2] = 3$. By Theorem 4.8, if ρ_1 would be exactly 0, the set

$$\Omega := \left\{ (x, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^{s(6)} \; : \; \begin{cases} M_3(\mathbf{y}) & \succeq 0, \\ M_2(g_j\,\mathbf{y}) & \geq 0, \quad j = 1, 2, \\ L_{\mathbf{y}}(X_i) & = x_i, \quad i = 1, 2, \\ y_0 & = 1 \end{cases} \right. \tag{4.14}$$

would be a semidefinite representation of $\mathbf{K}$.

At least in practice, for every linear polynomial $f \in \mathbb{R}[X]$, minimizing $L_{\mathbf{y}}(f)$ over Ω yields the desired optimal value $f^* := \min_{x \in \mathbf{K}} f(x)$, up to $\rho_1 \approx -10^{-11}$.

Indeed, let $f \in \mathbb{R}[X]$ be $\mathbf{f}^T X$ for some vector $\mathbf{f} \in \mathbb{R}^n$. In minimizing f over $\mathbf{K}$, one has $\mathbf{f} = \lambda_1 \nabla g_1(x^*) + \lambda_2 \nabla g_2(x^*)$ for some $\lambda \in \mathbb{R}^2_+$, some $x^* \in \mathbf{K}$ with $\lambda_i g_i(x^*) = 0$, $i = 1, 2$, and $f^* = \lambda_1 \langle \nabla g_1(x^*), x^* \rangle + \lambda_2 \langle \nabla g_2(x^*), x^* \rangle = \min_{x \in \mathbf{K}} \mathbf{f}^T x$. Let x be as in (4.14), arbitrary. Then

$$\mathbf{f}^T x - f^* \; = \; L_{\mathbf{y}}(f(X) - f^*) \; = \; \sum_{i=1}^{2} \lambda_i L_{\mathbf{y}}(\langle \nabla g_i(x^*), X - x^* \rangle).$$

If $\lambda_1 > 0$ so that $g_1(x^*) = 0$, use (4.12) to obtain

$$L_{\mathbf{y}}(\langle \nabla g_1(x^*), X - x^* \rangle) = L_{\mathbf{y}} \left(\rho_1 + \Delta_0(X) + \sum_{j=1}^{2} \Delta_j(X) g_j(X) \right) \geq \rho_1,$$

because $L_{\mathbf{y}}(\Delta_0) \geq 0$ follows from $M_3(\mathbf{y}) \succeq 0$, and $L_{\mathbf{y}}(\Delta_j g_j) \geq 0$, $j = 1, 2$, follows from $M_2(g_1\mathbf{y}), M_2(g_2\mathbf{y}) \succeq 0$. If $\lambda_2 > 0$ so that $g_2(x^*) = 0$, then from (4.10)

$$L_{\mathbf{y}}(\langle \nabla g_2(x^*), X - x^* \rangle) \; = \; L_{\mathbf{y}} \left(g_2(X) - \left\langle (X - x^*), \nabla^2 g_2(x^*)(X - x^*) \right\rangle \right) \geq 0,$$

because $L_{\mathbf{y}}(g_2) \geq 0$ follows from $M_2(g_2\mathbf{y}) \succeq 0$ whereas the second term is nonnegative as $\langle (X - x^*), -\nabla^2 g_2(x^*)(X - x^*) \rangle$ is SOS and $M_3(\mathbf{y}) \succeq 0$. Hence, $\mathbf{f}^T x - f^* \geq \lambda_1 \rho_1$. On the other hand, from $\mathbf{K} \subseteq \{x \; : \; (x, y) \in \Omega\}$, one finally obtains the desired result

$$f^* + \lambda_1 \rho_1 \; \leq \min \{ \mathbf{f}^T x \; : \; (x, y) \in \Omega \} \; \leq \; f^*.$$

5. Conclusion. As well-known, convexity is a highly desirable property in optimization. We have shown that it also has important specific consequences in polynomial optimization. For instance, for polynomial optimization problems with SOS-convex or strictly convex polynomial data, the basic SDP relaxations of the moment approach [13] *recognize* convexity and finite convergence occurs. Similarly, the set $\mathbf{K}$ has a semidefinite representation, explicit in terms of the defining polynomials (g_j).

The class of SOS-convex polynomials introduced in Helton and Nie [6] is particularly interesting because the semidefinite constraint to handle the semidefinite relaxation involves only the Hankel-like moment matrix which does *not* depend on the problem data! Hence, one might envision a dedicated SDP solver that would take into account this peculiarity as Hankel-like or Toeplitz-like matrices enjoy very specific properties. Moreover, if restricted to this class of polynomials, Jensen's inequality can be extended to linear functionals in the dual cone of SOS polynomials (hence, not necessarily probability measures).

Therefore, a topic of further research is to evaluate how *large* is the subclass of SOS-convex polynomials in the class of convex polynomials and, if possible, to also provide simple sufficient conditions for SOS convexity.

Acknowledgments. The author wishes to thank L. Tuncel and Y. Nesterov for helpful discussions on various characterizations of convex sets, and also two anonymous referees for several corrections as well as suggestions and remarks to improve a first version of this paper.

REFERENCES

[1] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*, MPS/SIAM Ser. Optim. 2, SIAM, Philadelphia, 2001.

[2] C.B. Chua and L. Tuncel, *Invariance and efficiency of convex representations*, Math. Program., 111 (2008), pp. 113–140.

[3] S.I. Birbil, J.B.G. Frenk, and G.J. Still, *An elementary proof of the Fritz-John and Karush-Kuhn-Tucker conditions in nonlinear programming*, European J. Oper. Res., 180 (2007), pp. 479–484.

[4] R.E. Curto and L.A. Fialkow, *Recursiveness, positivity, and truncated moment problems*, Houston J. Math., 17 (1991), pp. 603–635.

[5] D. Handelman, *Representing polynomials by positive linear functions on compact convex polyhedra*, Pacific J. Math., 132 (1988), pp. 35–62.

[6] J.W. Helton and J. Nie, *Semidefinite representation of convex sets*, Math. Program., to appear.

[7] J.W. Helton and J. Nie, *Sufficient and necessary condition for semidefinite representation of sets*, SIAM J. Optim., to appear.

[8] D. Henrion, *On Semidefinite Representations of Plane Quartics*, Research report 08444, LAAS-CNRS, University of Toulouse, Toulouse, France, 2008, submitted.

[9] D. Henrion and J.B. Lasserre, *GloptiPoly: Global optimization over polynomials with Matlab and SeDuMi*, ACM Trans. Math. Software, 29 (2003), pp. 165–194.

[10] M. Kojima and M. Maramatsu, *An extension of sums of squares relaxations to polynomial optimization problems over symmetric cones*, Math. Program., 110 (2007), pp. 315–336.

[11] C.W.J. Hol and C.W. Scherer, *A sum-of-squares approach to fixed order H_∞-synthesis*, in Positive Polynomials in Control, Lecture Notes in Control and Inform. Sci. 312, D. Henrion and A. Garulli, eds., Springer-Verlag, Berlin, 2005, pp. 45–71.

[12] J.L. Krivine, *Anneaux préordonnés*, J. Anal. Math., 12 (1964), pp. 307–326.

[13] J.B. Lasserre, *Global optimization with polynomials and the problem of moments*, SIAM J. Optim., 11 (2001), pp. 796–817.

[14] J.B. Lasserre, *Semidefinite programming vs. LP relaxations for polynomial programming*, Math. Oper. Res., 27 (2002), pp. 347–360.

[15] J.B. Lasserre, *Polynomial programming: LP-relaxations also converge*, SIAM J. Optim., 15 (2005), pp. 383–393.

[16] J.B. Lasserre, *Convex sets with semidefinite representation*, Math. Program., to appear.

[17] J.B. Lasserre, *Representation of nonnegative convex polynomials*, Arch. Math. (Basel), 91 (2008), pp. 126–130.

[18] J.B. Lasserre, *A semidefinite programming approach to the generalized problem of moments*, Math. Program., 112 (2008), pp. 65–92.

[19] A.S. Lewis, P. Parrilo, and M.V. Ramana, *The Lax conjecture is true*, Proc. Amer. Math. Soc., 133 (2005), pp. 2495–2499.

[20] Y.E. Nesterov and A.S. Nemirovski, *Self-concordant Functions and Polynomial Time Methods in Convex Programming*, report, Central Economical and Mathematical Institute, USSR Academy of Sciences, Moscow, 1989.

[21] P.A. Parrilo, *Semidefinite programming relaxations for semialgebraic problems*, Math. Program. Ser. B, 96 (2003), pp. 293–320.

[22] P. Parrilo, *Exact semidefinite representations for genus zero curves*, in lecture at the Banff workshop Positive Polynomials and Optimization, Banff, Canada, Banff International Research Station, 2006.

[23] M.D. Perlman, *Jensen's inequality for a convex vector-valued function on an infinite dimensional space*, J. Multivariate Anal., 4 (1974), pp. 52–65.

[24] M. Putinar, *Positive polynomials on compact semi-algebraic sets*, Indiana Univ. Math. J., 42 (1993), pp. 969–984.
[25] J. Bochnak, M. Coste, and M.-F. Roy, *Real Algebraic Geometry*, Springer-Verlag, Berlin, 1998.
[26] C. Scheiderer, *Positivity and sums of squares: A guide to recent results*, in Emerging Applications of Algebraic Geometry, IMA Vol. 2 Mathematics and Its Applications, Minneapolis, M. Putinar and S. Sullivant, eds., Springer, New York, 2008, pp. 271–324.
[27] K. Schmüdgen, *The K-moment problem for compact semi-algebraic sets*, Math. Ann., 289 (1991), pp. 203–206.
[28] R. Schneider, *Convex Bodies: The Brunn–Minkowski Theory*, Cambridge University Press, Cambridge, United Kingdom, 1994.
[29] M. Schweighofer, *Optimization of polynomials on compact semialgebraic sets*, SIAM J. Optim., 15 (2005), pp. 805–825.
[30] F.-H. Vasilescu, *Spectral measures and moment problems*, in Spectral Theory and Its Applications, Theta Ser. Adv. Math. 2, Theta, Bucharest, 2003, pp. 173–215.